



Evaluation of Humanitarian Action Guide

ALNAP is a unique sector-wide active learning membership network dedicated to improving the quality and accountability of humanitarian action.

www.alnap.org

Authors:

Margie Buchanan-Smith

John Cosgrave

Alexandra Warner

With additional help from:

Francesca Bonino

Neil Dillon

Paul Knox Clarke

Ralf Otto

The views contained in this report do not necessarily reflect those of the ALNAP Members.



ALNAP (2016) Evaluation of Humanitarian Action Guide.
ALNAP Guide. London: ALNAP/ODI.

© ALNAP/ODI 2016. This work is licensed under a Creative Commons Attribution-Non Commercial Licence (CC BY-NC 4.0).

ISBN 978-1-910454-49-7

Communications management by
Maria Gili and Alex Glynn

Design by Human After All
www.humanafterall.co.uk

Foreword by John Mitchell

Director, ALNAP



Twenty years ago, when ALNAP was first established, humanitarian evaluation was still in its infancy. Since then, EHA has grown to be an integral part of the way the humanitarian system operates, with countless good quality evaluations. As a result, learning and accountability in humanitarian action have improved significantly.

The knowledge generated from the practice of EHA is now part of an ever-growing global evidence base, which is being used to create baselines from which humanitarian performance is monitored over time, as shown in the ALNAP State of the System Report. The need for high quality evaluative material to improve the precision and level of confidence of performance reporting, will always be vital to the humanitarian community. I believe that this Guide will provide a great service in this respect.

Developing a truly comprehensive Guide has been a formidable task. It has taken five years to produce and covers many aspects of humanitarian evaluation, all of which are part of a highly complex operating environment. It has also benefited from an unusually high level of participation, with over 40 organisations testing and providing feedback through the pilot process. Without these contributions, we could not have created such clear, relevant and authoritative guidance that has already proved invaluable to evaluators. I would like to sincerely thank all those who have taken part.

We hope the Guide will be widely used and will add to the quality and enjoyment of conducting evaluations of humanitarian action. It will also continue to enhance accountability and promote better evaluative evidence for lesson learning and analysis, whether system-wide, or for individual organisations and their programmes. We would like to invite users to continue to provide feedback for the Guide as we are committed to ensuring the quality of the guidance and, subsequently, the impact of EHA.



ALNAP

Table of Contents

Introduction

03	Foreword
10	Acronyms
12	Acknowledgements
14	Why evaluation matters
14	Why do we need an EHA Guide?
16	Developing the ALNAP EHA Guide
18	How to use the ALNAP EHA Guide
20	How is the Guide organised?
22	Key to design features

1 / What is Evaluation of Humanitarian Action?

24	1.1 Humanitarian crises and humanitarian action
26	1.2 Evaluation of humanitarian action
30	1.3 Monitoring and evaluation in the humanitarian response cycle
32	1.4 Common challenges faced in EHA

2 / Deciding to do an evaluation

41	2.1 When is evaluation suitable for accountability and learning purposes?
47	2.2 Is evaluation the right tool for the job?
48	2.3 Balancing learning and accountability
49	2.4 Deciding what and when to evaluate
55	2.5 Ethics and EHA

3 / Think early and often about evaluation utilisation

62	3.1 What it means to be focused on use
63	3.2 How evaluations of humanitarian action are used
65	3.3 Identifying, understanding and engaging intended users of the evaluation
72	3.4 The role of the affected population in the planning stage
73	3.5 Factors affecting utilisation of evaluations

4 / Types of evaluation

- 78** **4.1** What type of evaluation are you going to undertake?
- 79** **4.2** Defining different evaluation types
- 87** **4.3** Real-time evaluations (RTEs)
- 89** **4.4** Joint evaluations

5 / Framing your evaluation

- 92** **5.1** The programme logic
- 97** **5.2** Theories of change
- 100** **5.3** Developing a logic model after the fact
- 101** **5.4** Evaluation frameworks and evaluation criteria

6 / Choosing evaluation questions

- 104** **6.1** Impact of question choice on quality
- 107** **6.2** Types of evaluation question
- 109** **6.3** Number of questions
- 110** **6.4** Unpacking overarching questions
- 111** **6.5** Evaluation criteria
- 116** **6.6** Selecting the best questions

7 / Terms of Reference and budgeting for an evaluation

- 118** **7.1** Terms of Reference (ToR)
 - 119** **7.1.1** The ToR and inception report
 - 121** **7.1.2** What does the ToR include?
 - 127** **7.1.3** ToR timeline
 - 127** **7.1.4** ToR length
- 128** **7.2** Budgeting for an evaluation

8 / Inception phase

- 136 8.1** Inception report and quality control
- 138 8.2** Activities in the inception phase
- 138 8.3** The evaluation matrix
- 142 8.4** The inception report

9 / Planning and managing your evaluation

- 147 9.1** The decision to evaluate
- 148 9.2** Advisory groups
- 153 9.3** Evaluation timeline
- 154 9.4** Internal and external evaluation
- 157 9.5** Contracting evaluators
- 168 9.6** Leadership and teamwork challenges
- 170 9.7** Managing conflict

10 / Desk methods

- 172 10.1** What is a desk review?
- 174 10.2** Why do a desk review?
- 175 10.3** How long should a desk review take?
- 175 10.4** Conducting a desk review
- 178 10.5** Tools for desk reviews
- 182 10.6** Summarising data through desk review

11 / Evaluation designs for answering evaluation questions

- 193 11.1** Qualitative and quantitative
- 194 11.2** Families of design
- 200 11.3** Selecting a design
- 209 11.4** Bias

12 / Sampling

- 215 12.1 Common sampling problems in EHA**
- 216 12.2 Non-random sampling**
- 224 12.3 Random sampling**
- 230 12.4 Sampling for mixed methods**

13 / Field methods

- 232 13.1 Interviewing as a technique**
- 236 13.2 Interpreters**
- 238 13.3 Qualitative interview methods**
- 246 13.4 Survey methods**
- 253 13.5 Observation**
- 257 13.6 Unobtrusive measures**
- 259 13.7 Learning-oriented methods**

14 / Engaging with the affected population in your evaluation

- 268 14.1 Engaging with the affected population**
- 271 14.2 Planning to engage with the affected population**
- 273 14.3 Designing the evaluation to engage with the affected population**
- 274 14.4 Methods for engaging with the affected population**
- 279 14.5 Particular challenges in engaging with the affected population in EHA**
- 280 14.6 Ethical issues in engaging with affected populations**

15 / Constrained access

- 283 15.1 What is constrained access?**
- 284 15.2 The importance of an evaluability assessment**
- 286 15.3 Ways to overcome constrained access**
- 291 15.4 Credibility of remote evaluation**
- 292 15.5 Other options to remote evaluation**

16 / Analysis

- | | |
|------------|--|
| 294 | 16.1 Big-n or small-n |
| 296 | 16.2 Analysing evidence to answer normative questions |
| 297 | 16.3 Analysing evidence for evaluative questions |
| 302 | 16.4 Analysing evidence for causal questions |
| 303 | 16.5 Analysing evidence to answer descriptive questions |
| 311 | 16.6 Statistical analysis of primary data |
| 316 | 16.7 Numerical analysis of secondary data |
| 318 | 16.8 From evidence to recommendations |

17 / Reporting and communicating evaluation findings with a utilisation focus

- | | |
|------------|--|
| 328 | 17.1 Key outputs |
| 330 | 17.2 Evaluation report |
| 339 | 17.3 Dissemination |
| 342 | 17.4 Different communication channels |
| 346 | 17.5 Facilitating take-up of an evaluation |
| 349 | 17.6 Evaluation synthesis, thematic reviews and meta analyses |

18 / Humanitarian impact evaluations

- | | |
|------------|---|
| 356 | 18.1 Why is impact evaluation important? |
| 356 | 18.2 What is impact in EHA? |
| 359 | 18.3 Planning an impact evaluation |
| 362 | 18.4 Approaches and designs |
| 367 | Glossary |

The bibliography and index are available at www.alnap.org/EHA

Acronyms

AAP	Accountability to affected populations
AAR	After-action review
ACF	Action Contre la Faim // Action Against Hunger
ACT	Action by Churches Together
ALNAP	Active Learning Network for Accountability and Performance in Humanitarian Action
CAQDAS	Computer Assisted Qualitative Data Analysis
CAR	Central African Republic
CBO	Community based organisations
CDAC	Communicating with Disaster Affected Communities
CDA	CDA Collaborative Learning Projects
CDR	Community-driven reconstruction
CERF	Central Emergency Response Fund
CFW	Cash for work
CHS	Common Humanitarian Standard
CRS	Catholic Relief Services
CwC	Communicating with Communities
DEC	Disasters Emergency Committee
DFID	UK Department for International Development
DRC	Democratic Republic of Congo
DRC	Danish Refugee Council
DRR	Disaster risk reduction
ECB	Emergency Capacity Building Project
ECHO	Directorate General for Humanitarian Aid and Civil Protection
EHA	Evaluation of Humanitarian Action
FAO	United Nations Food and Agriculture Organization
FGD	Focus Group Discussion
FFS	Farmer Field School
HAP	Humanitarian Accountability Partnership
HC	Humanitarian Coordinator
HCT	Humanitarian Country Team
IAHE	Inter-Agency Humanitarian Evaluation
IASC	Inter-Agency Standing Committee
ICRC	International Committee of the Red Cross
IDP	Internally displaced person
IFRC	International Federation of Red Cross and Red Crescent Societies
INGO	International non-governmental organisation
IRC	International Rescue Committee
JEEAR	Joint Evaluation of Emergency Assistance to Rwanda
JHIE	Joint Humanitarian Impact Evaluation

LRRD	Linking Relief Rehabilitation and Development
M&E	Monitoring and evaluation
MEAL	Monitoring, evaluation and learning
MRM	Management Response Matrix
MSC	Most Significant Change
MSF	Médecins Sans Frontières
NGO	Non-governmental organisation
NNGO	National non-governmental organisation
NRC	Norwegian Refugee Council
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
ODI	Overseas Development Institute
OECD-DAC	Organisation for Economic Co-operation and Development – Development Assistance Committee
OFDA	Office of US Foreign Disaster Assistance (USAID)
PRA	Participatory rapid appraisal
PDA	Personal digital assistant
RAPID	Overseas Development Institute's Research and Policy in Development Programme
RCT	Randomised control trial
RfP	Request for proposals
ROI	Region of Origin Initiative in Afghanistan
ROMA	RAPID Outcome Mapping Approach
RTE	Real-time evaluation
SAVE	Secure Access in Volatile Environments
SCHR	Steering Committee for Humanitarian Response
Sida	Swedish International Development Cooperation Agency
SOHS	State of the Humanitarian System
TEC	Tsunami Evaluation Coalition
ToC	Theory of change
ToR	Terms of Reference
UNDP	United Nations Development Programme
UNEG	United Nations Evaluation Group
UNEG HEIG	United Nations Evaluation Group's Humanitarian Evaluation Interest Group
UNHCR	United Nations High Commissioner for Refugees
UNICEF	United Nations Children's Fund
WASH	Water, sanitation, and hygiene
WFP	World Food Programme
WHO	World Health Organization
WVI	World Vision International

Acknowledgements

The development of Evaluation of Humanitarian Action Guide was an extensive and participative process made possible thanks to the openness, hard work and dedication of numerous individuals and organisations. The ALNAP Secretariat would like to acknowledge a few in particular.

First, the Guide would not have been possible without the continued positive engagement of the ALNAP Membership. The ALNAP Secretariat is extremely grateful for the network's honesty and keenness: it would not have been possible to create such a practical resource without their support.

The ALNAP Secretariat would also like to thank John Cosgrave and Margie Buchanan-Smith for their continued engagement and passionate work on drafting, improving, promoting and using the EHA Guide as well as previous EHA training materials over the course of the last several years.

A very special thank you also goes to Wendy Fenton, from the Humanitarian Practice Network (HPN), who initially launched a survey among HPN Members. Many of them recognised the need for new guidance on the topic of evaluating humanitarian action.

We acknowledge the valuable feedback provided by members of the advisory group representing evaluation stakeholders from across the sector: Jock Baker, Tony Beck, Hana Crowe, Stefan Dahlgren, Wendy Fenton, Enrique Garcia, Josse Gillijns, Saul Guerrero, Babar Kabir, Peter Klansoe, Caroline Loftus, Rob McCouch, Joakim Molander, Jonathan Patrick, Nicoletta Pergolizzi, Riccardo Polastro, Hannah Reichardt, Peta Sandison and Vivien Walden. They helped guide the ALNAP Secretariat through the drafting process.

A huge thank you is due to the hundreds of people who piloted and used the EHA Guide and who took the time to engage with it and get back to us with insights, feedback, suggestions for improvement, or simply a word of acknowledgement and appreciation. We were amazed by the willingness of the ALNAP Network Members and beyond to take part in the pilot, to review the EHA e-learning course co-developed with UNICEF, and to share their views on how to make the Guide more useful and practitioner-focused. Their feedback was invaluable in the revision and finalisation of the EHA Guide.

Thank you to the select group of pilots and EHA practitioners who helped steer us through some of the more difficult, higher-level pieces of feedback during the end-of-pilot feedback validation workshop. Special thanks to Sarah Bailey, Lian Bradley, Annie Devonport, Kathy Duryee, Josse Gillijns, Langdon Greenhalgh, Hélène Juillard, Anne-Claire Luzot, Velina Mikova, Mikkel Nedergaard, Joanna Olsen, Juliet Parker, Koorosh Raffii, Anke Reiffenstuel and Bonaventure Sokpoh. We are grateful to Ralf Otto for facilitating this discussion and for representing the group as a peer-reviewer for the final version.

Warm thanks are owed to Sian Cook, who provided invaluable research assistance during the revision of the EHA Guide.

Special thanks go to Alexandra Warner for her invaluable work coordinating the active pilot process, gathering its feedback and managing the finalisation of this Guide as well as providing key content. Many thanks go to John Mitchell and Paul Knox Clarke for their continued leadership and guidance, and to the Communication team in the ALNAP Secretariat for their continuous support and creativity that succeeded in making the EHA Guide an approachable and unintimidating resource. Thanks go out in particular to Patricia Curmi, Franziska Orphal, Maria Gili and Alex Glynn.

And last but certainly not least, we are most grateful to Francesca Bonino, who managed this initiative for four years. Your energy and enthusiasm helped propel this project forward and it would not have been possible without your efforts.

Why evaluation matters

How do we know if our humanitarian efforts are successful? Evaluation is one important way of finding out. At its core, evaluation aims to make an informed judgement on the value of activities and their results. Have we made a difference? Did we indeed help to save lives and alleviate suffering? Did we do so in the best way possible? Good and robust evaluations enable us to make a considered and evidence-based judgement on the degree to which a programme was successful, and the nature of the success. It enables stakeholders to answer the question of ‘so what?’ (Morra Imas and Rist, 2009) – looking beyond broad statements on success to focus on whether and/or why a project, programme or policy was worthwhile and for whom (Buffardi et al., 2015).

The push in the 1990s to increase quality and accountability in the humanitarian sector (Knox Clarke and Darcy, 2013), as well as the more recent pressure to demonstrate value for money and make the most of stretched resources (ALNAP, 2015), has made quality evaluations of humanitarian action much more important. Well-planned, designed and executed evaluations are processes that can assist both learning and accountability at a number of levels. In these often fast paced-working contexts, evaluations are a valuable opportunity to pause and take stock, from an objective perspective, of what is working or has worked, and what needs to change, giving structured insight into the overall performance of a project, programme or response. They contribute to the body of evidence on ‘what works’ and what does not in these often very challenging contexts. Evaluations can hence answer that difficult question of ‘how are we *really* doing?’ and assist decision-makers in making the necessary course corrections or tough choices.

Why do we need an EHA Guide?

There has been a surge in evaluations of humanitarian action since the 1990s. This has led to the creation of a critical mass of collective knowledge and experience to draw upon. By 2009, however, there was still no comprehensive guide for the sector on how to evaluate humanitarian interventions. Work on the Guide started in response to feedback from members of the Humanitarian Practice Network. When asked what would be their one request for guidance material, EHA came back loud and clear.

Six reasons it's time for an EHA Guide:

1

There is increasing interest and investment in evaluations as concerns are raised about the accountability and effectiveness of international development and humanitarian action.

2

There is now a critical mass of collective knowledge to build on – ALNAP’s evaluation database alone contains over 2,000 evaluations covering the last 30 years.

3

There is a need to create a common language and understanding of EHA in order to facilitate discussions within teams, organisations and across organisations.

4

Although evaluations have become more common practice, relatively few are clear about their methodology. Where they are clear, humanitarian agencies and evaluators often restrict themselves to a small number of designs and methods. There is room for improvement in the range of designs and methods used and in the selection of the most effective methodology for the evaluation questions posed.

5

Similarly, evaluations are still a common tool for donors to assess accountability, but can also be used by organisations to learn and improve their programmes. Overall, there are still many opportunities for humanitarian agencies to make the most of evaluations that fit these needs.

6

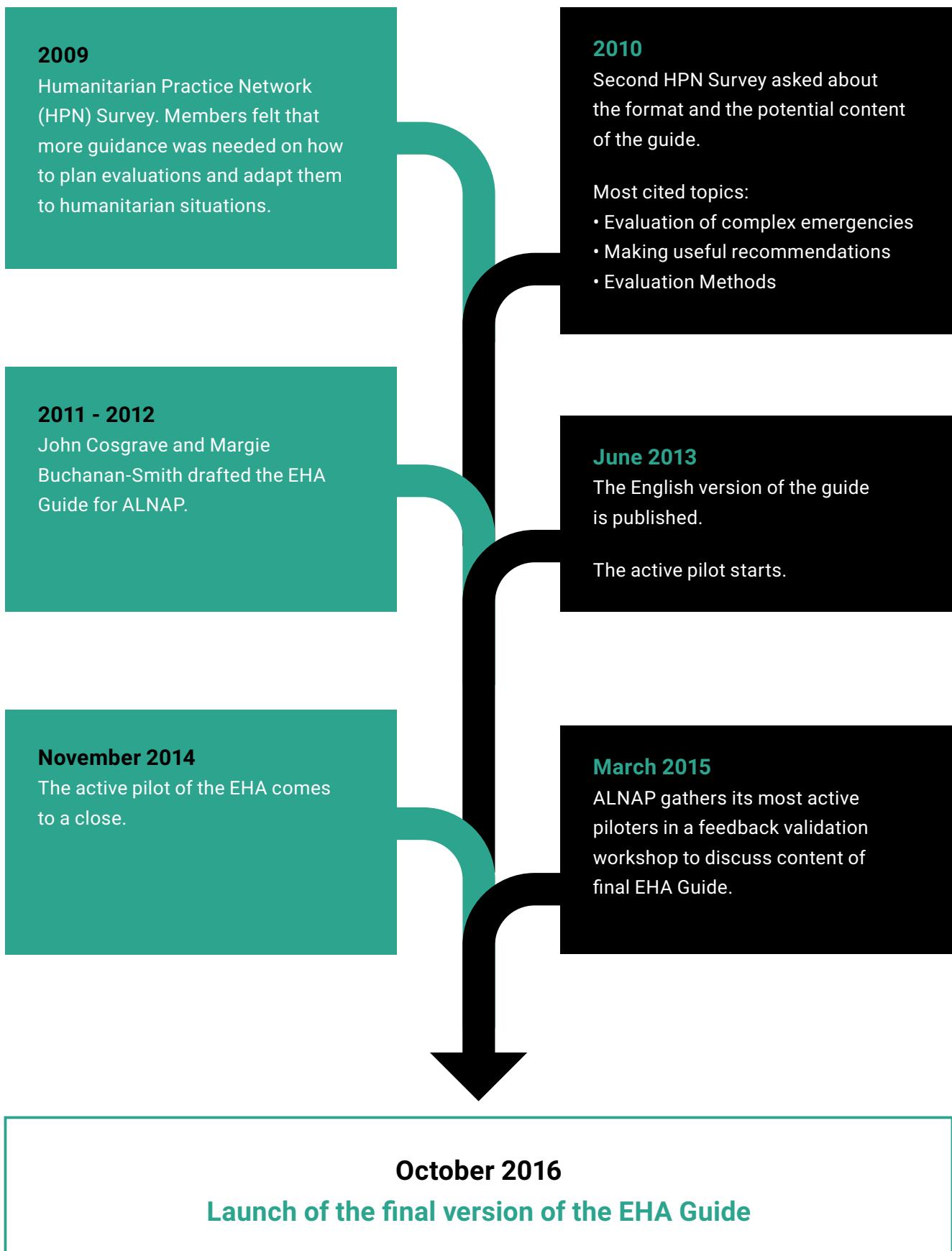
The commissioning of evaluations has shifted from agency head offices to field-based staff as agencies decentralise. Yet field-based managers often have little experience in planning and managing evaluations - especially EHA.

Developing the ALNAP EHA Guide

Developing the ALNAP Evaluation of Humanitarian Action Guide was based on extensive research and consultation. The Guide was:

- Based on training materials commissioned by ALNAP in 2001. These materials were tested, updated and adapted numerous times over the years.
- Drafted in consultation with an advisory group that represented evaluation stakeholders from throughout the sector.
- Actively piloted over a period of 18 months. Over 40 humanitarian agencies used and provided feedback on the Guide. This culminated in an end-of-pilot feedback validation workshop hosted by ALNAP, with 15 of the most active piloters and EHA practitioners. This group helped the ALNAP Secretariat decide how to address higher-level feedback in the final revision process.
- Reviewed and updated during the development of the ALNAP-UNICEF Introduction to EHA e-learning course, in collaboration with UNEG and EvalPartners. The revised sections of the Guide draw heavily on the extensive review process that took place to re-design it for this online format.
- Revised based on the feedback received from the pilot process. A review of evaluation reports and other evaluative resources was used to incorporate new examples.

EHA Guide Timeline



How to use the ALNAP EHA Guide

Who is the Guide for?

This Guide is intended to help all those thinking of planning, designing and implementing evaluations and/or intending to use them, drawing on that critical mass of knowledge and particularly on a large number of Good practice examples. The Guide aims to meet an expressed need, and to help evaluation commissioners, managers, designers, implementers and users make the most of the evaluation process, and ultimately to improve the quality of EHA, and how it is used.

Things to consider when using this Guide

This Guide attempts to support high-quality evaluations that contribute to improved performance by providing the best available evidence of what is working well, what is not, and why. To achieve this, there are a few considerations that readers should bear in mind when using the Guide.

First and foremost, evaluations of humanitarian action cost money that could otherwise be used in preventing deaths or relieving suffering. This money is well spent if it leads to improvements in humanitarian action, but this will happen only if the evaluation findings are of high quality and are used (Hallam and Bonino, 2013). An evaluation should not be a 'standalone' activity, but part of an agency's efforts to be accountable for its work, and to learn. Evaluations are likely to have greatest influence in contributing to improved performance when organisations, and the sector as a whole, is committed to understanding 'how they are doing', and to learning.

Thus, EHA can do so much more than fulfil an accountability function. This does not mean that evaluations are not an important tool for accountability – they are. Accountability is about more than simply reporting. It is also about internalising the contents of these reports, and acting on them to make improvements. Learning is an important element of this. As Irene Guijt states: 'You cannot be accountable if you do not learn' (2010: 277). Very often evaluations are perceived, even feared, as a form of criticism or solely as a donor requirement. But they can, and should, play a much more constructive and useful role. By tailoring evaluations to the needs of its intended primary users, they are crucial opportunities for learning and improving programming. This Guide has been written with a strong focus on utilisation.

Many of the challenges and complexities confronting those involved in humanitarian action also affect evaluations of humanitarian action, with lack of access to affected people and affected locations being perhaps one of the most testing characteristics (Hallam and Bonino, 2013: 12). So how can we carry out sufficiently rigorous and credible evaluations in such contexts? Answering this question is at the heart of this Guide, and is addressed in every section.

The EHA Guide presents a series of structured stages for evaluations to help you achieve quality results. Each of these stages needs to be adapted to suit the context of the evaluation (considering factors such as: crisis, response, country, region, project or programme, or team) as well as the organisational context (for instance, ‘buy-in’ from leadership, learning culture). The Guide presents examples from large-scale and small-scale evaluations. It is important to determine the level of ambition of an EHA on the resources available, in terms of both internal capacity and funds. To help with this, the EHA Guide offers some insights for smaller-scale evaluations.

The Guide draws on a number of ALNAP publications and resources but does not duplicate them. While it aims to be comprehensive, there are doubtless some current topics that are not covered, and more are likely to emerge as humanitarian action evolves. One example is the evaluation of humanitarian protection activities. The ALNAP Secretariat is currently piloting guidance on evaluating protection, the final version will be published as a companion guide. Over time, there may well be other new topics that will be addressed in companion guides.

How is the Guide organised?

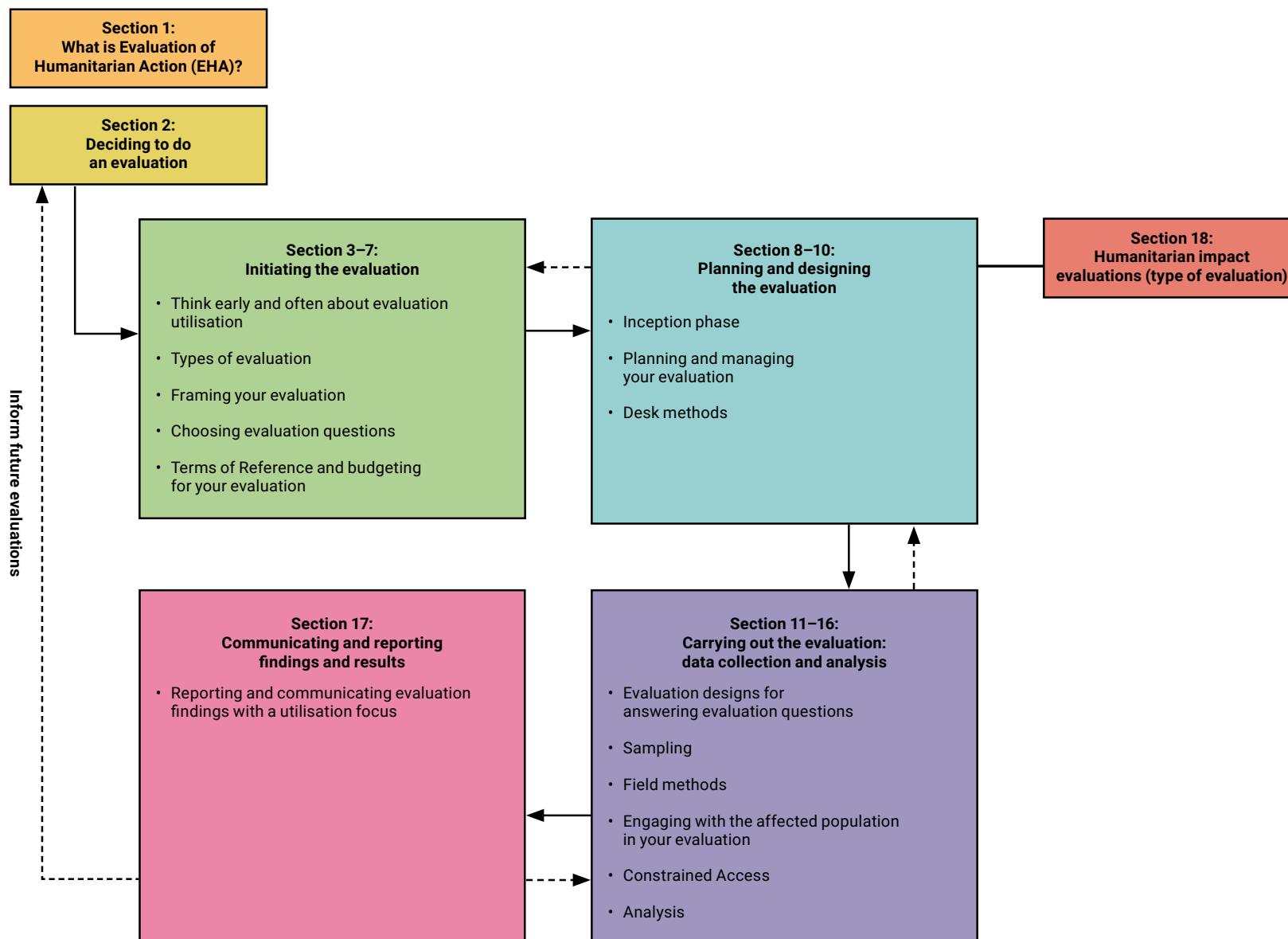
The EHA Guide is organised to reflect a typical evaluation process. It leads the user through the stages that the commissioning agency, evaluation manager and evaluator would undertake from when the decision is made to do an evaluation, all the way until the dissemination of results. The guide is organised into five chapters that reflect the important stages in the evaluation process, then broken down further into sections that walk through the specific activities in each of these stages. These stages and sections are shown in the EHA Guide map opposite. This map can serve as a navigational tool for the Guide, but also shows the evaluation process, complete with feedback loops.

Quality evaluation processes are not linear. Firstly, the EHA Guide emphasises the importance of having a utilisation-focus. Evaluation results should be feeding into future humanitarian work, by informing decision-making, facilitating learning, identifying areas for improvement and, notably, pointing to other possible research or evaluation questions. To reflect this, the evaluation stages are represented as a loose loop, rather than a timeline.

Secondly, evaluation processes have a number of feedback loops, and there may be some back and forth between activities. Within the guide, these links between sections are highlighted with cross-referencing, like this.

The bibliography and index are available at www.alnap.org/EHA.

EHA Guide map: How to navigate the EHA stages and sections



Key to design features

Icons



Key questions



Definition



In depth



Good practice example



Keep in mind



Tip

Tables



Tables in the colour of the section should be read as content of the Guide



Tables in grey are examples from practice that support the text

Diagrams

→ Direct

→ Indirect

What is Evaluation of Humanitarian Action?



1 / What is Evaluation of Humanitarian Action?

1.1 Humanitarian crises and humanitarian action

Humanitarian crises are triggered by many different causes and can have very different characteristics. Some of the most common ways in which they are categorised and described include:

1. **Cause:** for example, armed conflict; rapid-onset natural disaster such as an earthquake or flooding; slow-onset natural disaster such as drought; health crisis such as Ebola (IFRC, n.d.).
2. **Timescale and frequency:** for example, protracted, recurrent, rapid or slow onset (ALNAP, 2007).
3. **Area affected:** for example, urban, rural, mega (IFRC, 2010).
4. **State/national capacity to respond:** for example, if state capacity is limited, international agencies may perform a substituting function, but where state capacity is strong they may play a more collaborative role (Ramalingam and Mitchell, 2014).

There are various definitions of **humanitarian action** in response to crisis. One of the most comprehensive and widely used comes from the Good Humanitarian Donorship Initiative (2003).



Definition: Humanitarian action

The objectives of humanitarian action are to save lives, alleviate suffering and maintain human dignity during and in the aftermath of crises and natural disasters, as well as to prevent and strengthen preparedness for the occurrence of such situations.

The parameters of what is called humanitarian action have gradually expanded. It used to be thought of simply as saving lives, but the importance of saving livelihoods is now widely accepted as well. Humanitarian action includes both assistance and protection. While the protection of citizens is clearly the role of the state, the entire humanitarian community – not only institutions such as the specialised agencies of the United Nations, the International Committee of the

Red Cross (ICRC), and the wider Red Cross and Red Crescent Movement – are now expected to play a role. Our working definition of humanitarian action also refers to maintaining human dignity, of which being able to support oneself is an important part.



Definition: Protection

Protection comprises 'all activities aimed at obtaining full respect for the rights of the individual in accordance with the letter and the spirit of the relevant bodies of law'. (IASC, 2011)

Humanitarian action includes responding to a crisis, supporting preparedness and disaster risk reduction (DRR) before a crisis, and recovery and rehabilitation afterwards – although preparedness and recovery fall between humanitarian and long-term development activities. There is a growing recognition of the importance of addressing recovery needs in the immediate wake of a natural disaster. In conflicts and other protracted crises, it is often unclear when the emergency ends and recovery begins. In practice, both types of support are often needed and provided simultaneously.



Tip

Consider the scope of your evaluation carefully. What do you want to focus on? Are you more concerned about the first phase of the emergency response or a later phase? Do you also want to look at preparedness? Support for recovery and rehabilitation? The more focused your evaluation, the more usable your results are likely to be.

Humanitarian action should be guided by the principles of humanity, impartiality, neutrality and independence (see [Table 1](#)). These are intended to distinguish humanitarian action from other activities, including those undertaken by political and military actors, and are important for humanitarian action to be accepted by relevant actors on the ground – for example, in order to secure access to those affected by the crisis (OCHA, 2012). Many humanitarian organisations have adhered to these principles, often through expressing their commitment to the Code of Conduct for the International Red Cross and Red Crescent Movement and NGOs in Disaster Relief (IFRC and ICRC, 1994). Humanitarian principles are therefore a key reference point in evaluating humanitarian action. They should also guide how the evaluation is carried out (UNEG HEIG, 2016).

Table 1: The principles of humanitarian action

Humanity	Neutrality	Impartiality	Independence
Human suffering must be addressed wherever it is found. The purpose of humanitarian action is to protect life and health and ensure respect for human beings.	Humanitarian actors must not take sides in hostilities or engage in controversies of a political, racial, religious or ideological nature.	Humanitarian action must be carried out on the basis of need alone, giving priority to the most urgent cases of distress and making no distinctions on the basis of nationality, race, gender, religious belief, class or political opinions.	Humanitarian action must be autonomous from the political, economic, military or other objectives that any actor may hold with regard to areas where humanitarian action is being implemented.

Source: OCHA (2012: 2)

Humanitarian agencies also follow the principle of Do No Harm. In the Humanitarian Charter, this is captured in 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 (Sphere, 2011). This is further explained in Section 2: Deciding to do an evaluation. In common practice 'do no harm' has been used to mean avoiding or minimising any adverse effects of an intervention on the affected population – for instance, siting a latrine too close to a well (Christoplos and Bonino, 2016). As developed in Section 2: Deciding to do an evaluation, this principle should also be applied to how an Evaluation of Humanitarian Action (EHA) is conducted.

1.2 Evaluation of Humanitarian Action

The Development Assistance Committee of the Organisation for Economic Co-operation and Development (OECD-DAC) defines evaluation as: The systematic and objective assessment of an ongoing or completed project, programme or policy, its design, implementation and results... to determine the relevance and fulfilment of objectives, development efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors. Evaluation also refers to the process of determining the worth or significance of an activity, policy or programme. (OECD-DAC, 2002)

Drawing on this, we define EHA as:



Definition: Evaluation of Humanitarian Action (EHA)

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.

A closer look at some of the key terms in this definition reveals the following:

- **Systematic** – a planned and consistent approach, based on credible methods.
- **Objective** – stepping back from the immediacy of the humanitarian action and getting some perspective, basing findings on credible evidence.
- **Examination** – exploration or analysis to determine the worth or significance of the action.
- **Drawing lessons** to improve policy and practice and enhance accountability are the reasons for doing an evaluation.

There are two key purposes of evaluation: learning and accountability.



Definition: Learning

The process through which experience and reflection lead to changes in behaviour or the acquisition of new abilities.



Definition: Accountability

Accountability is the means through which power is used responsibly. It is a process of taking into account the views of, and being held accountable by, different stakeholders, and primarily the people affected by authority or power.

The extent to which an evaluation is truly independent depends on its purpose (see [Section 2: Deciding to do an evaluation](#)). This is more critical for accountability-oriented evaluations. It may be less achievable in learning-oriented evaluations if those doing the learning are involved in the evaluation and were responsible for implementing the humanitarian action being evaluated. Even in such cases, it is desirable to bring some level of objectivity into the process, such as by including an external facilitator or experienced and independent resource people in or leading the team.

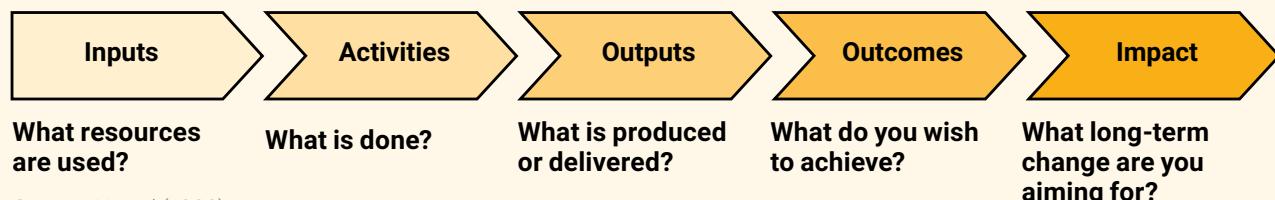


Keep in mind

All evaluations should aim to reduce bias.

Key evaluation concepts

Figure 1.1: Results chain



Source: Norad (1999)



Definition: Inputs

The financial, human and material resources used in the humanitarian action.



Definition: Outputs

The products, goods and services which result from an intervention.

Outputs are the result of inputs received and activities conducted by the actor or group of actors. An output must be fully attributable to an actor or group of actors – for example, water points provided by an aid agency in a camp of internally displaced persons (IDPs).



Definition: Outcomes

Intended or unintended changes or shifts in conditions due directly or indirectly to an intervention. They can be desired (positive) or unwanted (negative). They can encompass behaviour change (actions, relations, policies, practices) of individuals, groups, communities, organisations, institutions or other social actors.

An outcome is only partly attributable to the actor responsible for the intervention – for example, how the water from water points newly installed by an NGO is used (e.g. domestic consumption, animal consumption, or other livelihood activities such as brick-making).





Definition: Impact

Looks at the wider effects of the programme – social, economic, technical and environmental – on individuals, gender, age-groups, communities and institutions. Impacts can be intended and unintended, positive and negative, macro (sector) and micro (household, individual), short or long term.

Impacts can be positive and negative at the same time. For example, providing food aid may prevent households selling their productive assets, such as livestock, to buy food (a positive, intended impact), but it may also discourage local food production (an unintended and potentially negative impact).

Note: Throughout this Guide, ‘impact’ is used to describe the wider effects of humanitarian action, as depicted in Figure 18.1 on pg 357 (Smutylo, 2001), and in line with the OECD-DAC definition of impact.



Definition: Attribution

The ascription of a causal link between observed (or expected to be observed) changes and a specific intervention.

In complex humanitarian interventions, it is rarely possible to attribute a result to one specific cause. A food-aid agency may attribute reduced malnutrition to food distribution, but the reduction could also be caused by improved water quality, childcare practices, hygiene, health care, sanitation, and vector control, or even normal seasonal changes. As demonstrated in Figure 18.1 on pg 357, attribution becomes more difficult as you move along the results chain – it is thus harder to attribute impacts to a specific intervention than to attribute outcomes.



Definition: Contribution

Analysing contribution in evaluation refers to finding credible ways of showing that an intervention played some part in bringing about results. Contribution analysis is a kind of evaluative analysis that recognises that several causes might contribute to a result, even if individually they may not be necessary or sufficient to create impact.

It is usually much easier in EHA to assess contribution than attribution.

1.3 Monitoring and evaluation in the humanitarian response cycle

Evaluation of humanitarian projects and programmes is usually a one-off activity, undertaken at a key point in the humanitarian emergency response cycle in order to inform that cycle as well as future responses.¹ In some cases, a series of evaluations may be planned for different stages of the response cycle, as in the Darfur evaluation (Broughton et al., 2006) and the response to the earthquake in Haiti (Grunewald et al., 2010; Hidalgo and Theodate, 2012). Monitoring, on the other hand, should be ongoing throughout the implementation of humanitarian projects and programmes. Monitoring and evaluation (M&E) are complementary tools for helping determine how well an intervention is doing (IFRC, 2010:19). As the Swedish International Development Cooperation Agency (Sida) explains:

For an evaluation to be feasible, however, monitoring data may be necessary. If an intervention has not been properly monitored from start, it may not be possible to subsequently evaluate satisfactorily. Just as monitoring needs evaluation as its complement, evaluation requires support from monitoring.'
(Molund and Schill, 2007: 15).



Definition: Monitoring

A continuing function that uses systematic collection of data on specified indicators to provide management and the main stakeholders of an ongoing humanitarian intervention with indications of the extent of progress, achievement of objectives and progress in the use of allocated funds. (Based on OECD-DAC, 2002)

As a general rule, those who are implementing a programme are responsible for monitoring it in order to ensure that it remains on track.² For example, monitoring of an emergency food aid programme would be likely to focus on inputs, such as the quantities of food aid delivered, and on outputs, such as the number of people receiving food aid. Occasionally it also captures outcomes and impact – for example, whether people are selling food (i.e. how they are using it – the outcome) and the impact on market prices of food. In a cash-transfer programme, monitoring might capture how many people received the money as well as when and how much they received. This is focused on inputs and outputs. Monitoring may also extend to outcomes – for example, what people did with the cash transfer.

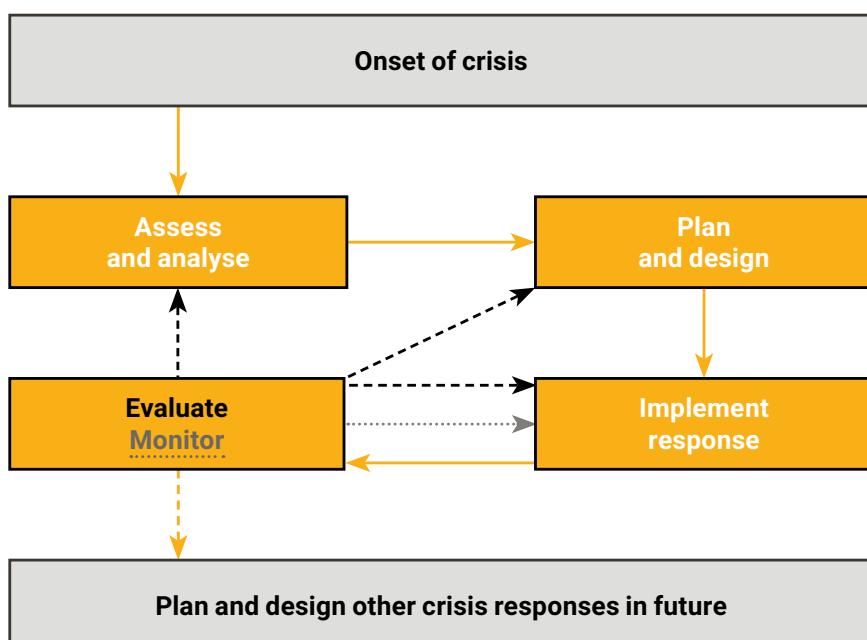
Monitoring is often weak in humanitarian action, and the lack of good monitoring data often creates problems for evaluation. An evaluation is usually conducted by people who are not involved in implementing the programme to give it more credibility.

Evaluation focuses particularly on outcomes and impact: in the example of the cash-transfer programme given above, it might look at the consequences of providing the transfer to women, and wider impacts such as the impact on market and trade activity.

Figure 1.2 shows M&E in the emergency response cycle. As Hallam and Bonino (2013) point out, the challenges confronting EHA mirror the challenges and complexities confronting humanitarian action. Some of the most common challenges facing EHA are described in Section 1.4: Common challenges faced in EHA.

Not all evaluations of humanitarian action relate to programming and the humanitarian response cycle, however. They may also focus on policy, for example (see Section 4: Types of evaluation). As is explored in Section 3: Think early and often about evaluation utilisation, it is important to consider where evaluation results fit within broader programme or organisational processes. To ensure that an evaluation is useful and is used, it should be scheduled when its results can best contribute to key decision-making moments.

Figure 1.2: Monitoring and evaluation in the emergency response cycle



1.4 Common challenges faced in EHA

EHA faces two broad sets of challenges: those that are common to all evaluations, often accentuated in humanitarian contexts, and those that relate specifically to evaluating humanitarian action, often in difficult environments. This Guide aims to address the specific challenges faced by EHA. The following list describes some of the most common challenges and their potential solutions, linking to where they are presented in the Guide.³

Evaluators should from the outset of the evaluation make clear the constraints they face and how they intend to deal with them from a risk-management perspective. The final report should indicate how these constraints affected the evaluation process and findings.

The urgency and chaos of humanitarian emergencies

A rapid humanitarian response is, by definition, planned quickly and often in extremis. Planning and monitoring documents may be scarce, objectives may be unclear, and early plans may quickly become outdated as the context changes or is clarified. So what do you use as your reference or starting point for evaluation?

Potential solution

Consider constructing a chronology of the crisis and of the humanitarian action to be evaluated. Use interviews with agency staff to identify actual or implicit objectives and how these might have changed over time. See Section 10: Desk methods.

Insecurity may mean lack of access in conflict environments

Security issues may make it difficult or impossible, especially in conflict environments, for evaluators to reach the affected population. How can you carry out an evaluation without access to the affected population?

Potential solution

Explore creative ways of carrying out the evaluation remotely. Be clear about any constraints you have faced in presenting your findings and writing up your report. Make sure not to generalise your findings to locations and populations that you have not been able to reach. See Section 15: Constrained access.

Lack of baseline data

Data may have been destroyed in the crisis, or may have become irrelevant – for example, when a large proportion of the population has been displaced. Within the short timeframe of an evaluation, how do you collect all the data you need, without a baseline reference?

Potential solution

Conduct interviews to ask crisis-affected people and local key informants about the extent to which conditions have changed and the reasons for the changes, i.e. use recall. Include similar questions in surveys. See [Section 14: Engaging with the affected population in your evaluation](#) and [Good practice example on pg 270](#).

High staff turnover

The high staff turnover in humanitarian action, especially international staff, can make it difficult for evaluators to find and interview key informants.

How do you find key staff for the period you are evaluating?

Potential solution

Invite key informants from among former or absent staff to participate in a telephone (or Skype) interview or online survey. Make greater use of national staff, who typically have lower rates of turnover – this may involve translating online surveys etc. See [Section 13: Field methods](#).

Humanitarian crisis in remote locations and with damaged infrastructure

Some humanitarian crises occur in remote locations and/or where infrastructure has been damaged, making access difficult. How do you reach those locations and how do you plan the fieldwork?

Potential solution

Carefully plan the fieldwork with someone who is familiar with the current situation, including likely travel times and access constraints, or consider letting the partner organisation plan the field visits subject to criteria set by the evaluation team. Build contingency time into your plan. See [Section 15 on Constrained access](#).

Conflicts polarise perspectives

Conflicts often intensify differences in perspective. Events, indeed the crisis itself, may be subject to widely differing interpretations. How can you do ‘objective’ evaluations in such contexts?

Potential solution

Familiarise yourself with the fault-lines in the conflict and gather as many different points of view as possible, particularly from groups on different sides in the conflict, and ensure these are represented in the final evaluation report. It can be hard to ensure the independence of the evaluation in these circumstances. If you feel the evaluation is compromised – for example, you have not been allowed access to certain contested areas – be sure to make this clear in the evaluation report and consider the implications for your findings. See [Section 14: Engaging with the affected population in your evaluation](#).

Breakdown in trust where there has been politicisation and trauma

In conflicts and other humanitarian crises that become politicised, and where there is widespread abuse, trauma and fear, trust breaks down within the affected population. How do you get in-depth and accurate information from the affected population during a short evaluation?

Potential solution

Design data-collection methods and ways of engaging with the affected population that are sensitive to trauma and fear. Consider gender-sensitive interviewer-interviewee pairing. See [Section 11: Evaluation designs for answering evaluation questions](#) and [Section 13: Field methods](#).

Humanitarian aid workers operating in pressured and stressful environments

Humanitarian organisations and their staff may be struggling to implement programmes in highly stressful environments. Making time to spend with evaluators may be a low priority, especially in [real-time evaluation \(RTE\)](#) early in the crisis. What methods are appropriate in such a context?

Potential solution

Be sensitive to the pressures that aid workers may be facing, and find ‘light’ ways to engage them. For example, short reflective-learning exercises can yield good insights if well facilitated. See [Section 2: Deciding to do an evaluation](#). Ask staff what they want to know about their intervention and in which areas they are keen to change how their organisation approaches a task.

Time pressure on the affected population

People may have little time to participate in an evaluation because their priority is their own survival. How can you ensure the affected population participates in the evaluation in an appropriate manner?

Potential solution

Avoid time-consuming methods such as focus group discussions, and opt instead for consulting members of the affected population while they are waiting for distributions or transport, and use direct observation. Consider these time pressures when designing the evaluation: for example, is it possible to delay the evaluation until after the height of the response so that people are calmer and not under so much pressure? See [Section 14: Engaging with the affected population in your evaluation](#).

Lack of clearly defined responsibility between humanitarian actors

The lack of clearly defined responsibilities among humanitarian actors can hinder accountability and attribution of impact, especially in a large-scale crisis involving many humanitarian actors. How do you attribute results and impact?

Potential solution

Focus on contribution rather than attribution. Consider carrying out a joint evaluation to explore the combined impact of a number of agencies, or of the entire international humanitarian response, rather than artificially attempting to isolate the impact of any one agency. See [Section 4: Types of evaluation](#).

Over-ambitious Terms of Reference (ToR) and limited resources

The ToR may imply an unrealistic workload, for example if it is expected that a task that should require the equivalent of two people to work over a period of a year is to be done with a budget for the equivalent of one person for four months (see [Section 7](#) which discusses ToRs). How do you manage expectations and produce a credible evaluation?

Potential solution

Use an inception report to refine the evaluation task so that it can be completed in the allotted period and budget. This is a useful way to manage expectations early on. See [Section 8: Inception phase](#).

Ethical constraints on experimental approaches

The hypothesis ‘if the drought-affected people had not received food aid, many of them would have died’ is often called a counterfactual. Ethical considerations would, of course, prohibit testing this hypothesis by withholding assistance from some people. [Section 2](#) and [Section 14](#) discuss ethical considerations, while other specific ethical considerations of EHA are presented throughout the Guide as .

Potential solution

Compare different means of assistance (for example, cash grants, vouchers, food aid). Take advantage of any natural ‘experiments’ that may have arisen, such as a group of people who received no assistance due to their isolation. See [Section 11: Evaluation designs](#) for answering evaluation questions.

Evaluating protection

Challenges here include lack of clarity about what protection is, and measuring the non-quantifiable – in particular, what did not happen? Such as violence or abuse as a result of action taken by humanitarian agencies. [See Companion Protection Guide](#) (Christoplos and Bonino, 2016) .

Despite these substantial challenges, EHA must often be undertaken quickly. EHA can thus require evaluation competencies beyond those of other evaluations (see [Section 9: Planning and managing your evaluation](#)), to ensure that rigour and credibility are not compromised (UNICEF, 2013: 3-4).

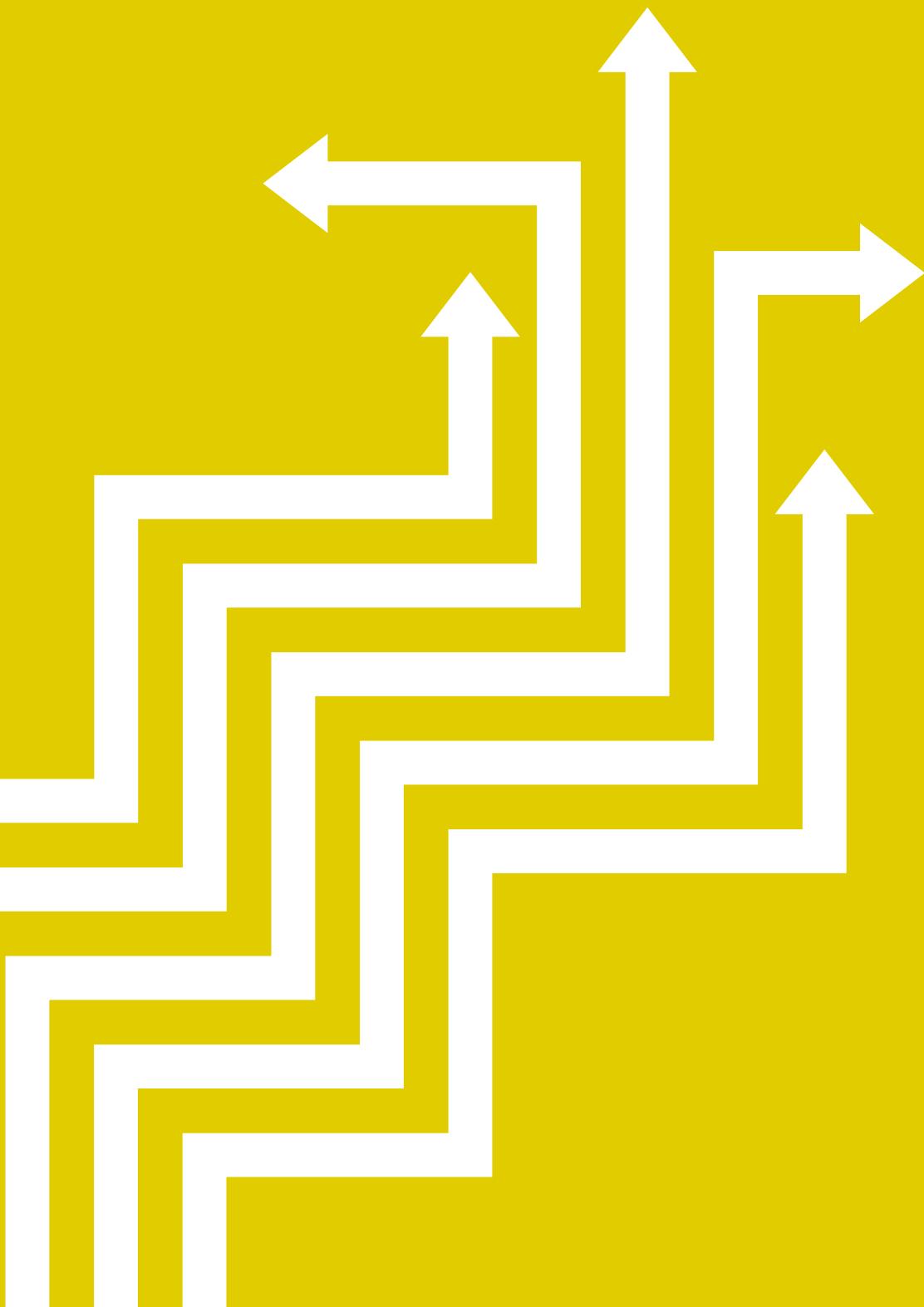
Endnotes

- 1.** This section is based on Morra Imas and Rist (2009) and Hallam (2011).
- 2.** There are external forms of monitoring, such as Citizen Report Cards (World Bank, 2004), but these are generally used in development rather than humanitarian contexts. Third-party monitoring is becoming more common in complex emergencies such as in Syria (see SAVE's work on this topic: <http://www.save.gppi.net/home/>).
- 3.** This section is adapted from Beck (2006).

Notes

Notes

Deciding to do an evaluation



2 / Deciding to do an evaluation

Sometimes commissioning an evaluation is a contractual obligation to a donor, but in other cases it might be weighed up against other options. This section presents a few options based on the purpose of this evaluative activity. If you are a programme manager or evaluation manager, this section will help you decide if an evaluation is indeed the best option, and if so when to do it. It also guides you on how to do an ‘evaluability assessment’.

2.1 When is evaluation suitable for accountability and learning purposes?

Accountability

Evaluation is one of several processes that can fulfil an organisation’s accountability requirements, in other words its responsibility to report to others, such as its board of directors, donors, or the affected population. You need to decide if it is the most appropriate and cost-effective option.

Table 2.1 summarises different types of accountability and suggests other options in addition to evaluation. For example, if financial accountability and compliance are the main concerns – whether of donors, senior management or the board of the organisation – an audit may be more appropriate. An audit reviews assurance and compliance with requirements, and is thus different from an evaluation that provides judgement of worth (IFRC, 2010). Accountability to the affected population (AAP) is now being given much greater attention, as discussed in Section 14: Engaging with the affected population in your evaluation. But this is mostly achieved through ongoing feedback and consultation during implementation (Darcy et al., 2013). Evaluations may assess AAP and should also aim to support this type of accountability (see HAP, 2010), but to consider this only once the EHA has been embarked upon is too late.

Accountability often has both external and internal dimensions. The management of an organisation may be accountable for the use of resources to donors (external) and to its board (internal), since poor use of resources may threaten organisational survival. Similarly, accountability to the affected population may reflect obligations that the management and the board have entered into.

According to Sandison (2006):

The impact of evaluation is enhanced, if not enabled, by being part of a broader menu of approaches to enhancing performance. Monitoring, for example, remains a poor cousin of evaluation and has yet to receive the same attention from decision-makers. Evaluation, in whatever form, is only one element of accountability.



Key questions

If accountability is the main purpose of your evaluation, ask yourself:

- Which type of accountability are you principally concerned about, and accountability to whom (with reference to [Table 2.1](#))?
- Is an evaluation the best way of fulfilling this?
- Should an evaluation be linked to any other accountability processes? (See [Good practice example below](#))



Good practice example: Linking accountability mechanisms: audit and evaluation

For two of its country-wide evaluations, the Food and Agriculture Organization (FAO) had completed an audit shortly before the independent evaluation (Somalia in 2012 and South Sudan in 2015). This sequencing worked well in that issues of cost-efficiency and compliance, especially related to the effectiveness and efficiency of management processes and procedures, had been covered by the audit before the evaluation. This freed the evaluation team to focus more on strategic programming issues, while drawing on some of the audit findings.

Table 2.1: Evaluation and other accountability processes

Type of accountability	Is evaluation appropriate?	Other accountability processes to consider
Accountability to the affected population	Yes, if the evaluation is designed appropriately, with full involvement of the affected population	Ongoing dialogue and consultation with the affected population, and feedback mechanisms used throughout the life of the project/programme may be more effective ways of being accountable to the affected population than through evaluation
Strategic accountability (for example, to agency's mandate and objectives)	Yes	Strategic review
Managerial accountability (for example, for use of resources within an agency)	Yes	Performance management and other management tools Performance audit
Financial accountability (for example, to donors), and compliance (for example, to senior management and the board)	Yes, especially if cost-effectiveness and efficiency are the main concerns	Audit (may be more appropriate if financial control and compliance are the main concerns)
Contractual accountability (for example, to carry out contracted tasks)	Yes, especially if there is a contractual obligation to do an evaluation	Audit; other processes specified in contract
Relational accountability (for example, to other agencies involved in an operation)	Yes, if this is included in the <u>Terms of Reference</u> for the evaluation	Institutional review
Legal accountability (for example, to local or international laws)	No	Legal (for example, labour law) compliance review

In practice, donor agencies are among the most powerful stakeholders. ‘Upwards accountability’ to donors thus tends to drive much evaluation activity (Brown and Donini, 2014). Accountability-oriented evaluations may also be used to report back to senior managers or a board of directors. The focus is on results, often looking at how well resources have been used to meet an organisation’s responsibility and mandate. For this reason, accountability-oriented evaluations usually take place halfway through or towards the end of a programme, by which time there should be results to report upon.

Learning

Learning-oriented evaluations are intended to facilitate, group, individual and/or organisational learning. Learning-oriented evaluations are valuable opportunities for learning based on real cases. They can be very effective in examining what worked, what didn’t, and how performance can be improved. The focus is often programmes that are in the initial and implementation phase, to identify what is working well and what can be improved. These may take place at any time throughout the programme cycle. In the humanitarian sector in particular, where resources for learning are scarce and there is a fairly poor record in learning (ALNAP, OECD-DAC Evaluation Network and UNEG, 2010; Patrick, 2011; Knox Clarke and Darcy, 2014), evaluations can be very useful in generating knowledge and initiating processes of organisational learning.

Evaluation is not, however, the only way to promote organisational learning, nor is it necessarily the most cost-effective. Other learning processes to consider are presented in Table 2.2. Some of these can be integrated into a learning-oriented evaluation – for example, After-Action Reviews. For more details on these and other learning-oriented methods see Section 13.7.

Table 2.2: Learning processes

Learning method	Brief description	How it can be used in evaluation
<u>After-Action Review</u>	A facilitated process for those involved in the programme to reflect on what happened, successes, challenges and learning.	This could be facilitated by the evaluators, as part of a learning-oriented evaluation, and the learning included in the evaluation report (USAID, 2006; Ramalingam, 2006: 64).
<u>Story-telling using metaphors</u>	Facilitated process whereby participants 'tell the story' of what happened and learning is drawn out of this.	Story-telling using metaphors can be used with staff, as described in <u>Good practice example on pg 260</u> , to encourage shared reflection and learning. Better Evaluation offers a number of useful references (Better Evaluation, 2014).
<u>'Most Significant Change' (MSC) technique</u>	In this participatory approach, 'significant change' stories are collected at field level. The most important of these are selected by panels of designated stakeholders as a way to capture project outcomes and impact (Davies and Dart, 2005).	MSC technique can be used with those affected by a crisis to identify changes that they attribute to the intervention (as used in the IFRC evaluation of the response to the 2007 Peruvian earthquake, Martinez, 2009).
Appreciative Inquiry	An approach focusing on solutions rather than problems, engaging key stakeholders in a reflective exercise on what worked, and what can be learned from successes.	Rarely used in EHA but Preskill and Catsambas (2006) discuss Appreciative Inquiry in detail and give examples of its use in evaluations in different sectors.

Some agencies, such as UNICEF and Tearfund, employ stand-alone learning processes in small to medium-scale emergency responses (for example, AARs) and try to encourage reflective-learning exercises. But for larger-scale emergency responses, UNICEF uses evaluation, including RTE (see Section 4: Types of evaluation).

An evaluation alone will not lead to organisational learning, although those who engage in and with the evaluation may learn as individuals. A learning organisation (Table 2.3) is most likely to benefit from an evaluation. ALNAP's work on 'Using Evaluation for a Change' (Hallam and Bonino, 2013) identifies

supportive leadership, a conducive organisational culture for evaluation and organisational structures that promote it, and the ability to secure adequate human and financial resources as fundamental to humanitarian organisations genuinely using evaluations.

Table 2.3: Evaluation and the three building blocks of a learning organisation

Building block	Distinguishing characteristics	Implications for evaluation
A supportive learning environment	Staff members feel safe enough to disagree with others, ask naive questions, own up to mistakes, and represent minority views. They recognise the value of opposing ideas, and are willing to take risks and explore the unknown.	Use evaluation methods that encourage openness – such as Appreciative Inquiry, the MSC technique, and After-Action Reviews.
Concrete learning processes	Formal processes exist for generating, collecting, interpreting, and disseminating information.	Consider asking staff members to serve on the evaluation team. Establish a formal process for involving staff in analysing findings and drawing conclusions.
Leadership that reinforces learning	The organisation's senior management demonstrate willingness to entertain alternative viewpoints; signal the importance of spending time on problem identification, knowledge transfer, and reflection; and engage in active questioning and listening.	Provide strong management support and adequate time for evaluation. Create space for feedback and the discussion of results with senior management.

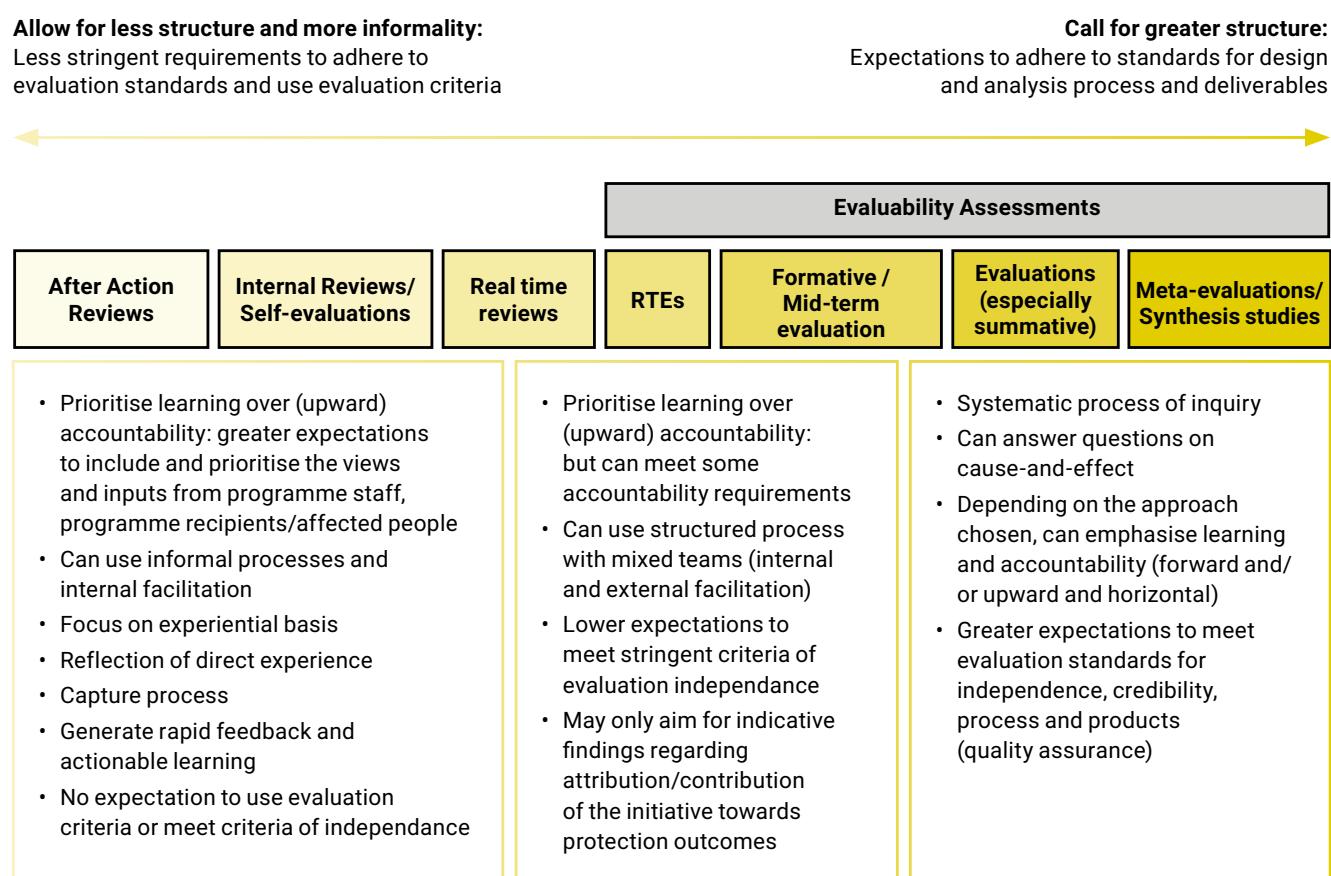
Source: Adapted from Garvin et al. (2008)

2.2 Is evaluation the right tool for the job?

Are you certain you should do an evaluation? Evaluation costs money that could otherwise be used for humanitarian action, preventing deaths or relieving suffering. Money for evaluation is well spent if it leads to improvements in humanitarian action, but it can do this only if the findings are acted upon. It is therefore important to choose the right tool for the job.

Figure 2.1 illustrates the spectrum of options from more informal reflective learning exercises on the left-hand side to more formal and structured evaluation approaches, required especially for accountability-oriented evaluations, on the right-hand side. When you have clarified the overall purpose of the evaluation, you may want to consider where your evaluative activity might sit.

Figure 2.1 Spectrum of reflective learning exercises to more formal evaluative options



Source: Christoplos and Bonino (2016) expanded from Scharbatke-Church (2011: 7)

2.3

Balancing learning and accountability

As [Figure 2.1](#) suggests, many evaluations serve in practice the dual purpose of accountability and learning. As Irene Guijt states: 'You cannot be accountable if you do not learn. And you need to know how well you live up to performance expectations in order to learn' (2010: 277), although achieving both purposes can be difficult in practice. Nevertheless, it is important to determine which is the more important or dominant purpose of the two as this will help determine the choice of evaluation design.

When planning and designing an evaluation, these are the key questions to ask:



Key questions

- Which purpose is more important – accountability or learning? Where on the spectrum between accountability and learning does it lie? It should not be in the middle, or you risk losing focus.
- If accountability is the main purpose, accountability to whom and for what?
- If learning is the main purpose, learning by whom? And how is the learning supposed to happen?

Characteristics of accountability and learning-oriented evaluations

When you have clarified the main purpose, you can design the evaluation process and methods to achieve it. Certain characteristics correspond to each evaluation purpose. For example, an accountability-oriented evaluation is likely to place greater emphasis on objectivity and independence and adopt a more investigative style, seeking to attribute responsibility for both success and failure. But this may not be conducive to learning if it makes those who need to learn feel defensive. Learning needs a safe psychological environment (Edmondson, 2014) where it is acceptable to acknowledge difficulties and admit mistakes. Thus a learning-oriented evaluation is likely to use a more facilitative style, encouraging those involved in implementation to participate and reflect. These differences explain why it can be difficult to combine both learning and accountability in the purpose of one evaluation. See [Table 2.4](#) for a description of how the evaluation approach and style might vary between accountability-oriented and learning-oriented evaluations.

Table 2.4: Accountability- and learning-oriented evaluations

Evaluation element	Accountability-oriented evaluation	Learning-oriented evaluation
Terms of Reference (ToR)	Based on input from external stakeholders as well as programme managers	Should be set by those directly involved in the programme who want or need to learn
Team membership	Generally independent external staff	Internal staff, perhaps with an external facilitator or leader, or mixed internal and external staff
Emphasis in approach	Methods of data collection and analysis (more objective)	Process of reflection and reaching conclusions (more participatory)
Methods	Mix of quantitative and qualitative methods that will provide robust evidence	Participatory methods involving those who are to learn
Management	Those responsible for accountability	Those responsible for knowledge management and learning
Management style	More directive	More facilitative
Circulation of report	Usually in the public domain	May be limited to the organisation to encourage open and honest participation

2.4 Deciding what and when to evaluate

To ensure that an evaluation is useful and is used, it should be scheduled when its results can best contribute to key decision-making moments. As Hallam and Bonino (2013) explain, an important influence on the demand for evaluation is the timing of the evaluation itself and of the presentation of its findings. It is important for those commissioning evaluations to consider this. Many potential users of evaluations complain that they often arrive too late to be of any use in decision-making (Weiss, 1983b; 1990).

A report on DFID research and evaluations notes:

The most common criticism of evaluations among the interviewees was timing: although they mark important rhythms for Country Offices, they were seen to generally take too long to be relevant to policy teams, and insufficient attention is paid to tying them into policy cycles and windows of opportunity. (Jones and Mendizabal, 2010: 11)

It is counter-productive to commission evaluations for every project or programme if these are likely to overload the organisation and if effective monitoring systems are in place. Strategically selected evaluations are more likely to meet key knowledge needs in a timely manner.

Key questions to consider for strategically selected evaluations are:



Key questions

- How can the evaluation process add value to the organisation as a whole?
- How many evaluations does the organisation have the capacity to absorb?

Sida's evaluation unit has successfully experimented with taking a strategic approach to development and humanitarian evaluations. This approach could similarly be applied only to EHAs.



Good practice example: Strategic selection of evaluations

Sida's evaluation planning cycle starts with the evaluation unit discussing with all operational units what they would like to know and how evaluation could help. A list of around 100 initial evaluation ideas is generated, from which the unit chooses 15. These are in addition to about 80 evaluations carried out each year at the operational level (Sida, 2011).

The Norwegian Refugee Council (NRC) prioritises evaluations 'on the basis of relevancy – evaluating the right things at the right time', identified through an annual evaluation plan (NRC, 2015).

It may be particularly appropriate to conduct an evaluation:

- For a programme with unknown or disputed outcomes
- For large and expensive interventions
- For pilot initiatives, or to test a new programme idea
- Where the agency has a strategic interest
- Where stakeholders are keen on an evaluation.

It is inappropriate to conduct an evaluation:

- When it is unlikely to add new knowledge
- When security issues or lack of data would undermine its credibility.

Weiss (1998) adds some important caveats about when it may be futile to conduct an evaluation because of the internal organisational dynamics.

Examples include if the evaluation manager (or similar) places strict limits on what the evaluation can study, excluding some important issues, or if there is not enough money or staff sufficiently qualified to conduct it.

Planning for evaluation early in the implementation of a humanitarian programme means that relevant monitoring systems can be established from the outset. This may also inform and influence the design of the final evaluation.

Some organisations have triggers for certain types of evaluation, especially the UN agencies. For example, the threshold for triggering an evaluation in the FAO is expenditure of \$10 million or more, and for UNICEF it is: 'major humanitarian emergencies'. These triggers relate principally to accountability, when large amounts of resources have been spent, and to some extent to when there may also be reputational risk.

Many of these aspects may be captured in an evaluation policy. An evaluation policy can also guide many of the decisions highlighted in this section about the purpose of evaluation, whether and when to do one, and some of the ethical considerations and quality assurance issues to be taken into account. If you are commissioning an evaluation, you should indicate your organisation's evaluation policy in the ToR for the evaluation. Evaluators should acquaint themselves with the evaluation policy from the outset as it may contain guidance and requirements regarding design and methods. See Good practice example, pg 52, of the NRC's succinct but comprehensive evaluation policy.



Good practice example: NRC's Evaluation Policy

NRC's evaluation policy 'clarifies why, how and when NRC uses evaluations. It also outlines standards for conduct and sets out how evaluations contribute to organisational learning and accountability' (p.3). This is an 11-page document for staff commissioning, planning and using evaluation. It sets out four principles to underpin its evaluation work: appropriateness, flexibility, transparency and participation, in recognition of the specific challenges in evaluating NRC's work with IDPs, who may have experienced traumatic events, as well as other common challenges in EHA such as constrained access and rapidly changing contexts. The policy indicates the minimum requirements for evaluation, for example that all country offices should commission at least one external evaluation every three years. It provides guidance on ensuring evaluations are useful, and indicates quality standards to be followed.

Source: NRC (2015)

Carrying out an evaluability assessment¹

It is useful to conduct an evaluability assessment before deciding to launch an evaluation, to consider some of the issues raised above. The assessment is a descriptive and analytical process intended to produce a reasoned basis for proceeding with an evaluation (Schmidt, Scanlon and Bell, 1979), and may also be used to suggest options to maximise the feasibility and usefulness of an evaluation to be commissioned at a later stage (Davies, 2013). The main purpose of the assessment is both to decide whether the evaluation can be undertaken or not, and to ensure steps are taken during implementation so that the conditions are in place to facilitate the evaluation process (UNEG, 2011: 17). If it is concluded that the evaluation should go ahead, the work done as part of the evaluability assessment can directly inform the ToR for the evaluation.

Although rare in EHA, an evaluability assessment could be carried out soon after starting a humanitarian intervention in order to improve the project design and thus its subsequent evaluability, for example by establishing strong monitoring systems and tracking defined indicators. More common in EHA are evaluability assessments to establish whether the timing is right for the evaluation, or whether it should be postponed to a later stage or until the context is more opportune, for example when there is greater security.

Rick Davies (2013) has synthesised the available guidance on evaluability assessments to produce the following outline in Table 2.5 of the main issues covered and steps commonly suggested.

Table 2.5: Generic outline for evaluability assessments

Main issues/steps	Steps commonly suggested
Define the boundaries of the project	<ul style="list-style-type: none"> • Define time period, geographical extent, and relevant stakeholders • Agree on expected outputs of the evaluability assessment
Identify the resources available	<ul style="list-style-type: none"> • Identify documents • Identify stakeholders
Identify and review documents	<ul style="list-style-type: none"> • The programme logic/theory of change/results chain and the clarity, plausibility and ownership involved • Information systems • Availability, relevance and quality of data, capacity of systems and staff to deliver what is needed • Examine implementation relative to plans
Engage with stakeholders	<ul style="list-style-type: none"> • Identify stakeholders' understandings of programme purpose, design and implementation, including areas of agreement and disagreement • Identify their expectations of an evaluation: objectives, process and use • Clarify and fill in gaps found in document review
Develop conclusions and make recommendations	<p>Evaluability assessment conclusions and recommendations should cover:</p> <ul style="list-style-type: none"> • Project logic improvements • M&E systems and capacity development (if/as required) • Evaluation questions of priority interest to stakeholders • Possible evaluation designs only if/as required*
Feedback findings and conclusions to stakeholders	

* Davies cautions against the temptations of those who commission or carry out evaluability assessments to venture into the evaluation design territory by starting to make suggestions to possible designs. Source: Davies, 2013: 16

For EHA, evaluability assessments can usefully cover four key areas:

- 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** – this is particularly important for outcome and impact evaluations that make use of theory-based designs to understand causation, and for mixed-methods designs and outcome-based approaches that look at the contribution to results in multi-actor or networked interventions (e.g. outcome mapping; outcome harvesting; RAPID Outcome Mapping Approach – ROMA).

- 3. Availability of data** or feasibility of generating data with the resources allocated, so that the evaluation can answer the chosen evaluation questions.
- 4. Conduciveness of the context** to carrying out an evaluation, for example in terms of access, logistics and security, and also the local office's ability to host the evaluation team, as well as the organisational 'climate' and leadership support for the evaluation.

The fourth area – conduciveness of the context (see [Section 15: Constrained access](#)) – is the most likely reason for deciding not to go ahead with an evaluation or to delay it. Evaluability assessments can be useful when it is not clear if it is possible to undertake a credible evaluation in the circumstances – for example, the Somalia evaluability assessment (Cosgrave, 2010). They are sometimes combined with initial programme reviews, as in the Programme Review and Evaluability Study of UNICEF's Education in Emergencies and Post-crisis Transition Programme (Barakat et al., 2010) and the preparatory review for the evaluation of the Global Education Cluster (Reid et al., 2010).

In practice, a dedicated evaluability assessment is generally used in large-scale humanitarian programmes and for multi-agency evaluations. For most EHAs, and especially for small-scale evaluations, it is usually enough for the evaluation team to undertake a 'rapid evaluability scan' during the inception phase. This can recommend improvements to the evaluation scope and focus. Once the evaluation team has been appointed it is usually too late to review the conduciveness of the context and to consider delaying or postponing the evaluation. If you are commissioning a small-scale EHA and have concerns about the conduciveness of the context, you should consider a 'rapid evaluability scan' before appointing the evaluation team.



Good practice example: 'Rapid evaluability scan'

During the inception phase of a joint evaluation commissioned by the Communicating with Disaster Affected Communities (CDAC) Network of 'Communicating with Communities' (CwC) initiatives in Nepal in response to the earthquake in 2015, it became apparent to the evaluation team that the aim and objectives set out in the [ToR](#) required a research study to test hypotheses rather than an evaluation. This was discussed with the steering group for the 'joint evaluation' during the [inception phase](#), and the consensus was that it should indeed be considered a research study, not an evaluation.²

An evaluability assessment or 'rapid evaluability scan' can thus ensure that the EHA is adapted to the context – for example, the type of crisis and type of response being evaluated, as well as the organisational context, such as the extent to which there is a learning culture and engagement by its leadership.

2.5 Ethics and EHA

Various evaluation bodies have produced guidance on evaluation ethics. For example, the American Evaluation Association's Guiding Principles for Evaluators emphasise systematic enquiry, competence, integrity, honesty, respect for people, and responsibilities for general and public welfare in their ethical guidance.³ The 2008 Ethical Guidelines produced by the United Nations Evaluation Group (UNEG) concentrate on the intentionality of evaluation, the obligations of evaluators, the obligations to participants, and the evaluation process and product.

Some humanitarian agencies have developed ethical guidelines specifically for EHA. For example, the International Federation of Red Cross and Red Crescent Societies (IFRC) refers to the principle of 'do no harm', which is further elaborated on pg 56, and also refers to Principle 5 of the Code of Conduct for the International Red Cross and Red Crescent Movement and NGOs in Disaster Relief: 'we will endeavour to respect the culture, structures and customs of the communities and countries we are working in' (IFRC, 2010).

NRC's evaluation ethics include:

- Doing no harm to informants and stakeholders
- Following international standards when interviewing children and survivors of gender-based violence
- Focusing on evaluation questions and avoiding distressing people (for example, not asking them to relive traumatic events) (NRC, 2015).

Considering these ethical guidelines may be important in carrying out an evaluability assessment. For example, can the fieldwork be conducted in such a way that informants and stakeholders are not put at risk? This is elaborated in the in depth box on pg 57. The ethics of engaging affected people in EHA is further explored in Section 14: Engaging with the affected population in your evaluation.

The evaluation should comply with the 'Do No Harm' principle

There are a number of ethical considerations for EHA, but it should always follow the principle of 'Do No Harm', just as should the humanitarian operations being evaluated. This is particularly relevant in evaluations carried out in conflict and insecure settings. The starting point should be to consider how engaging in the evaluation process might affect those taking part or being consulted.⁴

Whether an evaluation concerns a response to a conflict or natural disaster, evaluators should be aware of how it could exacerbate tensions by:

- Raising expectations that taking part in the evaluation (e.g. during data collection) will lead to more aid being provided, which could result in frustration
- Triggering heated discussions, for example between different groups in the affected population, during the data-gathering process, raising issues that reinforce tensions and divisions
- Being perceived to be involved in gathering intelligence for one of the parties in conflict
- Unwittingly presenting a biased analysis by inadequately representing the views of different stakeholders.

To avoid such risks and to conduct the evaluation in a 'conflict-sensitive' manner, evaluation managers and evaluators should from the outset consider:

- Assessing whether any steps in the evaluation process could contribute to tensions (this will need to focus on data-gathering and dissemination of the report in particular)
- For conflict settings, carrying out new (or updating existing) conflict analysis, to inform the planning and design of an EHA (see Chigas and Goldwyn, 2012)
- Revising evaluation plans in light of this analysis to ensure they do not contribute to tensions (where possible and within the organisation's mandate, trying to minimise them).

Source: Adapted from Christoplos and Bonino (2016)



In depth: What is 'Do No Harm'?

The meaning of the term 'Do No Harm' differs in the fields of humanitarian action and conflict sensitivity. 'Do No Harm' is derived from the medical principle that a medical practitioner should cause no harm to the patient. It has been adopted and adapted in other fields.

From a humanitarian perspective, 'Do No Harm' is a widely used term but is often not well defined. In the Sphere Handbook it is captured in Protection Principle 1: 'avoid exposing people to further harm as a result of your actions', which includes both violence and rights abuses and 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 the siting of a latrine too close to a well has sometimes been described as a 'harm'.

From a conflict-sensitivity perspective, 'Do No Harm' is a specific seven-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 the most widely used 'tool' for assessing conflict sensitivity.

Conflict sensitivity means ensuring that an intervention does not inadvertently contribute to conflict, and where possible, contributes to peace (within the scope of an organisation's mandate). In this definition, 'Do No Harm' relates to conflict-related risks, including many protection-related risks, since these are closely intertwined.

It is worth noting that there are many other tools in the conflict-sensitivity toolbox beyond 'Do No Harm', and there is much practice and analysis that relate to conflict sensitivity more widely.

Source: (Christoplos and Bonino, 2016)

Endnotes

1. Heavily based on Christoplos and Bonino (2016).
2. See the CDAC Network website: www.cdacnetwork.org.
3. See www.eval.org/p/cm/ld/fid=51.
4. These points are also covered in the 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.

Notes

Notes

Initiating the evaluation



3 / Think early and often about evaluation utilisation

This section aims to help evaluation commissioners and managers to consider from the outset what they need to do to ensure that the evaluation findings and recommendations will be used. It is strongly based on ALNAP's work on evaluation utilisation, which aims to improve how humanitarian agencies use and take up EHA findings (Sandison 2006; Hallam, 2011; Hallam and Bonino, 2013).

3.1 What it means to be focused on use

A utilisation-focused evaluation is done with the intended primary users in mind and for specific, declared, practical uses.



Sandison, drawing on Patton (1997) and Carlsson et al. (1999), offers the following definition of utilisation: 'an evaluation has been utilised if users with the intention and potential to act have given serious, active consideration to its findings, identifying meaningful uses according to their own interests and needs' (2006: 100-101).

It is essential to identify the intended users early on to help them decide what they want to achieve with the evaluation. Their involvement in determining the EHA purpose, objectives and scope should guide the choice of design and methods. The utilisation focus and collaboration with primary users should continue to guide the evaluation process from planning through to implementation. This way of thinking and working requires commitment and time, but ensures that the EHA should contribute to enhancing its users' knowledge and helping to bring about change and improvements in practice (Hallam and Bonino, 2013).



Keep in mind

It is important to remember that evaluation costs money, money that could otherwise be used for humanitarian action, preventing deaths or relieving suffering. Money for evaluation is well spent if it leads to improvements in humanitarian action, but evaluation can only lead to improved humanitarian action if the findings are acted upon.

3.2 How evaluations of humanitarian action are used

EHAs are generally used in two ways: for accountability (also called summative) purposes; and for learning, usually focused on a specific project (formative) but sometimes on general, system-wide knowledge (developmental). Most EHAs contain a mix of elements, though one may predominate (see [Table 3.1](#)).

Table 3.1: Evaluation purposes

Use	Questions	Examples
Summative Judging the merit or worth of a programme (for example, to fulfil its accountability to stakeholders or inform funding decisions)	Does the programme meet needs? Does it have merit? What are its outcomes?	The joint donor evaluation of humanitarian and reconstruction assistance to Afghanistan between 2001 and 2005 (DANIDA, 2005).
Formative To enhance learning (for example, to improve a programme)	What does and does not work? What are current strengths and weaknesses?	The Organisational Learning Review of Caritas Internationalis' Response to the Tsunami Emergency (Otto et al., 2006) facilitated the process of learning, emphasising openness and the participation of key stakeholders. Real-time evaluations usually have a primary learning objective (Saavedra, 2013).
Developmental To contribute to new concepts, ideas and ways of thinking, whether to an organisation or to the sector as a whole	Does the programme take real-world events and limitations into account? What are general patterns across programmes?	The Joint Rwanda Evaluation (Borton et al., 1996) introduced new ideas about the accountability of humanitarian agencies and precipitated major policy innovations such as the Sphere standards.

The ways in which EHAs are used can be further categorised as follows:

- **Instrumental use:** direct implementation of the findings and recommendations by decision-makers, for example to make changes to an ongoing programme, or to target practice. Most EHAs are designed for this type of use, whether they are accountability-oriented or learning-oriented.
- **Conceptual use:** evaluation results and conclusions are assimilated by the organisation in the form of new ideas and concepts, as described in the ‘developmental’ category in Table 3.1. This is less common and often occurs incrementally. Conceptual changes may not be attributable to a single evaluation, but evaluation syntheses may be particularly useful in triggering conceptual change.
- **Process or learning use:** engagement and participation in the evaluation itself can lead to individual learning and trigger changes in behaviour.
- **Legitimising use:** the evaluation legitimises an existing decision or understanding from an organisation or country office by providing independent and objective evidence that may be used to justify subsequent actions (see Sandison, 2006; Hallam and Bonino, 2013: 18).

Sometimes evaluations are not used effectively for various reasons:

- **Ritualistic use:** the evaluation serves a purely symbolic purpose, for example to fulfil a contractual obligation to the donor agency, but with little or no commitment on the part of the organisation to use it.
- **Misuse:** findings are suppressed, misrepresented, or distorted to serve a personal or political agenda.
- **Non-use:** findings are ignored because users find little or no value in them (owing, for example, to poorly formulated recommendations), or they are not aware of them, or the context has changed dramatically. Non-use may be a consequence of a lack of management buy-in, poor evaluation design, or an evaluation’s failure to answer its own questions or to provide compelling evidence for its conclusions (Sandison, 2006).

3.3 Identifying, understanding and engaging intended users of the evaluation

This sub-section offers tips for those commissioning and managing an evaluation, for the design and planning stage, to ensure it is utilisation-focused.

There are three key steps:

1. Identify the stakeholders of the evaluation, and within that group the primary intended users.
2. Consult the intended users about their information needs while keeping the scope of the evaluation realistic and manageable.
3. Find ways to involve the intended users throughout the evaluation.

This may also be useful to the evaluation team. If these steps have not been done by the time the evaluation team is appointed, the team should facilitate these prior to moving forward in the evaluation process. For example, during the inception phase the evaluation team could facilitate a stakeholder mapping exercise with the evaluation manager (see pg 66), and could consult the primary intended users to find out what they want from the evaluation.

Identifying the stakeholders

The first step is to identify the stakeholders and users of the evaluation. A stakeholder is an individual or organisation that has an interest or stake in the evaluation because they may be affected by decisions arising from its results or recommendations.

Not all stakeholders will necessarily use the final evaluation, however. A good example of this is the affected population. They are stakeholders because the evaluation is an opportunity to express their views, and may result in programme improvements, but they are unlikely to use its final results.

Getting to your intended users

The intended users are the main audience of the EHA. The following four questions will help you identify the primary intended users, which can also be done as a visual participatory exercise or a power-interest stakeholder analysis.

1.

Who are your stakeholders for the evaluation?

Remember these are stakeholders of the evaluation, not of the humanitarian action being evaluated.

The stakeholders may range from the evaluation unit commissioning the evaluation, to the programme implementing staff, to local authorities, to the affected population.

2.

Stakeholders with a direct interest

Those with a direct interest should be fully engaged in the evaluation. These could include funders and programme implementing staff.

2.

Stakeholders with an indirect interest

Those with an indirect interest include those who should be influenced by the evaluation or are consulted (for example, staff working for other organisations in the area), and those who should be consulted, for example the affected population.

3.

Which of the stakeholders with a direct interest are your intended users?

In most cases, if they are expected to learn from the evaluation findings, the programme implementing staff are the intended users. Funders may also be the intended users in an accountability-oriented evaluation.

4.

Who are your primary intended users?

This question is particularly important if there are several primary stakeholders with different and competing interests. Unless you clearly identify and prioritise among the intended users, the competing purposes may become unmanageable, and you may not adequately achieve any of them.



Tip: Consider how indirect stakeholders can be influenced or encouraged to ensure that the intended users do in fact make use of the evaluation findings (e.g. senior management).

Funders may have a particular interest in effectiveness and cost-efficiency of the programme, while programme staff may want to learn about what did and did not work and how they could do better, and senior managers may be interested in the wider policy implications of the evaluation's findings.

Your evaluation report should make explicit who are your primary intended users. This can be very simple or more elaborate, as the two examples below illustrate.

- ‘The evaluation’s primary users will be stakeholders at the MFA and Danida’s implementing partners, while the parliament and general public are likely to be secondary users’ (Mowjee, Fleming and Toft, 2015: 16).
- ‘There are multiple stakeholders for this evaluation. First and foremost is the Concern Worldwide office in DRC [Democratic Republic of Congo], to inform their current and future programming in Masisi. Second is Concern Worldwide Headquarters, to contribute to their learning from DRC and its potential application to other contexts. The third set of stakeholders is Concern’s donors in DRC, namely OFDA [USAID’s Office of US Foreign Disaster Assistance], which is funding the evaluation, and ECHO [the European Commission’s Humanitarian Aid and Civil Protection Department] and Irish Aid, who also finance Concern’s humanitarian activities in North Kivu. Finally, the evaluation may be useful for other aid actors interested in practical lessons and evidence on cash-based interventions’ (Bailey, 2013: 10).

For larger and more complex evaluations you might want to consider analysing the stakeholders by using a classic ‘power-interest’ matrix, as demonstrated in Figure 3.1. This matrix was constructed during the inception phase of the ‘Inter-Agency Humanitarian Evaluation of the Response to the Central African Republic’s Crisis 2013 -2015’. It helped the evaluation team understand the diverse stakeholders, to engage them without overburdening them, and to promote structured user engagement. Based on a detailed assessment of evaluation stakeholders according to power and interest (see Annex 2 of Inception report), the matrix proposed a user engagement strategy to the evaluation managers, OCHA in this case. For instance, the evaluators aimed to ‘monitor’ the Central African Republic (CAR) government and affected population, assuming their expected lower interest in the evaluation (they could still benefit from a chance to contribute views, learn about the response, measure its success, and inform future strategy), and lower power to enable the process (evaluators still needed their participation in data collection and formal acceptance) (Lawday, 2015; personal communication with Lawday November 2015).

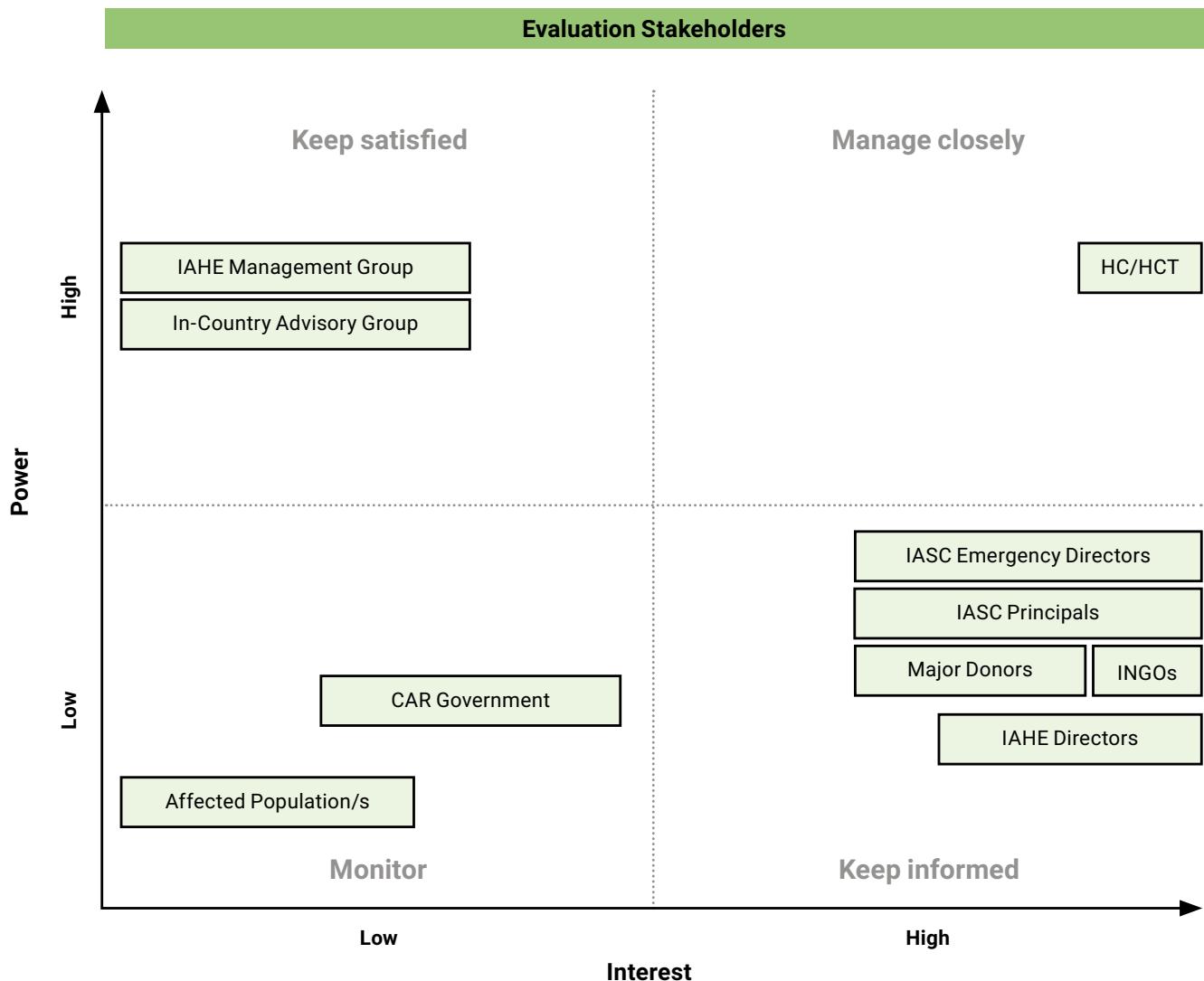
Stakeholder mapping

Different groups of stakeholders can be mapped on a series of concentric circles, distinguishing between the primary stakeholders, who are expected to be fully engaged in the evaluation, and those with an indirect interest, who may play a different role. Those with an indirect interest are further divided between those who should be influenced by the evaluation as well as consulted and those who should merely be consulted.

One way of doing this type of mapping is to write down all the different stakeholders on sticky notes, draw circles on a flip-chart, and then place each note in the appropriate circle according to the stakeholder's interests and proposed engagement in the evaluation. This can be a quick and participative exercise, and can be done for small evaluations.



Figure 3.1 Analysis of evaluation stakeholders



Source: Lawday (2015) For more on this tool, see Start and Hovland (2004: 26).

What are the users' information and evidence needs?

Once you have identified your primary intended users, ask them: What do you need to know to enable you to better decide what to do and how to do it? Usually asked by the evaluation commissioner or manager, this question helps to clarify and articulate the goals of the evaluation.

The scope of an evaluation can easily become very broad if you ask stakeholders what they would like to know, which may lead to an unmanageable 'shopping list' of issues. Encourage a utilisation focus by emphasising 'knowledge for action' by asking primary intended users questions like:
 What do you need to know that would make a difference?
 How will you use the evaluation findings?

The answers should bring you closer to the purpose of the evaluation and its scope, and give an insight into how information and evidence gathered through the evaluation might be best presented to facilitate use. For example, what should be the format and/or length of the report? What other forms of dissemination would work for the intended users? Are there key meetings or decision-making fora at which the evaluation findings could be presented, and in a way that is immediately relevant? For more discussion on these options see [Section 17: Reporting and communicating evaluation findings with a utilisation focus](#).

Involving your primary intended users

People are more likely to use an evaluation if they understand and feel ownership of the process and findings, which is in turn more likely if they have been actively involved throughout (Patton, 1997). This means there should be repeated interactions between the intended users and the evaluation managers and evaluators. This requires an investment of time and resources.



Key question

Think of your evaluation process. What steps or specific activities could be made more participative or could include users?

Here are a few ideas for involving primary intended users that are further elaborated in the guide:

- Involve potential users in compiling the [ToR](#).
- Consider including some users in the [evaluation team](#). (Agencies have different policies on this, and some prefer to involve intended users in external reference groups rather than in the evaluation team).
- Form a [reference group](#) for the evaluation that includes primary intended users.
- Hold a workshop in which primary intended users and the evaluation team jointly design the evaluation (see [Good practice example, pg 72](#)).
- Share the [draft report](#) with everyone who was interviewed and give them an opportunity to comment on it.
- Involve stakeholders in the [data analysis](#).
- Hold a workshop to [present evaluation findings](#).
- Ask the users to participate in [drafting recommendations](#).
- Ask the users to design a [dissemination strategy](#).

It is also important to consider activities that span the whole evaluation process, for example:

- Ensure that the evaluation team communicates regularly and transparently with intended users throughout the evaluation process. This will help to maintain expectations and avoid unwelcome surprises. Keeping intended users informed of the evaluation progress and changes made to the scope or methods, for instance, will facilitate use of the findings.
- Involve the primary intended users in key decisions, for example about refocusing the evaluation during the inception phase if the original ToR prove to be too ambitious for the resources available. Users being part of the reference group for the evaluation can facilitate this.

Evaluations can provoke anxiety or resistance among those whose work is being evaluated. This may be exacerbated in high-profile humanitarian crises that have attracted international media attention. In such cases, it is even more important to build a sense of ownership. This can be balanced with the need to maintain the objectivity or independence of the evaluation if the evaluators adhere to professional standards and guidelines. Objectivity does not mean disengaging from the users' needs. The evaluators need to have the necessary interpersonal skills to establish constructive relationships with key users and build trust. Frequent communication throughout the evaluation helps. Engendering a commitment to evaluation often involves promoting openness to change.



Good practice example: Involving key stakeholders throughout the evaluation

The Emergency Capacity Building Project evaluation of the response to the Yogyakarta Earthquake (Wilson et al., 2007) had a strong utilisation focus. Key aspects included:

- discussion at the outset with in-country steering committee members to find out what they wanted from the evaluation
- dialogue between the team leader and the steering committee before the team leader arrived in the country
- discussion between the evaluation team and the steering committee about evaluation methods – for example, the most appropriate interpretation of the DAC criteria.



Good practice example: Helping key stakeholders clarify their objectives

An example of stakeholders working together to determine evaluation objectives is the joint learning study regarding housing reconstruction following the Yogyakarta Earthquake in 2006. Agencies that participated in the study were Swiss Red Cross, Medair, Caritas Switzerland, Solidar, Heks and Swiss Solidarity.

Before the evaluation, intended users, in particular agency staff from the respective head offices, the evaluators, ebaix and the donor, Swiss Solidarity, attended an externally facilitated workshop to develop and identify key learning topics. Participatory methods, such as Open Space, were used to enable participants to select learning topics relevant to their current and future work in humanitarian reconstruction programmes, which then defined the areas of investigation. Participants contributed project examples for the fieldwork (Otto, 2014).

3.4 The role of the affected population in the planning stage

As mentioned above, there are usually more appropriate ways for agencies to be accountable to the affected population than through evaluations (see [Section 14: Engaging with the affected population in your evaluation](#)), and there are still few examples of consultation with the affected population about the focus of the evaluation or the questions to be asked. [The Good practice example on pg 73 provides an interesting but rare example of this being done for a Joint Humanitarian Impact Evaluation \(JHIE\).](#)



Good practice example: Consultation with affected population about the potential of JHIE

In order to assess whether there was support for a JHIE, and to explore possible approaches, consultations were carried out between February and November 2010. This included consulting the affected population in 15 communities in Bangladesh, Haiti and Sudan, as well as humanitarian actors and government representatives in each country. ‘This is perhaps the most systematic attempt to consult with the affected population during the design phase of a major evaluative exercise’ (Beck, 2011: iv). Some members of the affected population expressed their desire to participate actively in the evaluation process and to receive feedback about the results of the JHIE so that they could play a role in validating the findings. Some expressed scepticism about whether such an evaluation would make any difference: ‘As one community noted in South Sudan, they were concerned by the fact that they have participated in several evaluations, but there had been no change in the mistakes that they had identified’ (*Ibid*: iv).

Source: Beck (2011)

If there has been effective consultation and dialogue between agencies and the affected population throughout the implementation stage, then those planning and designing the evaluation may have a clear idea of the kinds of issues that should be explored. For example, if the affected population has raised protection concerns, then it may be appropriate to include an evaluation question about how effectively the agency or programme has addressed protection. For more on engaging the affected population in your evaluation, see [Section 14](#).

3.5 Factors affecting utilisation of evaluations

‘Change is not a rational process, but one that occurs through a complex variety of interactions across an organisation. As a result, harnessing the full power of EHA for change requires attention to a whole range of interconnected issues’ (Hallam and Bonino, 2013: 18-19). This sub-section presents two frameworks to help you think through the factors that influence evaluation utilisation. The first is adapted from the Overseas Development Institute (ODI)’s RAPID Context, Evidence, Links framework for assessing research and policy linkages (Crewe and Young, 2002) and can be used to think through the wider context and how to promote the use of your evaluation within

that context. The second is ALNAP's Evaluation Capacity Areas framework (Hallam and Bonino, 2013), which focuses on evaluation utilisation at an organisational level.

Context, Evidence, Links framework

Evaluation quality – The quality of the entire evaluation process, not just the final product, is a critical factor determining its usefulness and the likelihood that it will be used. This is the factor over which the evaluation team has the greatest control and has a substantial impact on credibility. As Sandison concluded: 'The picture of utilisation that emerges from [studies on the topic] is complex and often unexpected. One of the few certainties is that how and why an evaluation is carried out significantly affects the likelihood of it being used' (2006: 91).

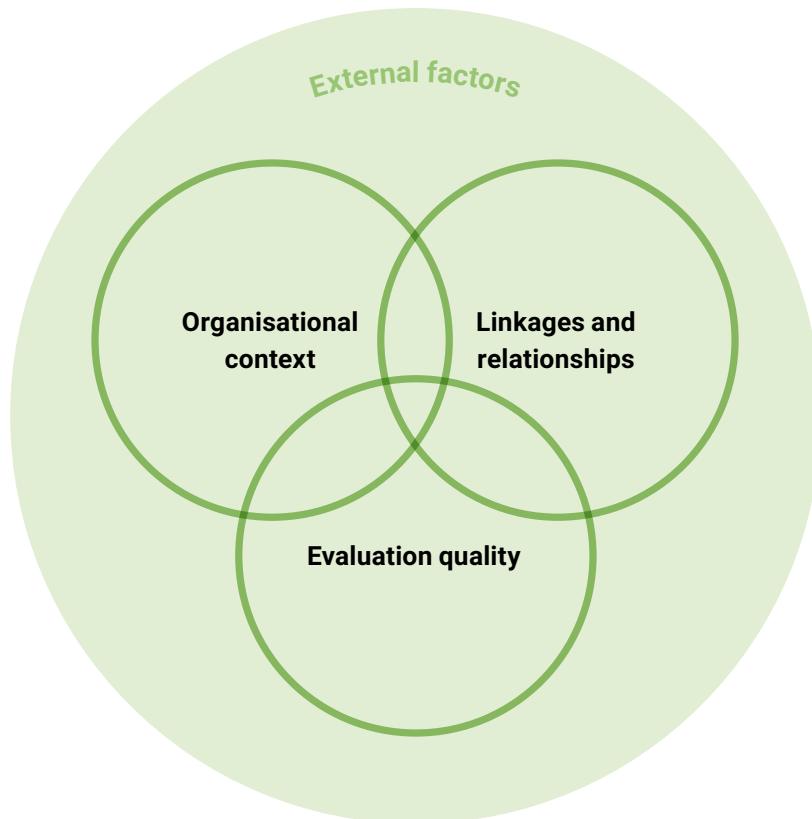
Organisational context – This encapsulates broader issues such as the organisational culture of learning and knowledge management, and the organisational structure in terms of the proximity of evaluation units to decision-makers. Evaluations are much more likely to be used in an organisation that actively seeks information on performance in order to improve programme management and implementation. This can be greatly influenced by whether key leaders support learning or place a low value on evidence-based decision-making and are inclined to be defensive in the face of evaluation.

Relationships and linkages – This includes the relationship between evaluators and evaluation users. The personal factor – for example, the presence of an evaluation champion(s) among the key stakeholders with whom it is critical for the evaluators to build and maintain a relationship – can be crucial in determining whether or not an evaluation is used. The formal position and authority of such champions may matter less than their enthusiasm and interest.

External factors – These may include the public or media, for example in a high-profile humanitarian response, asking questions about how large sums of fundraised money have been used. This was the case for the response to the 2006 Indian Ocean tsunami, for instance. External factors may also refer to the wider humanitarian policy context and whether there is openness to change. Some observers believe that the humanitarian system is 'most responsive to change under pressure when the push factors are high' (Buchanan-Smith, 2005: 98), as was the case when the multi-agency Rwanda evaluation was carried out (see Good practice example pg 75). On the other hand, the Tsunami Evaluation Coalition (TEC) synthesis concluded that lack of external pressure for change was a key reason why there had been so little improvement in performance in the ten years preceding the tsunami response (Telford and Cosgrave, 2006).

These are represented as a three-circle model in [Figure 3.2](#), with external circumstances represented by the light green shaded area around the circles. In planning your evaluation, take the time to consider how each of these circles could facilitate or hinder use of the evaluation findings, what you can do to remove any blockages and promote utilisation.

Figure 3.2 Factors affecting utilisation of an EHA



Note: For more on this framework see Start and Hovland (2004: 6).



Good practice example: The Joint Evaluation of Emergency Assistance to Rwanda

The Joint Evaluation of Emergency Assistance to Rwanda in 1996 (Borton et al., 1996) is one of the most influential evaluations of humanitarian action ever undertaken (Buchanan-Smith, 2003). How and why this happened was explored some years later, and found its success could be explained in terms of the three-circles model ([Figure 3.2](#)):

- This was a high-quality evaluation, because of the thoroughness and rigour of the work, the calibre of the team and the way it made recommendations. It had high credibility with potential users.



- In terms of organisational culture and structure, the international humanitarian system was open to change at that moment, partly because of widespread unease about the highly variable performance among humanitarian agencies.
- A number of well-placed individuals championed use of the evaluation and skilfully created or exploited alliances or networks on its behalf.
- In terms of relationships, the evaluators had strong ties to key policy-makers in the sector.
- External influences were also important. After the shock of the Rwanda crisis and the intense media coverage, policy-makers were more open to change.



Good practice example: Federal Foreign Office and the Federal Ministry for Economic Cooperation and Development's 2009 evaluation of Germany's humanitarian assistance

In 2009, the Federal Foreign Office and the Federal Ministry for Economic Cooperation and Development undertook an independent evaluation of Germany's humanitarian assistance in order to gain insights that could be used for management purposes. The process attracted much attention both because it was the first time that the two ministries had undertaken an evaluation together and because it was the first time that Germany's official humanitarian assistance had been externally evaluated.

Completed in 2011, the evaluation has had significant effects on the shape of German humanitarian assistance. This was facilitated by a number of factors. First, the evaluation coincided with a major reform of the government's humanitarian system. The newly appointed leaders were committed to the results of the evaluation and used its findings in this process. The findings informed the drafting of the very first humanitarian strategy of the Foreign Office, which launched an unprecedented quality initiative, resulting in numerous activities related to issues addressed in the evaluation.

The evaluation's recommendation to prioritise stronger partners, countries and sectors remains important in how official humanitarian assistance is set out. For instance, the Foreign Office is also investing in ways to better identify priority areas of work in terms of sectors and countries.

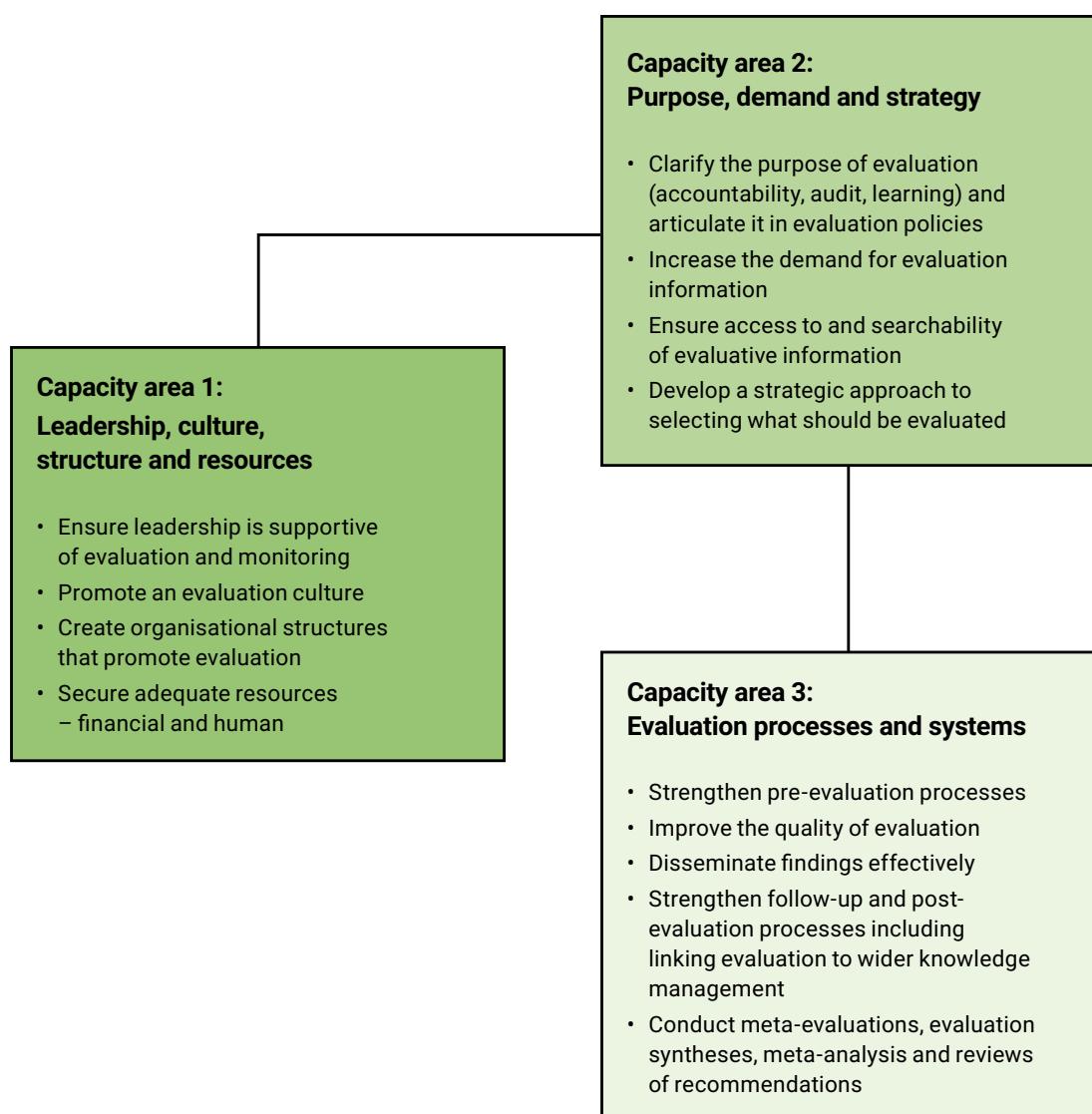
Source: Weingärtner et al. (2011)

ALNAP Evaluation Capacity Areas framework

Hallam and Bonino (2013) present a simple analytical framework to facilitate analysis and discussion of the factors that influence whether evaluations lead to impact within an organisation. As the authors explain:

Evaluative thinking should be rewarded, and should permeate throughout the organisation. Yet, people must be willing to change, and be rewarded for doing so, for improvements to occur in their programmes. They need to have the tools and the time to analyse what they do, and to explore the findings in a way that will encourage change. The organisation as a whole needs to accept that this will involve challenge to existing ways of working and critical debate. Leadership and culture are of great importance, as are formal and informal systems and reward structures' (Hallam and Bonino, 2013: 18-19).

Figure 3.3 ALNAP Evaluation Capacity Areas framework



4 / Types of evaluation

Once you have decided who the primary intended users are, what they want and need to know, and the overall purpose of the evaluation in terms of accountability and/or learning, then you must decide on the type of evaluation (IFRC, 2010: 15-16).

4.1 What type of evaluation are you going to undertake?

In identifying the type of evaluation you are going to undertake, ask the following questions:

- 1. What is the scope of the evaluation?**
Is it focused at the project, programme, sector, thematic or global level?
- 2. What level of results are you interested in?**
Do you want to evaluate process, outcomes or impact? How far down the results chain do you want to go? See Section 5: Framing your evaluation for more on this.
- 3. What is the timing of the evaluation in relation to implementation?**
Is the evaluation intended to influence programming at an early stage, for example a real-time evaluation (RTE)? Or is it a mid-term evaluation? Or is it taking place after the programme has finished (ex-post)? The timing will be largely influenced by the overall purpose of the evaluation: whether it is learning-oriented or accountability-oriented.
- 4. How many actors are involved in the evaluation?**
Is this a single-agency evaluation, a joint evaluation involving one or more actors, or a system-wide evaluation looking across the entire humanitarian system?
- 5. Who is involved in carrying out the evaluation?**
Most EHAs are carried out by teams who have not been involved in implementation, or by mixed teams of external and internal evaluators. Occasionally those who have been involved in implementing the programme undertake a self-evaluation. Evaluations may also be carried out in a participatory way and involve the affected population from start to finish, but this is extremely rare in EHA.

6. Are there any other distinguishing aspects of the evaluation?

For example, is the evaluation focused on policy or is it a meta-evaluation?

These categories are not watertight and EHAs typically belong to several categories simultaneously. For example, a programme evaluation (scope) could be an impact evaluation (level of results), conducted ex-post (timing) as a joint evaluation (involving a number of actors) and carried out independently (in terms of which stakeholders do it). The evaluation type, or combination of types, also helps determine the evaluation design, and the kind of evaluation questions that are appropriate.

The answers to the above questions should be clearly reflected in the evaluation's ToR.

4.2 Defining different evaluation types

Evaluations with different scopes



Definition: Project evaluation

Evaluation of a single humanitarian intervention with specific objectives, resources, and implementation schedule, which often exists within the framework of a broader programme.

An example of this type of evaluation is the evaluation of Tearfund and Tear Netherland's shelter projects after the 2009 earthquake in Padang, Sumatra (Goyder, 2010). Sometimes a small group of related projects by the same agency may be evaluated, as in the case of ECHO-funded projects in Pakistan implemented by the NRC (Murtaza et al., 2013).



Definition: Programme evaluation

Evaluation of a set of interventions with a unifying humanitarian objective.

An example of this type is the evaluation of Danida's assistance to IDPs in Angola (Cosgrave, 2004). Programmes sometimes emerge from a set of interventions with common objectives, as in this example, rather than being designed from the start as a coherent programme, as was the NRC's Palestine Education Programme (Shah, 2014).



Definition: Cluster evaluation

Evaluation of multiple projects within a larger programme, OR evaluation related to the UN Cluster Coordination System.

Examples of a cluster evaluation include AusAid's evaluation of a cluster of NGOs after the Pakistan earthquake (Crawford et al., 2006) and of NGO work in the Pacific (Crawford and Eagles, 2008).

Increasingly, the term cluster evaluation refers to the evaluation of the UN Cluster Coordination System, such as the initial evaluation by Stoddard et al. (2007), the Global Logistics Cluster Evaluation, the Global Food Security Cluster Evaluation (Steets et al., 2013) and the Nutrition Cluster in the response to the 2010 floods in Pakistan (Nutrition Cluster, 2011).



Definition: Partner evaluation

Evaluation of a set of interventions implemented by a single partner.

The Norad evaluation of the work of the NRC is an example of this type (Ternström et al., 2013), as are the ECHO evaluations of its partnerships, such as, the UN relief and works agency. (Grünewald and de Geoffroy, 2009).



Definition: Sector evaluation

Evaluation of a group of interventions in a sector, all of which contribute to the achievement of a specific humanitarian goal. The evaluation can cover part of a country, one country, or multiple countries (UNICEF, 2013).

An example of a sector evaluation is an evaluation of the water, sanitation, and hygiene sector in DRC (van der Wijk et al., 2010). For some agencies, sector evaluations have now been overtaken by the second type of cluster evaluation described above, such as the IFRC evaluation of the shelter and non-food item cluster in the 2010-2011 response to tropical cyclone Giri in Myanmar (Hedlund, 2011).



Definition: Thematic evaluation

An evaluation of a selection of interventions that all address a specific humanitarian priority that cuts across countries, regions, and possibly agencies and sectors.

An example of a thematic evaluation is the evaluation of the role of food aid for refugees in protracted displacement (Cantelli et al., 2012) or the evaluation of needs assessment after the December 2004 Indian Ocean tsunami (de Ville de Goyet and Morinière, 2006). This was one of five thematic studies for the Tsunami Evaluation Coalition. The UN High Commissioner for Refugees (UNHCR) review of refugee education is an example of a single-sector thematic evaluation (Dryden-Peterson, 2011).



Definition: Humanitarian portfolio evaluation

An evaluation of the whole humanitarian portfolio of an agency.

Such evaluations are conducted by donor agencies to review their entire humanitarian portfolio. Examples include the evaluation of Finnish humanitarian action (Telford, 2005), of Swiss humanitarian action (de Ville de Goyet et al., 2011) and of Danish humanitarian strategy (Mowjee et al., 2015).

Evaluations at different levels of results



Definition: Impact evaluation

An evaluation that focuses on the wider effects of the humanitarian programme, including intended and unintended impact, positive and negative impact, macro (sector) and micro (household, individual) impact.

A good example of an impact evaluation is the evaluation of the assistance provided by the US State Department's Bureau of Population, Refugees and Migration to Burundian refugees (Telyukov et al., 2009). Another excellent example is the impact assessment of community-driven reconstruction in Lofa County in Liberia (Fearon et al., 2008). Impact evaluations are discussed further in [Section 18](#).





Definition: Process evaluation

An evaluation that focuses on the processes by which inputs are converted into outputs; may also examine the intervention as a whole.

Process evaluations do not seek to assess outcomes or impact but rather the processes behind humanitarian action. For example, UNHCR evaluated its relationship to and use of the Central Emergency Response Fund, to assess how that had affected the funding of UNHCR programmes and contributed to funding trends over a five-year period (Featherstone, 2014).



Definition: Normative evaluation

An evaluation that compares what is being implemented with what was planned or with specific standards.

Normative evaluations are relatively rare in EHA. They evaluate against a set of normative standards and should not be confused with evaluations of normative work, which evaluate the work to establish norms and standards (UNEG, 2013). An example of a normative evaluation was the Disasters Emergency Committee (DEC) evaluation of the response to the 2002-2003 Southern Africa crisis (Cosgrave et al., 2004) against the Code of Conduct for the ICRC and NGOs in Disaster Relief (SCHR and IFRC, 1994).

Evaluations with different timing



Definition: Real-time evaluation (RTE)

An evaluation of an ongoing humanitarian operation as it unfolds.

Examples include the IFRC evaluation of the response to tropical cyclone Haiyan in the Philippines (Greenhaigh et al., 2014), and UNHCR's evaluation of its response to the Syria Crisis (Crisp et al., 2013).



Definition: Mid-term evaluation

An evaluation performed towards the middle of an intervention.

Examples include the mid-term review of the Syria Needs Assessment Project (Featherstone, 2013) and the mid-term evaluation of the European Commission Directorate General for Humanitarian Aid and Civil Protection's Regional Drought Decision for the Greater Horn of Africa (Wilding et al., 2009).



Definition: Ex-post evaluation

An evaluation performed after an intervention has been completed.

An example is CARE's evaluation of its tsunami relief response in two districts of Sri Lanka (Bhattacharjee et al., 2007).



Definition: Ongoing evaluation

A series of evaluations designed to run throughout an intervention.

Two examples of ongoing evaluation are the pair of IASC evaluations of the response to the Haiti earthquake at three and 20 months after the earthquake (Grünewald et al., 2010; Hidalgo and Théodate, 2012), and a UNHCR evaluation that ran for six years after the end of an intervention (Skran, 2012).



Definition: Ex-ante evaluation

An evaluation performed before an intervention begins.

Such evaluations, rare in the humanitarian sector, are based on the lessons learned from similar operations. One example is the Review Concerning the Establishment of a European Voluntary Humanitarian Aid Corps (Bruaene et al., 2010).

Evaluations involving a different number of actors



Definition: Single-agency evaluation

An evaluation carried out by the agency that implemented the intervention.

Examples of this common type of EHA include CARE's evaluation of its Pakistan earthquake response (Kirkby et al., 2006) and Oxfam GB's evaluation of cash transfers in Sudan (Bush and Ati, 2007).



Definition: Joint evaluation

An evaluation carried out by two or more agencies, evaluating the work of two or more agencies.

Joint evaluations might involve a donor and a recipient agency, multiple agencies with similar missions, or different actors working in the same sector. Examples include joint evaluations carried out by NGO partners in an emergency capacity-building project after the Yogyakarta earthquake (Wilson et al., 2007) and the CARE and Save the Children evaluation of their Haiti response (O'Hagen et al., 2011). Donor joint evaluations include the joint follow-up evaluation on linking relief rehabilitation and development (Brusset et al., 2013). UN examples include the joint WFP/UNHCR evaluation of the contribution of food aid to refugees in protracted displacement (Cantelli et al., 2012).



Definition: System-wide evaluation

An evaluation of the international humanitarian system's response to a humanitarian crisis, open to all actors in the system.

Examples of this are the Tsunami Evaluation Coalition's evaluation of the international response to the 2004 Indian Ocean tsunami (Telford et al., 2006), and the multi-agency evaluation of the response to the Rwanda crisis (Borton et al., 1996).

Evaluations carried out by different stakeholders



Definition: External or independent evaluations

An evaluation carried out by evaluators who are outside the implementing team.

External or independent evaluations are usually a requirement of accountability-oriented evaluations because they bring an objective perspective.



Definition: Self-evaluation

An evaluation carried out by those who design and deliver an intervention, in other words an internal evaluation.

Most self-evaluations are focused on lesson-learning rather than accountability, and often produce internal documents. One published example is the evaluation of Medair's response to the 2004 Indian Ocean tsunami in Sri Lanka, which was conducted by a member of the initial team in Sri Lanka who later returned to carry out the evaluation (Lee, 2005).



Definition: Participatory evaluation

An evaluation in which stakeholders, including the affected population, work together to design, carry out, and interpret an evaluation (also see participatory design in Section 11).

Participatory evaluations are very rare in the humanitarian sector. King (2005) highlights a common misconception that simply involving programme staff, programme recipients, or programme participants in any way – most commonly collecting data from them – makes it a ‘participatory’ evaluation. Mere contact between an evaluator and programme recipients or participants, he notes, is not sufficient. What makes an evaluation ‘participatory’ is the role that stakeholders, including programme participants, play and the degree of their involvement throughout the evaluation process (Alexander and Bonino, 2014).

Other evaluation types



Definition: Technology evaluation

An evaluation of a specific technique or technology.

These evaluations are not common. Examples include the review of the use of the British Red Cross Mass Sanitation Module after the 2010 Haiti earthquake (Fortune and Rasal, 2010) and of the shelter kit used after the Nigerian Floods of 2012 (Bravo et al., 2014). Such evaluations may also cover innovative approaches, such as the use of Community-based Management of Acute Malnutrition in Nepal (Guerrero, 2010).



Definition: Institutional evaluation

Evaluation of the internal dynamics of implementing organisations, their policy instruments, service delivery mechanisms and management practices, and the linkages among these.

Examples include the two evaluations of UNICEF's programme for humanitarian capacity-building, funded by the UK's DFID (Bhattacharjee et al., 2010; Brown et al., 2005), and the Independent Review of UNICEF's Operational Response to the January 2010 Earthquake in Haiti (Bhattacharjee et al., 2011).



Definition: Policy evaluation

An evaluation that examines the understandings, beliefs, and assumptions that make individual projects possible as well as desirable. It may evaluate both the efficacy of the policy itself and how that policy has been implemented.

Policy evaluations are often quite wide-ranging and tend to be reviews rather than evaluations. An example is UNHCR's review of its urban refugee policy in New Delhi (Morand and Crisp, 2013). Sometimes such policy reviews cover the whole humanitarian sector, such as the UN's Humanitarian Response Review (Adinolfi et al., 2005), and DFID's Humanitarian Emergency Response Review (Ashdown, 2011).



Definition: Meta-evaluation

An evaluation designed to aggregate findings from multiple evaluations, or an evaluation of the quality of one or more evaluations.

One example is ALNAP's meta-evaluation of joint evaluations (Beck and Buchanan-Smith, 2008). Scriven (2011) provides a checklist for meta-evaluations of the second type and the meta-evaluation of the quality of US evaluations of foreign assistance (Kumar and Eriksson, 2012) is an example, as are the reports of UNICEF's Global Evaluation Reports Oversight System (Universalia, 2013). UNICEF has published the standards by which evaluations are measured (UNICEF Evaluation Office, 2013). For more discussion on meta-evaluation, see [Section 17.6](#).

4.3 Real-time evaluations (RTEs)



In depth: Real-time evaluations (RTEs)

Primary objective: to provide feedback in a participatory way, during fieldwork, to those implementing and managing the humanitarian response (Cosgrave et al., 2009b).

Use: There has been a huge growth in the number of RTEs in the humanitarian sector in recent years as agencies attempt to strengthen evaluations' contribution to learning, especially the immediate learning during implementation.

Advantages of RTEs:

1. **Timeliness** – RTEs are carried out when key operational and policy decisions are being taken. They can highlight important issues that have been overlooked in the rush to meet immediate needs. For example, Catholic Relief Services' 2010 Pakistan RTE took place only nine weeks after the start of the agency's response and included a one-day reflection with staff and partners at different locations, during which immediate action plans were drawn up (Hagens and Ishida, 2010).



2. Facilitate interaction between staff and with evaluators –

RTEs involve sustained dialogue with staff, both in the field and at the head office, and can provide a channel for communication between both sets of staff that bypasses bureaucratic obstacles.

For example, the 2010 RTE of the Inter-Agency Standing Committee's response to floods in Pakistan included a second visit in which findings, conclusions, and recommendations were discussed with stakeholders in a participatory fashion (Polastro et al., 2011b).

3. Perspective – An RTE team can approach an emergency from various angles, talking with staff at all levels, in different countries, with the affected population, and with partner agencies and government officials. This can provide a view of the operation that is less influenced by solving day-to-day problems.

Calibre of evaluators: Team members must have sufficient operational experience to understand the most likely outputs, outcomes and impacts of current actions.

Challenges and potential solutions:

- Completing the RTE report before leaving the field, unlike standard evaluations.
Factor writing time into the fieldwork schedule.
- RTEs' primary stakeholders are the field team and those managing the operation at the head office, most of whom are working under considerable pressure with little time to read a conventional evaluation report.
Include time for debriefings in the field in the evaluation plan, e.g. one mid-fieldwork to raise emerging issues with the field team, and a final debriefing.

4.4 Joint evaluations



In depth: Advantages of joint evaluations¹

Primary objective: evaluation across agencies to understand their collective impact, and/or promote learning across agencies, and/or evaluate aspects of the relationship between agencies. The stated purpose of joint EHAs is usually both accountability and learning, but in practice participating agencies, for example in the Tsunami Evaluation Coalition, have often found them to be most useful for learning.

Use: Joint evaluations are becoming more common in humanitarian aid, reflecting a wider trend towards aid ‘jointness’.¹ Joint evaluations can be found in multi-donor funding channels, for example the evaluation of support for peace-building in South Sudan (Bennett et al., 2010); in NGO alliances, for example the multi-NGO evaluation of the Yogyakarta earthquake response (Wilson et al., 2007) and the Action by Churches Together (ACT Alliance) evaluation of the Haiti earthquake response (McGearty et al., 2012); and in the concept of the UN’s transformative agenda as in the IASC RTEs (Cosgrave et al., 2010; Grünwald et al., 2010; Polastro et al., 2011a). They can also be found in system-wide evaluations, as in the case of the 2004 Indian Ocean tsunami (Telford et al., 2006).

Advantages:²

1. **Focus on the big picture** – Joint evaluations encourage a focus on the big picture, on strategic issues, and with more of a policy focus.
2. **Well-suited to evaluating impact** – Joint evaluations avoid attributing impact to any one agency and can be used to evaluate the collective impact of the humanitarian response.
3. **Stronger record in consulting the affected population** – With greater resources than most single-agency evaluations, joint evaluations have a stronger record in consulting the affected population (See Beck and Buchanan-Smith, 2008).
4. **Utilisation** – Joint evaluations often have greater credibility and can be useful for advocating change. They can also play a valuable learning function across agencies, helping participating agencies to understand each other’s approaches and to exchange good practice.



5. **Managerial and financial** – Joint evaluations pool evaluation resources and capacity, and can reduce transaction costs if they reduce the number of single-agency evaluations consulting the same stakeholders, especially in the affected area.

Calibre of evaluators: Complex joint evaluations require a combination of technical, political, and interpersonal skills, especially in the team leader. The pool of evaluators with such skills is small. Planning and tendering early will help to secure qualified candidates.

Challenges and potential solutions:

- In view of the number of stakeholders involved in a joint evaluation, negotiating the ToR is likely to take much longer than a single-agency evaluation.
It is important to allow extra time for negotiating the ToR. Strong chairing and facilitation skills are essential for this initial planning stage in order to manage the negotiations and to ensure that the final ToR are clear and concise, rather than becoming a long list of unprioritised questions.
- The complexity of management structures and high transaction costs associated with joint evaluations can be a barrier to the involvement of smaller organisations, for example smaller NGOs, including national NGOs.
Identifying engagement options for smaller NGOs that are less demanding than those for larger stakeholders – for example, only during key moments in the evaluation and/or with smaller contributions of funds than larger agencies – will help to ensure their involvement.

5 / Framing your evaluation

This section looks at how to frame your evaluation. Framing refers to establishing or identifying a structure that will help to translate the overarching evaluation questions into specific questions. This translation can be based either on assumptions about how the intervention is meant to achieve the desired change (causal questions) or in terms of standards for such interventions (normative questions).

The choice of causative or normative framing will depend on the context. Sometimes it is more appropriate to focus attention on normative questions, for instance when it is not possible to establish attribution and it is difficult to assess contribution.

Essentially, framing is about enabling the evaluators to identify and concentrate on more significant issues rather than on those that are less important for the intervention to achieve its objective.

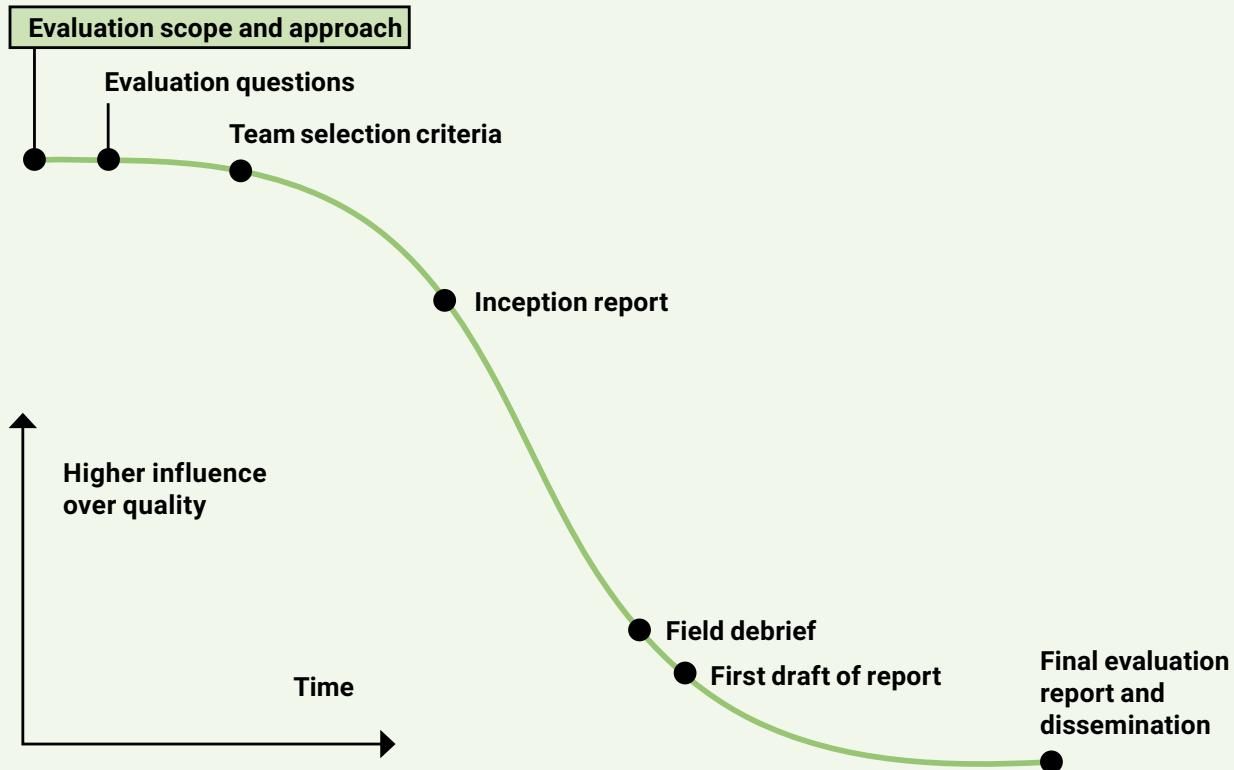
How can you ensure the quality of your evaluation? A carefully thought-out and structured approach helps. There are a number of key quality control points in the evaluation. Too often the main emphasis on quality control is on the evaluation report. It is much better to focus on quality control earlier on in the process.

Throughout the guide we will highlight the points in the evaluation process at which evaluation managers can best assess the quality of the deliverables and make adjustments to the process accordingly.

These key quality control moments are presented along an S-curve as the level of influence an evaluation manager has over quality goes down as the evaluation process progresses.



Quality checkpoint: Evaluation scope and approach



The first key quality control point you should have is around framing your evaluation. This includes determining the scope and approach of your evaluation.

5.1 The programme logic

At the design stage of interventions, the planned actions should ideally be based on some theory as to how it will achieve the desired end result. Explicit theories are usually presented as some sort of logic model.



Definition: Logic model

A table or diagram presenting the programme theory (the way in which inputs are expected to contribute to the overall goal) for an intervention.

Logic models range from the logical frameworks to complex diagrams and theories of change (for a good overview, see Funnell and Rogers, 2011).

Conceptual frameworks

Conceptual frameworks are logic models, usually based on extensive research, that illustrate the key factors in a particular issue. This research base means that conceptual frameworks are much more robust than the other logic models presented here. While other logic models may be based on assumed causal linkages, conceptual frameworks are based on linkages established by research.

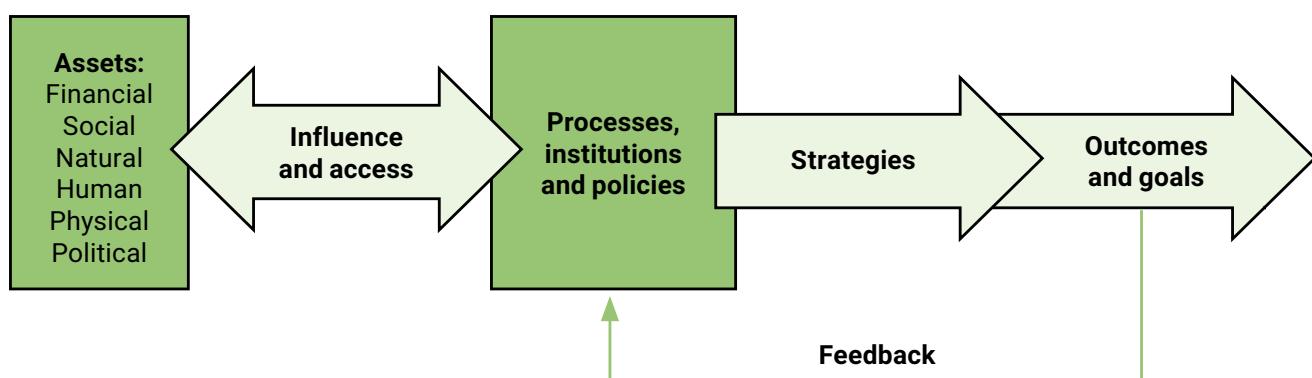


Tip

Check if there is a conceptual framework covering the theme of your evaluation. Conceptual frameworks are very useful tools for ensuring that no key area is overlooked.

The sustainable livelihoods framework looks at rural livelihoods. Lautze and Raven-Roberts (2003: 10) proposed the following variant of the framework for complex humanitarian emergencies.

Figure 5.1: Livelihoods Analytics Tool



Source: Lautze and Raven-Roberts (2003: 10)

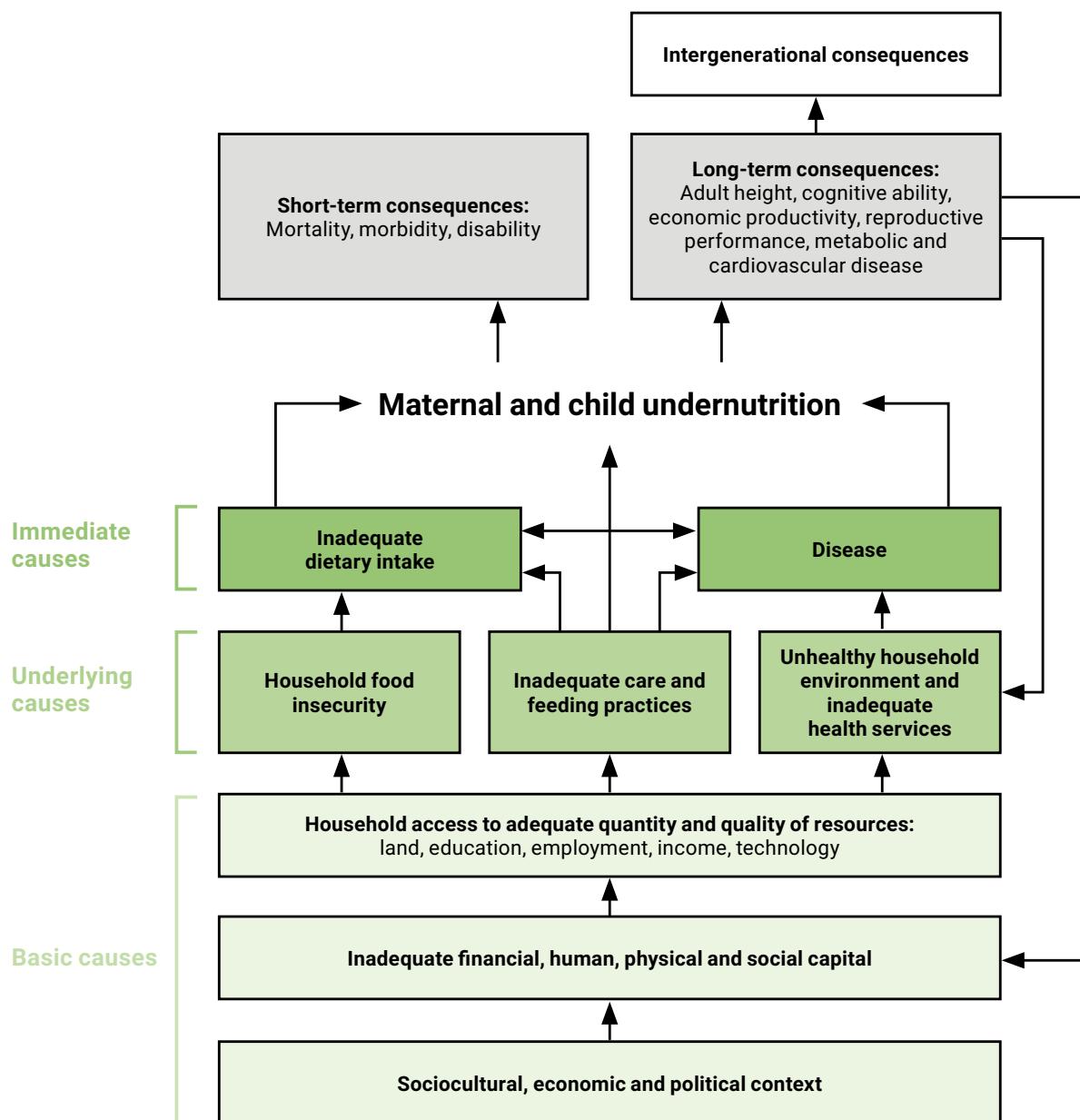
The evaluation team used this framework in evaluating the FAO's programme in Somalia (Buchanan-Smith et al., 2013). This meant that while the team looked at how FAO has supported livelihoods through assets, particularly for instance in supporting livestock holdings with distribution of drugs and fodder, they also looked at its action on the policy context. One policy issue was the ban placed on livestock from the region by countries in the Middle East, which had a huge influence on livelihoods.

Another conceptual model, which can be particularly useful in evaluating humanitarian action, is UNICEF's conceptual model of malnutrition, especially if the focus is on food aid programmes or medical interventions.

The evaluation of Oxfam GB's Emergency Food Security and Urban Livelihoods programme (Macauslan and Phelps, 2012) used this conceptual model as a base for the causal framework for malnutrition and its effect on livelihoods.

Even if a conceptual framework was not used in developing an intervention, evaluators can still make use of a conceptual framework as it contains proven causal linkages. It may also hint at relevant evaluation questions. For example, in evaluating a nutrition programme, poor results would not be surprising if childcare or health issues had not been addressed.

Figure 5.2: UNICEF Conceptual framework of the determinants of child undernutrition



Source: UNICEF (2013).

Logical frameworks

Logical frameworks are one of the simplest forms of logic model, with activities linked directly to outputs. The concept was first introduced in the development field by USAID in 1970 (Rosenberg et al., 1970) and was then more broadly adopted. Table 5.1 presents the key elements of the logical framework. Agencies may use different names and precise definitions of the elements of a results chain.

Table 5.1: Sample logical framework

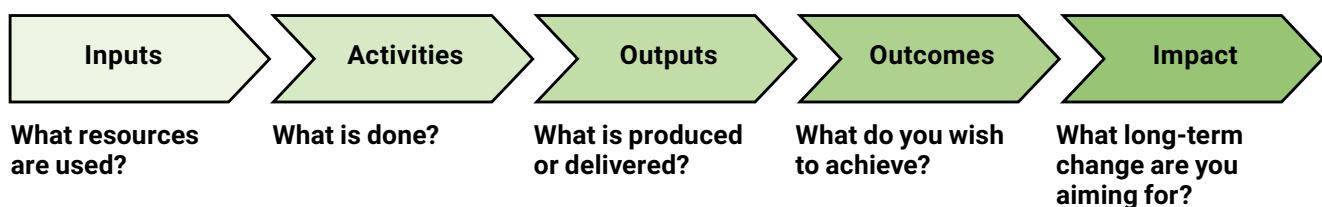
Results hierarchy	Indicators	Assumptions
Goal The highest level objective towards which the project is expected to contribute	Indicators measuring progress towards the goal	Assumptions relating to the sustainability of the goal
Purpose The effect that is expected to be achieved as a result of the project	Indicators measuring progress towards the purpose	Assumptions related to the achievement of the goal given that the purpose is achieved
Outputs The results that the project management should be able to guarantee	Indicators measuring the extent to which outputs are produced	Assumptions related to the achievement of the purpose once the outputs are in place
Activities Actions undertaken to produce the outputs	Inputs, such as goods and services necessary to undertake the activities	Assumptions related to the production of outputs

Source: adapted from Norad (1999)

Results chain

The logical framework approach has been criticised as being too rigid, especially in fast-changing humanitarian contexts. In addition, the results hierarchy can be confusing, in that the goal of a project may be the same as the purpose of a programme of which the project is a component.

Norad eventually moved towards a more flexible approach using a results chain, as shown below.

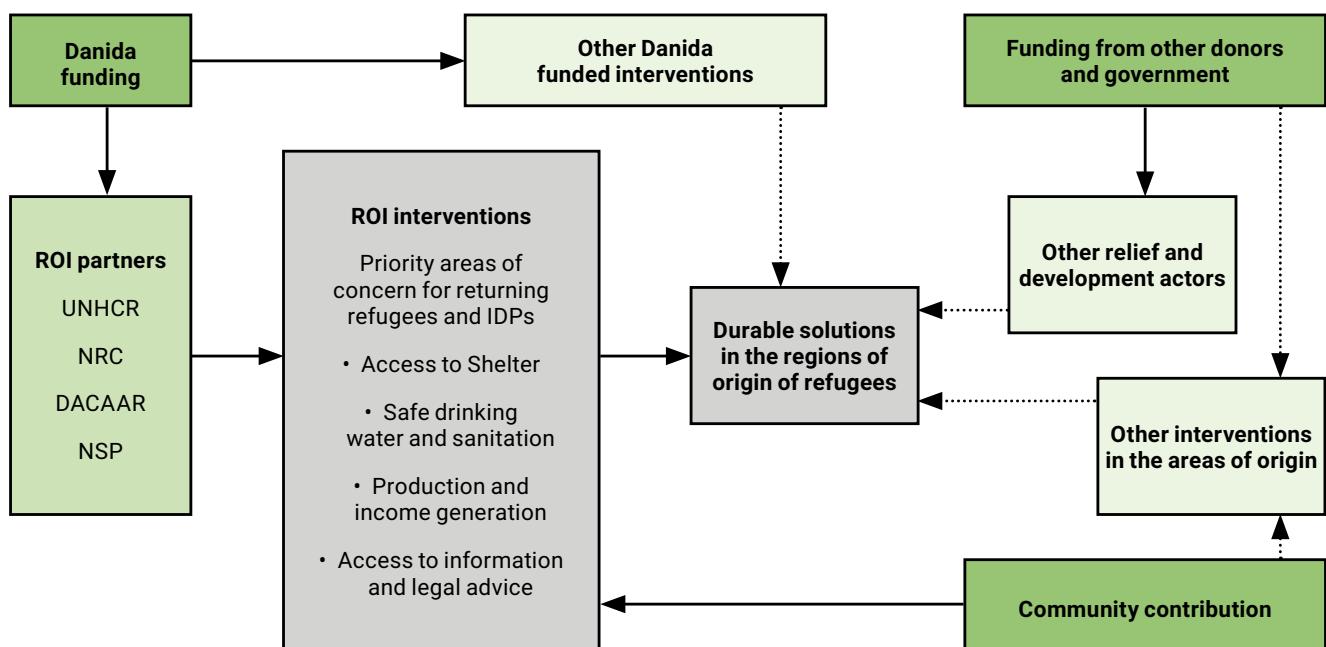


The results chain may have different names, including causal pathways, but the central idea is to show how inputs are intended to lead to the intended long-term change.

Graphical logic models

Models can also be presented graphically, as in the following example for an evaluation of Danish Region of Origin Initiative (ROI) in Afghanistan. Such graphic presentation can help to highlight linkages to others' actions.

Figure 5.3: Evaluation of Danish assistance to Afghanistan



The Kellogg Foundation has published a guide to developing logic models (WK Kellogg Foundation, 2004), which is framed around quite simple results chains.



Tip

Models are convenient simplifications of complex reality. This simplification is their greatest strength, as it allows the manipulation of complex concepts and situations – but also their greatest weakness, as it may lead you to overlook a key factor. If you use a model in a validation, examine the extent to which the model fits the real situation you are examining and modify it as necessary.

Software packages such as DoView can facilitate the drawing of simple pipeline or outcome-chain logic models.

5.2

Theories of change



Definition: Theory of change

A theory of change is a description of the central mechanism by which change comes about for individuals, groups and communities.

A theory of change (ToC) expresses the cause and effect linkages or chains of linkages between the intervention and the desired end-result. It is an increasingly popular approach, and some donors have adopted ToC in place of logic models (although some require both). Vogel notes that they are seen as a ‘more realistic and flexible thinking tool than current logical framework approaches’ (2012: 9). Keystone’s ToC guidance compares ToC and the logical framework (2009, 30), while Rogers (2012: 7) lists logical frameworks as one of four ways to represent a ToC.

Theories of change are most useful in complex environments because they make it possible to break up the results chain into a series of causal linkages that can be tested by the evaluation.

There is even less agreement about ToC than about other forms of logic model, however. Stein and Valters note (2012: 5) that there ‘is a basic problem that different organisations are using the term ToC to mean very different things’.

A ToC can be represented as a statement (for examples see OECD-DAC, 2012: 85), or a table or diagram. The following example from the inter-agency evaluation of food assistance for protracted refugees in Bangladesh shows a tabular presentation of a ToC as a ‘simplified logic model’ (Nielsen, 2012: 4).

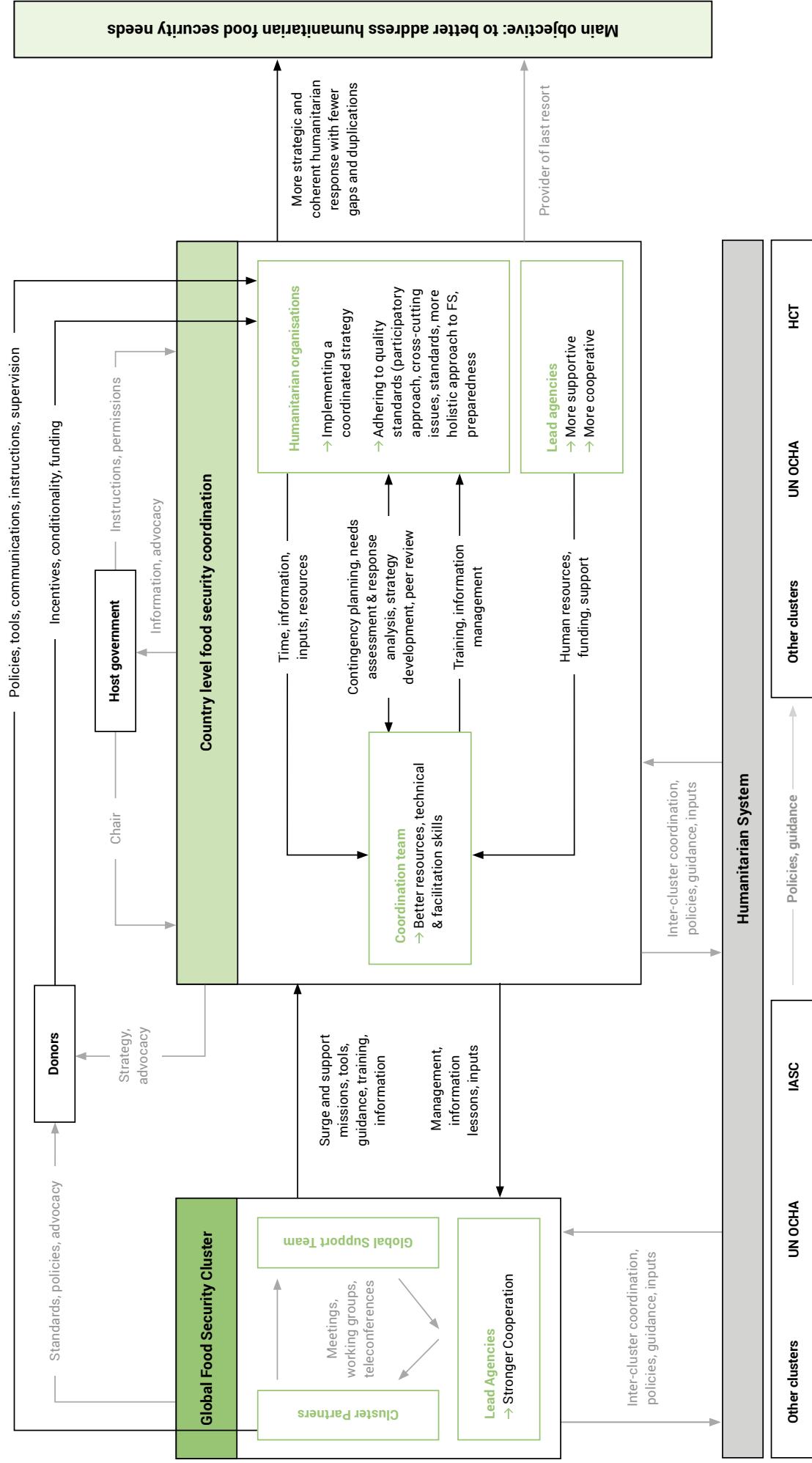
Table 5.2: Example from inter-agency evaluation of food assistance for protracted refugees

Time	Food assistance	Assumptions	Expected results
Short term	General food distribution – full rations	Emergency response assistance	Lives saved; improved food consumption; safety and protection provided. Minimal level of self-reliance.
Medium term	Food assistance decreases (partial rations)	Transition from emergency response; complementary social service interventions are available, e.g. water, sanitation, education, housing, etc.	Improved food basket, improved nutritional status (acute and chronic malnutrition). Increased capacity of affected people to establish livelihoods.
Long term	Food assistance decreases (partial rations)	Livelihood interventions available; asset building	Refugee self-reliance; local integration; resettlement or repatriation.

The inception report for the Food Security Cluster evaluation developed a graphic ToC (Steets et al., 2013: 6) as shown in [Figure 5.4](#) on the page overleaf.

Resources on developing a ToC include Funnell and Rogers (2011) and Care International (2012), and the 2011 Comic Relief review of ToCs provides a useful overview. The annex to the OECD-DAC guidance on the evaluation of peace-building interventions includes one on understanding and evaluating ToCs (OECD-DAC, 2012: 80-86).

Figure 5.4: Food Security Cluster - Theory Of Change



5.3

Developing a logic model after the fact

Ideally, the programme intervention should have been designed around a ToC in order to produce a clear statement of how it should work. However, this is not yet common practice, and the end result is that evaluators may be asked in the ToR to develop an explicit ToC to fit the one implicit in the intervention.

In principle, an implicit ToC underpinned decisions in the case of the Danida Regions of Origin Initiative in Afghanistan logic model described above, but failing to make the ToC explicit can lead to problems:

- An evaluation covering more than a single project or programme is unlikely to have a single overall logic. This was the case in the evaluation of the Danish strategy on humanitarian evaluation (Mowjee et al., 2015: 18).
- One of the most useful aspects of an explicit ToC is that it provides a framework for the evaluators to test the implementation. If the evaluators develop the ToC there is the risk of using the same data both to generate the theory and to test it.
- If the evaluators find that a particular causal linkage (as postulated in the ToC they developed) is invalid, this could be because their model rather than the intervention is invalid.

An ex-post ToC can be only a pale substitute for one developed before the programme begins and used to guide it. Developing a ToC after the actions is rather like trying to do ex-post strategic planning. Where you have to develop an ex-post theory of change, one approach is to use a stakeholder workshop to elaborate the implicit ToC under which they operated, or to validate an explicit ToC that the evaluators have elaborated from the desk study.

5.4

Evaluation frameworks and evaluation criteria

The second type of framing is through normative standards for humanitarian action. Humanitarian contexts are often complex, and there is a danger that some vital aspect may be overlooked, that the evaluation may lack conceptual rigour, or that it may not meet accepted standards. Normative frameworks can help with these issues in the following ways:

- They provide a structure for breaking up the evaluation task into smaller elements that are easier to manage.
- They reduce the risk that key elements will be overlooked by systematically directing the team's attention to all elements of the evaluation subject.
- They provide a structure for the evaluation that stakeholders will recognise. Stakeholders may have had little exposure to the standard evaluation criteria, but are often familiar with the structure of conceptual frameworks or standards in common use in their sector.
- They can provide a baseline of generally accepted good practice against which the project can be tested.

The OECD-DAC criteria described in [Section 6: Choosing evaluation questions](#) offer one possible framework for an evaluation, but all frameworks are better for checking than for developing questions.

Other frameworks may add specificity and detail to the evaluation and make it more accessible to users. Such frameworks can be used in addition to the OECD-DAC criteria, or even substituted for them if appropriate. Evaluators are often asked to use one of the following types of framework:

- Broad normative frameworks that reflect the norms that define the humanitarian endeavour. They include international humanitarian law, the humanitarian principles (UNEG HEIG, 2016), and various conventions.
- Standards and guides can be used both as standards to evaluate against and as a way to break down humanitarian actions into components that are easier to examine. They include system-wide, sector-specific standards and agency guides and manuals.

Frameworks and their uses are summarised in [Table 5.3](#). This list is not exhaustive and new frameworks are always being developed. The framework chosen should fit the context.

Table 5.3: Examples of frameworks and their use

Frameworks and examples	Possible use
<p>Broad normative frameworks</p> <ul style="list-style-type: none"> International humanitarian law such as the Protocol on the Protection of Victims of Non-International Armed Conflicts (ICRC, 1977) Humanitarian principles (Wortel, 2009) Convention on the Rights of the Child (UN General Assembly, 1989) Convention relating to the Status of Refugees (UN General Assembly, 1951) 	<p>Normative frameworks can be used during the inception phase to see if there are any aspects of the intervention that might raise concerns and need to be more closely examined. Some frameworks also provide checklists or standards against which to review policies and performance.</p>
<p>System-wide standards and guidelines</p> <ul style="list-style-type: none"> Principles and Good Practice of Humanitarian Donorship (Good Humanitarian Donorship, 2003) NGO/Red Cross/Red Crescent code of conduct (Borton, 1994) The Core Humanitarian Standard on Quality and Accountability (HAP International, 2014) Guiding Principles on Internal Displacement (OCHA, 2004) Fragile States Guidelines (OECD-DAC, 2007) Core Humanitarian Standard on Quality and Accountability (HAP, 2014) Quality Compas (Groupe URD, 2009) 	<p>These standards can provide a checklist or reference point against which to evaluate performance, a basis for breaking down the evaluation into manageable tasks, and a structure for the report. They are most effective when an agency has made a formal commitment to adhere to them; otherwise, it may be difficult to justify using them. The inception phase is the best time to propose using a particular standard.</p>
<p>Sectoral standards</p> <ul style="list-style-type: none"> Sector-specific elements of the Sphere standards (Sphere Project, 2011) Minimum Standards for Education in Emergencies (INEE, 2006) Livestock Emergency Guidelines and Standards (LEGS Project, 2009) Minimum Economic Recovery Standards (the SEEP Network, 2010) Minimum Standards for Child Protection in Humanitarian Action (Child Protection Working Group, 2012) 	<p>Sectoral standards are a good basis for organising sectoral evaluations. In some cases, they are based on general standards: for example, the Sphere standard on consultation with affected populations is meant to apply to all sectors.</p>



Frameworks and examples	Possible use
Agency standards and guides <ul style="list-style-type: none"> • United Nations High Commissioner for Refugees' Handbook for Emergencies (UNHCR, 2007) • World Food Programme's Emergency Field Operations Pocketbook (WFP, 2002) • UNICEF's Emergency Field Handbook (UNICEF, 2005) 	These documents can provide a good basis for checking compliance (accountability) and also for breaking down and organising the evaluation task.
Local standards and guides <ul style="list-style-type: none"> • National laws, standards, and policies • Locally agreed guidelines (e.g. at the Cluster level) 	These documents may often be based on broader international models, but are adapted for the local context. They are thus usually more appropriate for the context than broader international standards and guides.



Tip

Follow the lead of the ToR and project documents. If the ToR or project planning documents refer to particular conceptual frameworks or standards, it might be appropriate to use one of these as a framework for the evaluation.

A problem sometimes arises when an evaluation report tries to use several frameworks at once. This can lead to a long report as each section from framework A has sub-sections for framework B (and sometimes even framework C).



Tip

Use only one framework to structure the report. If an evaluation uses a framework and this is central to the evaluation, this may provide useful structure for the report. If the evaluation uses several standards, it may be advisable to choose one for the main framework and address the others in short annexes to the main report.

6 / Choosing evaluation questions

This section focuses on choosing the evaluation questions. These should be based on what your primary intended users' needs to know that would make a difference in their work, as explained in Section 3: Think early and often about evaluation utilisation. Questions can be divided as follows:

- The top-level questions, such as 'How effective was our response?'
- The actual evaluation questions, unpacked from the top-level questions.
- Questions asked of interviewees, focus groups and survey subjects, in interviews, topic guides and survey instruments.

This last category is addressed in Section 13: Field methods.

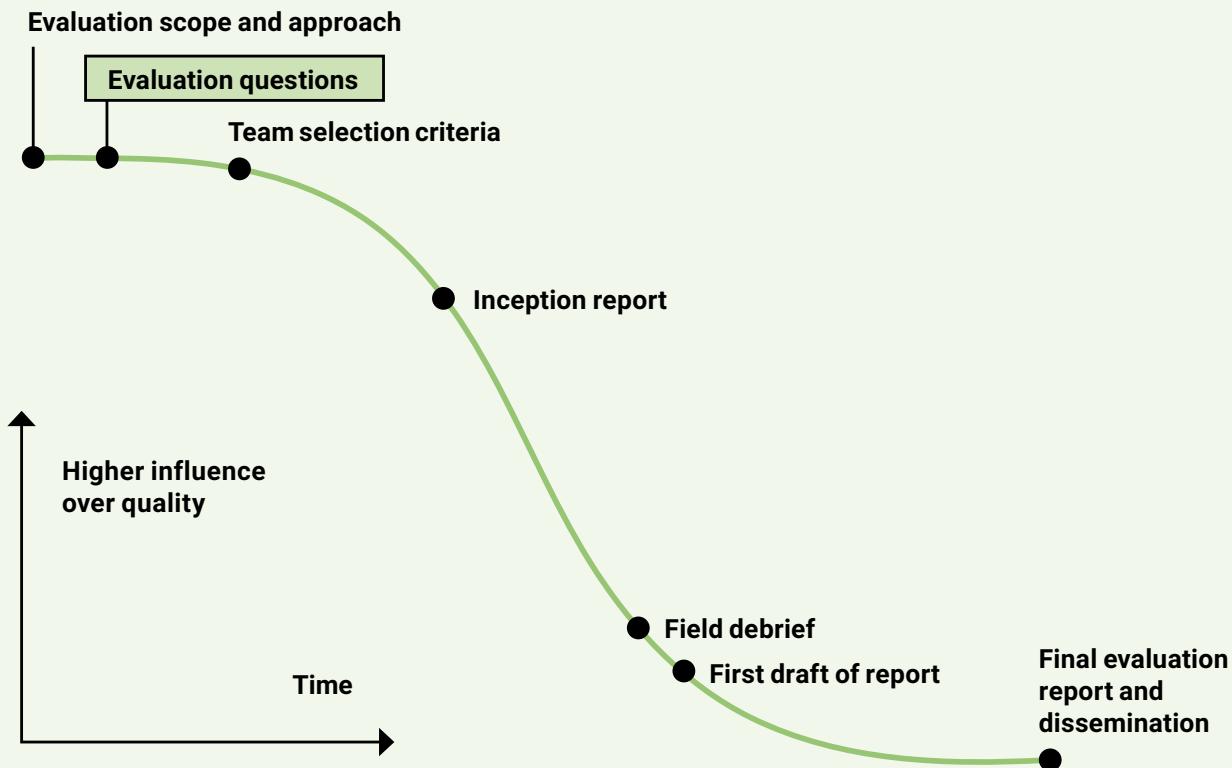
6.1 Impact of question choice on quality

The choice of evaluation questions has a critical effect on the quality of the evaluation.

The following factors can lead to a poor evaluation:

- Too many questions will mean the evaluation cannot go into any depth.
- Questions that are a poor fit with the scope and the approach.
- Questions that are best answered by other means (such as an audit or thematic study).
- Questions that are not finding answers that can be used.

Quality checkpoint: Evaluation questions



The questions you ask determine which research designs and methods are feasible and which are the most appropriate. The questions also determine what budget will be needed. Critically, the choice of questions has a strong influence on the utilisation of the evaluation.

Not all evaluations pose questions

The Participatory Evaluation of the 2008 Farmer Field School Programme in Lira in Uganda did not include any questions, only evaluation objectives. The evaluator then drew up an information requirements table for the data needed to meet the objectives (Foley, 2009). In the case of the evaluation of the Danish engagement with Somalia, no questions were given and the evaluators were asked to 'deepen and specify the evaluation question for the issues presented above' in an evaluation matrix (Gartner et al., 2012: 72). While this approach is still relatively rare, the EU has now adopted it as a standard practice for its evaluations.

The evaluation questions drive the whole evaluation

The evaluation questions have a critical role in shaping the evaluation. It is worth making the effort to get the questions right to ensure the quality of the evaluation, as:

- They drive what type of evaluation is chosen, for example impact evaluation, process evaluation or RTE.
- They determine the most appropriate design or designs. For example, if the question is about the effect of assistance, you may need to conduct a comparison study such as the joint impact evaluation studies on food aid and the protracted refugee crises case study for Rwanda (Sutter et al., 2012).
- They determine which methods need to be used. For example, if the ToR require the evaluation to 'determine overall beneficiary satisfaction' there may be a need to conduct a survey, as in the case of the IFRC evaluation of transitional shelter in Indonesia (Tango, 2009: 38).
- They drive the generation of the choices in the evaluation matrix. This is discussed in Section 8: Inception phase. See Buchanan et al. (2009: 37-39) for an example of how the questions shape the evaluation matrix.
- The questions should drive the budget for the evaluation. For example, questions that could be answered only by a survey will increase the cost.
- They influence both the type and extent of sampling choices.
The review of UNICEF's Education in Emergencies and Post-Crisis Transition used a purposive sampling strategy to meet the aims of the study (Barakat et al. 2010: 194).
- The evaluation questions determine the recommendations. As the OECD-DAC guidance on the evaluation of development assistance notes: 'The findings and conclusions of the evaluation are the answers to the questions raised and selected for evaluation. The lessons learned and the recommendations provide the link between the results of the evaluation and future policy and programme development' (1991: 10).

6.2 Types of evaluation question

Table 6.1 presents five types of evaluation question drawing both on BetterEvaluation (2013) and Morra Imas and Rist (2009: 223-229).

Table 6.1: Types of evaluation question

Question type	Example	Commentary
Descriptive	How did affected people use the shelter kits provided?	This is a fairly common type of question in EHAs. But what do you do with the results?
Normative	To what extent did the shelter provided meet the Sphere Standards?	Other normative questions may ask about achievement of project targets or agency standards, or about benchmarking against other agencies. They are focused on <u>accountability</u> rather than learning.
Causal	To what extent did the provision of assistance at the village level discourage migration to the regional capital?	Often the most difficult type to answer in EHA as it is difficult to assess attribution or even contribution given the chaotic, complex, and complicated nature of humanitarian action. Impact questions are causal by definition. See the discussion of causality in <u>Section 16: Analysis</u> .
Evaluative	Was our policy to only provide shelter kits to those with proof of plot ownership appropriate?	Asks the evaluators to make a judgement about the worth of a programme. See the discussion of evaluative reasoning in <u>Section 16: Analysis</u> . 'By definition, evaluations answer evaluative questions, that is questions about quality and value' (Davidson, 2014: 1). However, most EHAs go beyond this and ask for <u>recommendations</u> : judgements of worth do not in themselves lead to recommendations.
Action-oriented	How could we better support vulnerable persons to rebuild their shelters?	Action-oriented questions are particularly suitable for EHA as they link the question directly to what actions the agency could take in future.

Action-oriented questions are complex, and the evaluation team usually has to answer several sub-questions to answer them. In the example given about options for better supporting vulnerable persons to rebuild their shelters, the implicit sub-questions include:

- What are the features of the project? (Descriptive)
- Who is classified as vulnerable here? (Normative)
- What is better support? (Evaluative)
- What factors drive better support? (Causal)

Only after answering these will the evaluation team be able to list options or make recommendations.



Tip

Where possible, frame overarching evaluation questions as action-oriented ones

Action-oriented questions are well suited to EHA because they, and the relevant sub-questions, are strongly focused on how the findings will actually be used. Action-oriented questions also make excellent choices for framing overarching questions.

The process by which evaluators answer descriptive, normative, causal, and evaluative questions and then make recommendations is usually somewhat haphazard. Asking action-oriented questions means that recommendations are not left to chance or to the personal preferences of evaluators, but clearly indicate where you want them to give you guidance for future programming.



Tip

Consider asking evaluators for options rather than single recommendations.

Decision-making is the realm of management. Asking for options requires evaluators to discuss their respective advantages and disadvantages rather than making a single recommendation. This does not stop the evaluators from identifying their preferred option(s), but the presentation of options can be more useful for management. It is a somewhat similar approach to asking the evaluators to present conclusions and then develop recommendations in a stakeholder workshop with the evaluation users.

6.3 Number of questions

Often, the ToR for an evaluation set out too many questions. While there are a few examples where this is not the case, such as the evaluation of the effect of food aid on protracted refugees (Sutter et al., 2012), which had only four evaluation questions, it is not unusual for the ToR to identify 20 or more.

In a review of the quality of DFID evaluations, Manning recommended limiting the number of evaluation questions ‘in order to avoid reports that are too long and discursive’ (2009: xi). In a related review, Perrin found that even light-touch DFID evaluations were expected to answer dozens of questions and that the ‘lack of focusing frequently led to superficial attention to too many considerations and ultimately often to a long list of findings where the key findings and implications are hard to identify’ (2009: vii).

The problem of asking too many evaluation questions is even greater for EHA, which are often conducted on a relatively small budget.

There are several good reasons to limit the number of questions:

1. **To focus the evaluation.** As Perrin (2009: vii) noted in his review of DFID evaluations, ‘When everything is a priority ... nothing is’. USAID (1996: 2) also notes, ‘Asking too many questions can result in an unfocused effort’. The number of evaluation questions should therefore be limited to the priority areas of interest. If there are too many questions the evaluation team may give priority to those of most interest to them, which may not coincide with the priorities of the commissioning agency.
2. **To ensure that the evaluation team addresses the questions in sufficient depth.** If there are 25 questions, it will be possible to spend only limited time on seeking to address each one. This leads to shallow evaluation reports.
3. **To make the evaluation more useful.** Few organisations can deal with more than half a dozen recommendations at a particular level. Each evaluation question needs an answer, and the answers are usually reported as findings, which lead to conclusions. While a single recommendation may draw on several conclusions, it is more common for one conclusion to lead to several recommendations. Too many recommendations will make it harder to use the evaluation.

Evaluation departments are gradually beginning to limit the number of evaluation questions in order to address these issues (Riddle, 2009: 16).

Hallam and Bonino (2013) note that:

ICRC used to request all interested parties to identify the range of questions and issues they would like included in an evaluation. The evaluation department then reframed this into 'evaluable' questions. However, the scope of the evaluations always grew, until it became difficult to manage the process. To mitigate this, the evaluation department now tries to focus on just three key questions for each evaluation.

The software used by the NRC to generate evaluation ToR allows for only three evaluation questions. Method Labs created a list of guiding questions on how to narrow the scope to data collection (Buffardi, 2015).

6.4 Unpacking overarching questions

How do you go about evaluating high-level questions like: 'Have we got the right strategy?' Such questions are difficult to answer in a rigorous way.

One approach to unpacking such questions is to consider how the answer to the high-level question might be used. Clearly, the reason to ask such a question is to decide whether to maintain the current strategy or to improve it in some way. There may also be an interest in benchmarking strategies against those developed by others.

The underlying questions are probably:

- How can we improve our current strategy? (Action-oriented)
- What are the advantages or disadvantages of our current strategy compared to those employed in similar contexts? (Normative)

In summary, look at what the commissioning agency might do with the answer to a question and then see if it can be refined to provide a better basis for action. Even the first of these underlying questions is probably too broad. The current strategy will consist of a mix of elements.

So the questions may be:

- What additions, deletions, or variations could we make to increase the impact or reduce costs? (Action-oriented)
- How does our strategy compare to those of other agencies and/or contexts? (Normative)

In this case it would make sense to first tackle the comparison with strategies in other contexts, since this lends itself to a desk study. The desk study may generate lines of enquiry for the examination of elements of the current strategy.

This process produces questions that are much easier to answer in a rigorous way. The underlying evaluative questions still need to be answered, because comparing strategies or alternatives entails evaluative judgements, but the judgments are now framed within practical actions. See Section 16: Analysis for a discussion on how to answer evaluative questions.

In some cases the evaluation team is left to unpack the high-level questions into evaluation questions that they strive to answer, and present these in the inception report, often in an evaluation matrix. However, the commissioning agency may be better placed to determine how the answers to evaluation questions might be used.

6.5 Evaluation criteria



Tip

The evaluation questions come first.

The evaluation criteria and other frameworks are useful tools once you have decided on your evaluation questions. Criteria and frameworks are good servants but poor masters. Avoid listing your criteria first and then picking questions under each one, and remember to maintain a strong focus on how the findings and recommendations will be used.

We engage in evaluation every day. Whether buying apples at a supermarket or choosing a sandwich in the canteen, we make a judgement about the options before us based on our personal criteria of quality and value. A formal EHA uses a much more structured approach but essentially looks at similar criteria.

In 1991, the OECD-DAC, focusing on the most common problems noted in development projects, proposed four quality criteria – relevance, effectiveness, sustainability and impact – and the value criterion of efficiency (OECD-DAC, 1991). A few years later, it adapted these criteria for EHA in complex emergencies (OECD-DAC, 1999), adding coverage and coherence, suggesting appropriateness as an alternative to relevance and connectedness as an alternative to sustainability, and proposing two new criteria: coordination and protection. These criteria reflected the biggest problems seen in humanitarian action in the 1990s.

Although the OECD-DAC criteria are not perfect, there are good reasons to use them to check your evaluation questions:

- Using standard criteria makes meta-evaluation (the drawing of lessons from a wide range of evaluations) much easier.
- Standard criteria are likely to capture common weaknesses in humanitarian action, based on experience and research.
- Evaluations that use standard criteria are easier for evaluation managers and other evaluators to work with.



Tip

You don't need to cover all the criteria. Your evaluation questions may touch on only one or two criteria.

At the EHA planning stage, first identify what you need to know and then place it within the evaluation criteria – not the other way around. The criteria are tools to think with and may suggest additional relevant questions. Ask evaluation questions only if you are ready to take action based on the answers, and use only the criteria that relate to the questions you want answered. It is the questions that matter, not the criteria.

A single evaluation question may span several criteria. For example the question: 'to what extent did our targeting strategy enable us to deliver timely assistance in a cost effective manner?' includes the criteria of coverage (targeting), effectiveness (timeliness) and efficiency (cost-effectiveness). You can break the question down further, but any questions about the effectiveness or efficiency of targeting inevitably include the coverage criterion.

In Table 6.2, the principal criterion is identified for each of the sample questions, with comments and an indication of what action an agency might take in response to the answers. In addition, an utilisation-focused alternative question is also given, which can help to make evaluation reports less discursive.



Tip

Use action-oriented questions for smaller evaluations. Action-oriented questions are particularly appropriate for small evaluations as they lead to a direct focus on what an agency might do to improve future performance.



In depth: Evaluation questions and their criteria

Table 6.2: Evaluation questions and their criteria

Sample questions	Principal criterion	Definition of criterion
To what extent did our programme meet immediate needs? How can we ensure that our kitchen sets are a better match with local needs?	Appropriateness	The extent to which humanitarian activities are tailored to local needs, increasing ownership, accountability and cost-effectiveness accordingly. (Replaces the relevance criterion used in development evaluations.)
To what extent did the programme achieve its objectives, including the timely delivery of relief assistance? What changes can we make to reduce the supplementary feeding drop-out rate?	Effectiveness	The extent to which an activity achieves its purpose, or whether this can be expected to happen on the basis of the outputs.
How cost-efficient was our shelter programme? How can we reduce waiting times at the health clinic?	Efficiency	The outputs – qualitative and quantitative – achieved as a result of inputs.
What has been the impact of the cash voucher programme, positive and negative? What measures could we take to reduce the damage caused by firewood collection by the refugees?	Impact	The wider effects of the project – 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). (This is not exactly the same thing as 'Impact' in the results chain.)





Table 6.2: Evaluation questions and their criteria (continued)

Sample questions	Principal criterion	Definition of criterion
<p>How has the provision of free livestock vaccines effected the cost-recovery approach of community animal health workers?</p> <p>What can we do to prevent the food distributions from discouraging farmers from planting?</p>	Connectedness	<p>The extent to which activities of a short-term emergency nature are carried out in a context that takes longer-term and interconnected problems into account. Replaces the sustainability criterion used in development evaluations.</p>
<p>To what extent have cash transfers benefited the most vulnerable households?</p> <p>How can we ensure that marginalised groups and individuals also have access to the shelter grants?</p>	Coverage	<p>The extent to which major population groups facing life-threatening suffering were reached by humanitarian action.</p>
<p>How coherent are agency policies on protection, and what are the implications?</p> <p>How could we advocate that other donors take human rights into account in funding decisions?</p>	Coherence	<p>The extent to which security, developmental, trade, and military policies as well as humanitarian policies, are consistent and take into account humanitarian and human rights considerations. (More focused on donor policy, but can also be applied to individual agencies on their own policy coherence.)</p>
<p>How well coordinated has food distribution been, across the region, with what consequences?</p> <p>How can we reduce gaps in water supply by the different agencies using water tankers?</p>	Coordination	<p>The extent to which the interventions of different actors are harmonised with each other, promote synergy, avoid gaps, duplication, and resource conflicts. (Often folded into effectiveness.)</p>

Source: Adapted from Beck (2006), who provides further information on using the DAC criteria in evaluation.

Avoid starting with the criteria and then selecting questions for each one, both because of resource limitations but also because an EHA should focus on users' needs.



The questions need to drive humanitarian evaluation. The evaluation criteria serve only as a checklist to make sure that no major problem area has been left out. Perrin (2009: xi) recommended that DFID evaluations 'should be focused on a limited number of key evaluation priorities and issues, using the DAC criteria intelligently rather than mechanistically to help in identifying a small number of evaluation questions that are realistic to address, taking into account the context, data availability, and the scope and resources provided for the evaluation'.

To ensure completeness and comparability of evaluations, some agencies may want to show which evaluation criteria are covered by particular questions. In some cases one question may address several criteria, but it is not always feasible to cover all the criteria.

Circulating the Terms of Reference

When the ToR are circulated for comment, the number of evaluation questions grows as each department/stakeholder adds a few suggestions. The wider the circulation, the larger the number of questions. Hallam and Bonino (2013) make the point that 'high-quality evaluations that address real information needs in turn, increases the demand for evaluations'. The questions added by circulating the ToR are often poorly considered and are likely to detract from rather than increase the demand for evaluation.

Strategies for controlling the number of questions include:

- Including only the overall evaluation objectives in the ToR and leaving the team to develop the evaluation questions in the inception report.
- Asking reviewers to rank suggested questions, and to rank and justify any additional questions they propose.
- Asking those suggesting additional questions to explain how they would use the answers and how this would justify any extra cost.



Tip

Circulate ToR as a PDF rather than as a readily editable document. This discourages the adding of questions.

6.6 Selecting the best questions

If there is a large number of questions you will need to rank them. One way to do this is to score the questions against different factors. This may include the extent to which the answers to the questions will meet users' needs.

For example, ask if the questions:

- Are central to the objective of the evaluation
- Can immediately have their answers applied
- Can be answered unequivocally by the evaluation
- Can be answered only by an evaluation
- Are central to your core mandate
- Could improve the quality of the service you provide
- Could reduce the costs per head of providing a service
- Would affect or be of value to almost all of the stakeholders.

The factors, or the weighting that you give them, will vary according to the aims of the evaluation. This can be presented as a rubric (see [Table 6.3](#)). The important thing is to decide in advance which factors are important and to rate the potential questions accordingly.

Table 6.3: Rubric for assessing evaluation questions

Factor question Apply factor question to your potential evaluation question		Score Pick the description that most closely matches your analysis of the potential evaluation question and apply the score (1-4)				Weight* How significant is the factor question to your evaluation?					Score for each potential question (score x weight)				
		Score = 1	Score = 2	Score = 3	Score = 4	Potential question 1	Potential question 2	Potential question 3	Potential question 4	Potential question 5	Total score for potential question				
A	If we had the answer to this question we could...?	Be better informed about our work	Use it to advocate for changes in the way we do things	Apply it in future programmes	Apply it immediately to our ongoing programmes										
B	Can the evaluation answer this question unequivocally with the available resources?	Unlikely	Possibly	Probably	Definitely										
C	Could this question be answered by other means?	Yes, but we have the budget for evaluation	An evaluation is the most convenient way of answering this question	An evaluation is the cheapest way of answering this question	No only an evaluation could answer this effectively										
D	How central is the question to our mandate?	This is peripheral to our core mandate	This is relevant to our core mandate	We expect this will be part of our core mandate in the future	This is central to our core mandate										
E	What is the potential impact of the answer to this question on the quality of the services we deliver?	Little impact on the quality of our services	Some impacted on the quality of our services	A large impact in the quality of the services we deliver	It could substantially improve the quality of services we provide										
F	What is the potential impact of the answer to this question on the cost of the services we deliver?	Little impact on the cost of our services	It could lead to some savings	Large savings in cost per head	It could significantly reduce the cost of services we deliver										
G	What proportion of our overall target group would be affected by the answer to this question?	Few	Few, but they are particularly vulnerable	Many, including those who are particularly vulnerable	Almost all										

*Some factors (A–G) may be more significant for some evaluations and may need to have their scores multiplied by a weighting factor.

7 / Terms of Reference and budgeting for an evaluation

This section covers the Terms of Reference (ToR)³ for the evaluation. The ToR are represented in a document setting out the evaluation task in sufficient detail for potential evaluators to understand what is expected.

This section also briefly covers budgeting, both for the evaluation manager to understand what is potentially feasible and for the evaluation team to prepare a bid.

7.1 Terms of Reference (ToR)



Definition: Terms of Reference

The ToR present: 'an overview of the requirements and expectations of the evaluation. It provides an explicit statement of the objectives of the evaluation, roles and responsibilities of the evaluators and the evaluation client, and resources available for the evaluation' (Roberts et al., 2011: 2).

The ToR form the basis of the contract between the evaluation team and the commissioning agency.

The ToR are critical to establishing the quality of the evaluation as choices about the scope and approach can have a large impact on how usable the result will be.

The ToR defines the scope, in terms of elements of the project or programme, the time period, and specific interventions most likely to provide you with usable answers. The ToR may also set out evaluation approach is to be used and the budget that should be used to answer the questions. It clarifies whether answers would best be delivered by an evaluation centred on staff, partner organisations, or affected people.

7.1.1

The ToR and inception report

Davidson (2012) underlines that evaluation is a joint enterprise between the commissioning agency and the evaluators.

High quality, worthwhile, actionable evaluation doesn't just depend on the technical competence and effective consultation skills of the evaluator.

Decisions made and actions taken (or not taken) by the client can make or break the value of evaluation for an organisation. High-value evaluation is the product of a fruitful interaction between a well-informed client and a responsive, appropriately skilled evaluation team.

The evaluation manager has two options for engaging with the evaluation team to negotiate the scope of the evaluation.

One approach treats the ToR as being in draft until after the evaluation team has been recruited. Outstanding issues are agreed as part of finalising the contact. This is good practice, but the financial systems in some agencies preclude this type of negotiation.⁴

Another approach is for the evaluation team to use the inception report to propose how it will address the ToR. The evaluation manager and the evaluation team can use the inception report as a basis for negotiating the final scope of the evaluation. An inception mission can contribute to the quality of the final evaluation scope.

The combined ToR and inception report define what is to be evaluated and why, and how it is going to be evaluated. This is why the UNEG checklist covers both the ToR and the inception report (UNEG, 2010b). The evaluation manager may either define some items in advance or leave them for the evaluation team to address in the inception report.

Table 7.1: Must-haves and nice-to-haves in an evaluation ToR and inception report

Items typically included in the ToR	<ul style="list-style-type: none"> • Context • Purpose and how it will be used • Objectives • Criteria • Scope • Audience • Roles and responsibilities • Milestones • Deliverables • Contents of the inception report
Items that may be included in the ToR or inception report	<ul style="list-style-type: none"> • Evaluation frame • Evaluation questions • Sources to be used • Evaluation matrix • Evaluation designs • Data-collection methods • Indicators to be measured • Data analysis methods • Contents of the evaluation report
Items typically included in the inception report	<ul style="list-style-type: none"> • Work plan • Allocation of work within the team • Data-collection tools

The ToR communicate the evaluation manager's intent to the evaluation team, and the inception report is the team's response to this. Thus the ToR and the inception report form a dialogue, supplemented by discussions on the inception report. Inception reports are appropriate even in smaller evaluations. The only exception to this is when the unit or department preparing the ToR will also conduct the evaluation.

As a general rule, unless there is some overriding issue (such as expecting that the evaluation team will be relatively inexperienced), it is better to leave as much as possible for the team to address in the inception report rather than defining it in the ToR. This is because by the inception report stage, the issues to be evaluated should be far more clearly identified. It is also better to leave the process of presenting the intended product to the evaluation team.

The ToR should be prepared even for internal evaluations, so that the objective, purpose, and scope are clearly set out. This avoids misunderstandings arising later.

7.1.2

What does the ToR include?

The ToR for even the smallest evaluation should include: (also see [Table 7.1](#))

- The context of the evaluation
- The objectives of the evaluation
- The scope in terms of the project or programmes covered, the sectors to be evaluated, the geographic extent, and the timeframe covered
- The milestones, deadlines and deliverables
- The roles and responsibilities of the evaluation team and of the evaluation manager
- What resources, including existing data, are available

In most cases the ToR will also include the evaluation questions to be answered. Some ToR state only the evaluation objectives and leave it up to the team to propose questions (to be agreed by the evaluation manager) in the inception report.

Guidance on writing ToR includes:

- The DFID ToR Template, which is short (two pages) and to the point (DFID, 2010)
- The UNEG quality checklist for ToRs and inception reports (UNEG, 2010)
- The World Bank's How-to Guide on ToRs (Roberts et al., 2011) provides a thorough guide
- USAID's guide to writing a Scope of Work (USAID, 1996)

Small evaluations may include contractual matters in the ToR

In the case of small evaluations, the ToR may also include the details of the tendering procedure. In this case the ToR will include the bid deadlines, the documentation to be submitted with the bid, on how the bids will be scored, and on any specific requirements (e.g. separate technical and financial bids). Large evaluations may set out these details in the formal Request for Proposals and present only the evaluation details in the ToR.

Table 7.2: Potential elements of the ToR

Element	Comments
Background	This requires only a couple of paragraphs; most of the evaluation team's contextual knowledge will come from background reading. Good sources for this include the introductions to project proposals or appeal documents.
Purpose of the evaluation	Is the evaluation mainly for learning or accountability? If both, which has precedence? For example, an evaluation may be intended to inform donors on how effectively their money was used (accountability) but might also examine which factors in project design and implementation led to the most effective projects (learning). All evaluations, whether primarily for accountability or learning, are learning opportunities.
Context of the evaluation	Why evaluate now? Are there any internal or external deadlines? Is the evaluation tied to decisions in a funding cycle or decisions about the agency's future strategy? Such deadlines and linkages should be clear if you have involved primary stakeholders from the outset, and should be made explicit in the ToR. Remember this relates to utilisation, see discussion in <u>Section 3</u> .
Scope of the evaluation	What sectors, geography, phase of the response, and timeframe will the evaluation cover? Is it focused on the policy level? The programme level? On specific operations or processes? Is it a real-time evaluation, mid-term evaluation, or ex-post evaluation? Is it a single-agency evaluation, self-evaluation, or joint evaluation? This is related to the type of evaluation, see <u>Section 4</u> .





Table 7.2: Potential elements of the Terms of Reference (continued)

Element	Comments
Users	<p>How are users expected to use the evaluation? The answer to this question should determine the length and readability of the evaluation outputs. See Section 3: Think early and often about evaluation utilisation.</p>
The evaluation frame	<p>Is there a conceptual model that you want evaluation team members to use when selecting the research methods and carrying out their analysis – for example, the livelihoods framework for complex humanitarian emergencies? What international standards are relevant to this evaluation – for example, Sphere standards or the Core Humanitarian Standards? See Section 5: Framing your evaluation.</p>
Main evaluation questions	<p>What are the main questions you want the evaluation to answer? Which OECD-DAC criteria do you want to use, and how do they relate? See Section 6 for discussion both on choosing evaluation questions and the OECD-DAC criteria.</p> <div data-bbox="568 1073 652 1156" data-label="Image"></div> <p>Tip: You may prefer just to state the objectives and leave the evaluation team to propose questions in the inception report.</p> <div data-bbox="568 1253 652 1336" data-label="Image"></div> <p>Tip: Keep the number of evaluation questions short in order to keep the evaluation focused.</p>
Inception phase	<p>The ToR should make clear what activities are expected in the inception phase. Is this limited to a desk study, or does it include an inception visit? The ToR should also set out the expected content of the inception report. See Section 8: Inception phase.</p>
Designs	<p>An evaluation may use a range of designs. A given type of question may be best answered by a particular design, but logistics and budget constraints may limit choices. In general, the team proposes designs in the inception report, but the commissioning agency may specify that a particular design should be used – to conform to overall agency policy, for instance. See Section 11: Evaluation designs for answering different evaluation questions.</p>





Table 7.2: Potential elements of the Terms of Reference (continued)

Element	Comments
Data-collection methods	<p>This should not be a detailed statement, but rather an indication of any methodological preferences – for example, if you want the team to consult specific people, such as government officials or affected people, or use particular methods, such as a formal survey. Usually, the team should develop the detailed methodology in the inception phase. If the evaluation manager specifies both the product and the process, the evaluation team may feel less sense of ownership and responsibility for the final result. See Section 10: Desk methods and Section 13: Field methods.</p>
Indicators	<p>The indicators (if quantitative indicators are used) are usually left to the team to propose in the inception report, but the commissioning agency may require the use of particular indicators (e.g. to permit comparison or synthesis with other evaluations).</p>
Data-analysis methods	<p>The ToR should state whether there are any requirements to use a particular analytical approach. Usually, the team proposes its analytical approach in the inception report. See Section 16: Analysis.</p>
Timetable	<p>Specify dates for key deliverables and deadlines. Ask stakeholders if there are any unsuitable times – for example, the rainy or harvest season, or administratively busy periods. See Section 9: Planning and managing your evaluation for a discussion of timelines.</p> <div data-bbox="568 1545 654 1644" style="background-color: #e0f2e0; padding: 5px; border-radius: 10px;"></div> <p>Tip: Allow enough time for circulation of a draft report (typically 14-21 days) and revision of the report (another 14-21 days).</p>
Roles and responsibilities	<p>Specify who is responsible for providing transport in the field, making appointments, and other tasks. Check with your colleagues in the field to make sure that they can provide the necessary support to the evaluation team.</p>





Table 7.2: Potential elements of the Terms of Reference (continued)

Element	Comments
Management arrangements	Specify whether there will be advisory, steering, or reference groups and what their composition and roles will be. See Section 9: Planning and managing your evaluation .
Skills and qualifications	What skills and qualifications does the evaluation team need to successfully carry out this evaluation? These may include broader context knowledge (of country, sector, or organisation), languages, or skills in particular methods or forms of communication. See Section 9: Planning and managing your evaluation for details on what skills an evaluator may need.
Outputs	<p>Outputs usually include an inception report, field debriefings (as a note or a presentation), a main report, an evaluation summary, and debriefings at general meetings or any other format considered useful for disseminating the findings. Specify the length and format if there is a preference. It is also useful to establish phased payments against specific outputs to encourage the evaluation team to maintain a goal-oriented approach. See Section 17: Reporting and communicating evaluation findings with a utilisation focus.</p> <p>If there is a house style to which the evaluation report should adhere, or strong preferences about format and style, specify these to avoid having to revise the report later. Specify the length of the report (number of words rather than number of pages) and what, if any, annexes it should include. You may wish to specify the broad structure of the report, although it is good to give the evaluation team some flexibility in this regard.</p>
Risk management	Describe the risks and challenges that are expected to arise in the evaluation and ask how the team proposes to deal with these. See Section 15: Constrained access .





Table 7.2: Potential elements of the Terms of Reference (continued)

Element	Comments
Budget	Give an indicative budget for the evaluation. See Section 7: Terms of Reference and budgeting for an evaluation .
Available data	Identify the main sources of documentary data that will be available to the evaluation team. Will they have access to the intranet, email records, and grey literature (e.g. previous evaluations of the project or programme)? Patricia Rogers notes that one reason evaluations fail is that evaluation teams are not given 'data, information that data exist, previous evaluations, concurrent evaluations, planned policy changes, forthcoming personnel changes, the dates of critical decisions and meetings' (comments posted to Davidson, 2012b).
Bid assessment	If there is no formal tender request document, you may want to include the basis on which bids will be assessed (what percentage of marks will go for price, team composition, data-collection and analysis methods, and other criteria) and the deadline for the receipt of tenders. See Section 9: Planning and managing your evaluation .
Key references	You may want to attach a list of key references for the evaluation. Many of these will be internal documents, but check with primary stakeholders whether other documents should form part of the initial set of reading. See Section 10: Desk methods .

7.1.3

ToR timeline

Developing a ToR can take as little a week in a small organisation or over a year in a large one. It usually takes longer to finalise a ToR when there are multiple stakeholders in different countries.

The IASC developed a pre-agreed draft ToR for Inter-Agency Real Time Evaluations of humanitarian emergencies in order to circumvent the need for negotiation among the stakeholders each time (IASC, 2010). This same draft ToR expected that contracts would be signed with the successful team 30 days after the ToR and requests for Expressions of Interest were published. See [Section 9: Planning and managing your evaluation](#) for a discussion of evaluation timelines.

Some agencies typically take longer than this. A review of eight EHAs in which the authors have participated found that the time between publishing the ToR and signing a contract ranged from two weeks to three months. Most EHAs allowed two months between publishing the ToR and contracting the team. One month is probably adequate for small evaluations.

7.1.4

ToR length

ToR do not have a standard length. A review of a convenience sample of 30 EHAs found that most ranged from two to six pages and three were over 15 pages. The shortest were for where only one external evaluator was being recruited or where the evaluation team was internal.

More complex evaluations had more extensive ToR.



Tip

The ToR should reflect the complexity of the task. For a small evaluation, the ToR should be correspondingly short.

The largest single cost for any evaluation employing external evaluators is the number of consultant days. This is usually a product of the size of the team and the number of days of fieldwork multiplied by the fees and daily subsistence rates. Estimating the total number of days required is the first step in estimating the cost of the evaluation. A budget can be broken down into preparatory, fieldwork and follow-up phases.

7.2 Budgeting for an evaluation

Inception phase

This can include an initial meeting to discuss the evaluation, background reading and a desk study, an initial interview, and writing an inception report.

- The initial meeting may require a couple of days including travel, but perhaps not all team members need to attend.
- Background reading, initial interviews, and the inception report may take from five to ten days for a small evaluation and several months or more for a large, complex evaluation.

A small single-agency evaluation of a single activity at a single site will typically allocate between three and seven days per evaluator for preparatory work, depending on the nature of the evaluation. Like report writing, proper preparation often takes longer than has been budgeted. It is important to leave some flexibility.

Fieldwork

Fieldwork costs – over and above staff salary and evaluation-related expenses (e.g. airfares, local transport, translation) – are determined by the length of the fieldwork, and the fees and subsistence costs for staff and/or external evaluators.

- As a general rule, quantitative methods demand more extensive fieldwork and larger field teams with technical statistical design and analysis skills; qualitative methods require more highly skilled fieldwork.
- A small evaluation would typically allow a week for observation and interviews at field sites along with half a week of initial briefing interviews and another half week for follow-up interviews and debriefing. More time is needed if there are multiple field visits, or if the team is expected to produce a draft report before leaving.
- A small evaluation would typically allow 12-14 days for fieldwork, including travel; a large evaluation may require months of fieldwork.
- A good rule of thumb is a minimum of one week for every country visited plus at least one week for each major site (e.g. province or district).
- If the team is to engage in detailed consultation with affected people, allow up to three weeks per site.
- Quantitative survey methods should be budgeted according to the sample size and the time it will take to process and analyse the data.

Table 7.3: Estimating the number of days for team leaders (TLs) and team members (TMs)

Activity	TL Days	TM days
Inception meeting including travel		
Initial interviews		
Desk study		
Inception field visit		
Drafting data tools		
Drafting analysis plan		
Evaluation matrix		
Inception report		
Meeting on inception		
Revising inception report		
Travel to field		
Initial briefing		
Meetings in capital		
Fieldwork 1		
Mid-term review meeting		
Fieldwork 2		
Preparing debriefing		
Debriefing		
Return travel		
Data collation		
Data analysis		
First draft of evaluation report		
Presentation of report		
Revision of first draft		
Revision of second draft		
Final editing		

After fieldwork

- The tasks consist primarily of collating and analysing the data and writing the report. Again, this phase is commonly under-budgeted. Collation, data analysis, and writing almost always take longer than the allocated time, depending on the complexity of the report and the amount of analysis needed.
- Debriefings tend to take a couple of days, but travel time can add to this. Writing the report may take anything from five to 20 days, and each review cycle from two to ten days.
- If there are many team members, a good rule of thumb is to add five days to the report preparation time to allow the team leader to incorporate their material.
- A small evaluation would typically allocate seven to 12 days for post-fieldwork activities.



Tip

Don't forget your report production and dissemination costs. Don't wait to have a final draft ready before realising that there are no resources for copy editing or including visual elements (pictures, infographics, etc.). Dissemination also has real costs (sending speakers to sectoral or regional meetings etc.).

Common budget elements are outlined in Table 7.4.

Table 7.4: Cost elements for an evaluation

Budget item	Possible elements
Personnel	Staff pay and allowances, allowances for partner agency staff and other staff
Evaluation consultants	Team leader, international, national and other consultants
Support staff	Pay and allowances for administration, background researchers, interpreters, drivers, security staff, and others
Travel	Visas, flights for evaluation team and accompanying staff, transport to attend briefings and debriefings, internal travel by team and accompanying staff
Subsistence allowances	Accommodation and per diem costs for consultants
Data entry	Data input and cleaning of data to remove nonsense responses, such as someone recorded as both male and pregnant
Meetings and workshops	Venue hire, meals and allowances
Report production	Copy editing, translation, artwork, graphic design, layout, printing, development of electronic media, distribution
Other products	Cost of writing, copy editing, graphics etc. for briefing note, infographics, or '10 things to know about...'
Dissemination costs	Travel costs and fees for presenting the results at national or regional level, or at sectoral meetings or other venues
Miscellaneous	Communications, mail and couriers, teleconferencing, licences and legal fees, security
Overheads	Often estimated as a percentage of other costs

Endnotes

4 / Types of evaluation

1. The OECD-DAC has produced guidance on managing joint evaluations (OECD-DAC, 2006). Joint evaluations of humanitarian action are the focus of a chapter in ALNAP's seventh review of humanitarian action (Beck and Buchanan-Smith, 2008).
2. See Breier, 2005.

7 / Terms of Reference and Budgeting

3. USAID calls the ToR a 'Scope of Work', which is probably a more accurate description.
4. For example, the DFID ToR Template states that 'Consultants will not normally be involved' in drafting the template (DFID, 2010:1).

Notes

Planning and designing the evaluation



8 / Inception phase

This section deals with the inception phase of the evaluation and particularly with the inception report.



Definition: Inception Phase

The inception phase of the evaluation goes from the selection of the evaluation team up to approval of the inception report.

During the inception phase the team tries to develop a full understanding of the evaluation task and prepares a report outlining the plan for the evaluation. The inception phase occurs prior to fieldwork; its output is the inception report. This report can be relatively short, as in the case of the Haiti Emergency Relief Response Fund evaluation (Moriniere, 2011a), and can be compared with the final evaluation report (Moriniere, 2011b).

Most ToR in EHA are fixed rather than negotiated between the evaluation manager and the evaluation team. The inception report allows the evaluation team to define a specific plan and agree it with the evaluation manager, as well as to raise concerns and address any ambiguities in the ToR. For example, the work implicit in the ToR may take a lot longer than is budgeted for; the inception report is an opportunity for the evaluation team to clarify what it is feasible to cover.



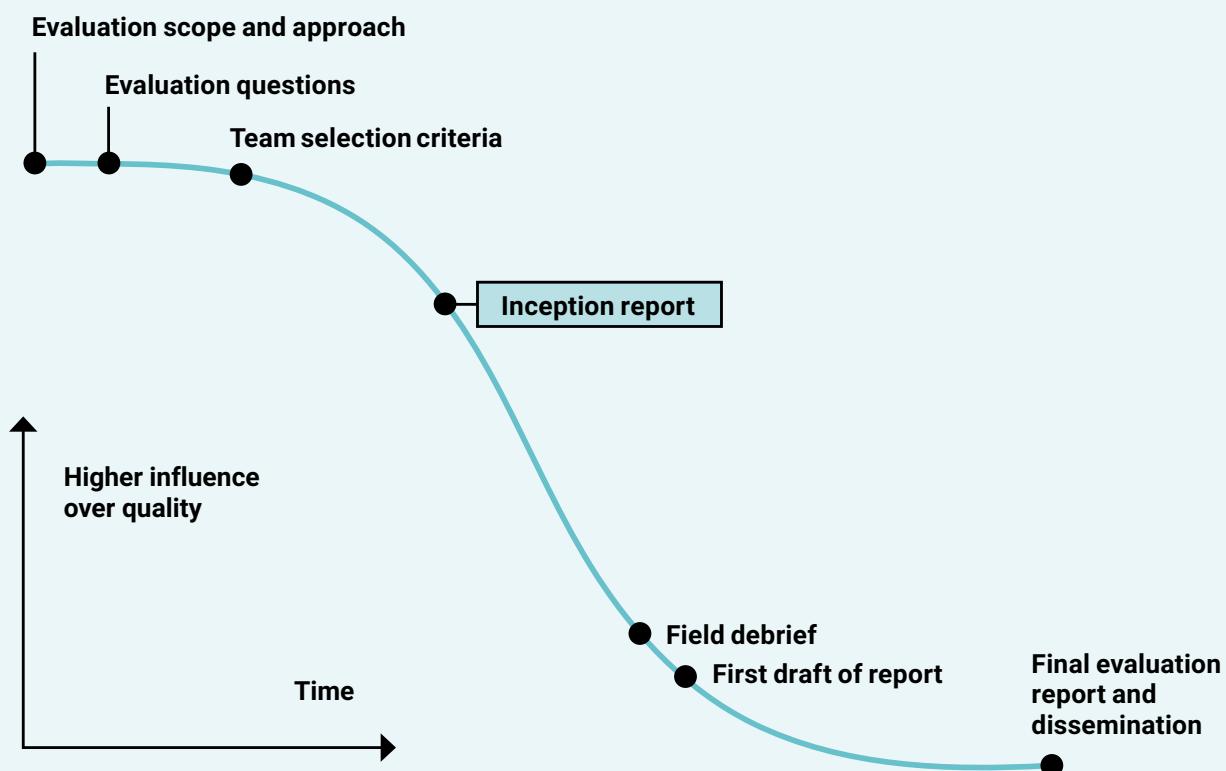
Tip

Do an inception report, even for the smallest evaluation, it is almost always worthwhile for the evaluation team to produce an inception report even if it is not a requirement to do so, as it helps the team to plan its work and minimise potential misunderstandings between the team and the evaluation manager.

8.1 Inception report and quality control

The inception report is probably the most useful tool in the quality control toolbox. As well as helping to establish a common understanding of the evaluation task, it can also indicate to the evaluation manager whether the evaluation team is unlikely to produce a quality evaluation. However, while a poor quality inception report is a good indication that the evaluation team may not be up to the task, a good quality report is no guarantee that they are. UNEG (2010) provides a checklist for ToR and inception reports.

Quality checkpoint: Inception report



The inception report provides the main quality check on the ability of the evaluation team to deliver the planned evaluation. This is the last chance for the evaluation manager to ensure the quality of the work done in the field.

The rise of the inception report

Inception reports and the related evaluation matrices are relatively new in EHA.

An inception report:

- Helps to get everyone started on the same basis
- Helps to identify over-ambitious evaluation scope
- Helps the evaluation manager to assess how the team understands and plans to approach the evaluation
- Provides an evaluation team with the opportunity to turn the ToR into a realistic plan that is agreed with the evaluation manager
- Gives the evaluation team a chance to clarify the ToR and highlight tensions that the commissioning agency needs to resolve (for example, conflicting internal expectations regarding the evaluation)
- Highlights flaws in the proposed design and methods
- Gives other stakeholders a chance to receive a clear statement of intent by the evaluation team so that they can quickly flag any concerns about the proposed approach
- Helps to identify teams that may not be up to the task
- Enables the evaluation team to clearly state what it will do and, sometimes more importantly, will not do.



Tip

For a complex evaluation, separate the inception and fieldwork phase. For complex evaluations it can be useful to issue separate contracts for the inception and fieldwork phases. Bidders are invited to tender for both phases, but the contract for the fieldwork is issued only on acceptance of the inception report.

8.2 Activities in the inception phase

The inception report should present a clear and realistic plan of work and timetable that takes existing constraints into account. The work plan should include the allocation of roles and responsibilities within the team, any deadlines for intra-team reporting, and detailed travel plans. The effort needed to establish a clear evaluation plan depends on the scope and complexity of the evaluation. The inception phase may include:

- Initial interviews with key informants to establish the context
- A desk study, including a literature review and the study of secondary quantitative data
- A workshop to present a draft inception report, which can be useful for validating the approach
- Developing and, if necessary, testing data-collection tools
- An initial scoping visit to the fieldwork country, which is good practice for large and complex evaluations.



Tip

Allow enough time not only for the team to conduct a desk review and initial interviews, but also between the submission of the inception report and the fieldwork to allow for fine-tuning.

8.3 The evaluation matrix

Most inception reports include an evaluation matrix, setting out a plan for answering each of the evaluation questions. Two examples, one from a large donor and another from a small NGO evaluation, are given below. Because evaluation matrices are relatively new, there is no single agreed format, and the column headings may vary.

At their simplest, evaluation matrices might have potential methods as column headings and questions as row headings. Intersections would be marked in some way (Todd et al., 2015).

This Guide advocates using a matrix with five columns:

- Evaluation questions
- Evaluation criteria (remembering that one question may address more than one criterion)
- The design to be used to answer the question
- The method or methods to be used to answer the question
- The sampling approach to be used for the question.



Tip

It is useful for evaluation managers to prepare their own evaluation matrix. Although preparing the evaluation matrix is the responsibility of the evaluation team, it can be useful for the evaluation manager to prepare an internal copy. This matrix should not be shared with the team that wins the bid until it has prepared its own version. The manager can compare the two versions to see if there are any major gaps.

Table 8.1: Example of an evaluation matrix for a small NGO cash transfer project

Question	Criteria	Designs	Methods	Sampling
What impact did the cash transfer have on household food security?	Impact, effectiveness	Difference in difference: comparing changes in household food security scores over time between recipient and non-recipient households*	Household food security survey	Random sampling using the initial food security assessment census as a sampling frame
		Non-experimental	Qualitative interviews with members of selected households	Purposive selection of households with highest and lowest scores in initial assessment

* See Section 11: Evaluation designs for answering evaluation questions for a description of the difference in difference design.

Note that one question may involve more than one design or method, and that the individual methods may help to answer more than one question – not shown here.

Once you have completed the evaluation matrix, this can be used as the basis for an evidence table, which can facilitate writing the report by gathering all the evidence about a particular question or theme in one place (for more on evidence tables, see [Section 16](#)).



Keep in mind

The planning process is always more important than the plan, and the team should have the opportunity to engage with the evaluation questions thoroughly in its own way.

Other formats for the evaluation matrix

Many evaluations use evaluation matrices that have only three key columns. These are:

1. **The questions to be answered** – these may be presented as key questions with subsidiary evaluation questions.
2. **How judgement will be formed** – the criteria or indicators against which the evaluation team will answer the question.
3. **Expected information sources and methods** – there are usually various sources and methods for each evaluation question. Sometimes these are split into two columns.

The following examples show extracts for the evaluation matrix for a large multi-country donor evaluation and for a single project evaluation, each of which has four columns.

An evaluation matrix of this sort shows how the team plans to answer each question, and reviewing it will allow the evaluation manager to see if the team has overlooked any major sources.

Some matrix formats are more complex. A completed evaluation matrix can be used as the basis for an evidence table, which gathers all the evidence about a particular question or theme in one place, and can be used to write the report (see [Section 16](#)). [Table 8.4](#) shows the headings of matrices for three inter-agency evaluations.

Table 8.2: Example of an evaluation matrix for a large evaluation

Core evaluation questions / sub-questions	Indicators	Analytical methods	Data sources
<p>1. How relevant and flexible is the Danish Humanitarian Strategy given the changing humanitarian context since 2010?</p> <p>1.1 Have the strategic priorities been relevant, given changing humanitarian challenges?</p>	<p>1.1a Number of strategic priorities covered by Danida-funded programmes</p> <p>1.1b Match between the strategic priorities and what Danida and its partners regard as key humanitarian challenges</p> <p>1.1c Partner anticipatory, adaptive and innovative capacities to deal with identifying and dealing with new types of threats and opportunities to mitigate them</p> <p>1.1d Evidence that Danida's funding and country-level strategies are flexible enough to enable partners to adapt to changing contexts</p>	Portfolio analysis, results tracking and comparative partner analysis to assess the coverage of the strategic priorities; Context Analysis	<ul style="list-style-type: none"> • Mapping of partner programmes against strategic priorities • Danida funding database • Partner reports • Stakeholder workshop discussion of current humanitarian challenges • Document review on international humanitarian context • Interviews with HCT and partners

Source: Mowjee et al. (2015: Annex C)

Table 8.3: Example of an evaluation matrix for an NGO project evaluation

Evaluation criteria	Key questions	Hypothesis	Sources of information
Coverage	How appropriate was the coverage of the Community-based Management of Acute Malnutrition programme?	The project has reached all the people equally within the designated areas	<ul style="list-style-type: none"> • Proposal • Assessments • National surveys • Evaluation reports • Key informant interviews

Source: Morán (2012: Annex 6).

Ideally the criteria should only appear after the key questions.

Table 8.4: Sample headings from matrices

Rwanda food for protracted refugees evaluation Sutter et al. (2012)	Bangladesh food for refugees evaluation Nielsen et al. (2015)	Evaluation of the response to the CAR crisis Lawday (2015)
<ul style="list-style-type: none"> • Evaluation questions • Sub-questions • Type of sub-question • Measure or indicators • Target or standard (normative) • Baseline data • Data source • Design • Sample or census • Data-collection instrument • Data analysis 	<ul style="list-style-type: none"> • Question type (primary or secondary) • Evaluation question • Guiding questions • Indicator category (preliminary) • Methods • Information sources • Comments or observations 	<ul style="list-style-type: none"> • Topic • Questions • Sub-questions • Judgement criteria, standards, guidelines, good practices • Sources • Methods • Analysis • Strengths/limitations

8.4 The inception report

The contents of the inception report vary depending on the context and the scale of the evaluation as seen in Table 8.5.

Some items may be included in the ToR rather than in the inception report. This depends on the policy followed by the evaluation manager. The inception report is to make the intent of the evaluation team clear to the evaluation manager. Together the ToR and the inception report should fully describe the evaluation task.

In a small evaluation the inception report may cover only some of the issues listed in the table, but at a minimum it should contain:

- The evaluation matrix (where this is left to the team to develop, as is the norm)
- The work plan (including work allocations if there is a large team)
- The data-collection tools to be used

Such a simple inception report would be suitable only for a very small evaluation.

Table 8.5: Must-haves and nice-to-haves in an evaluation ToR and inception report

Items typically included in the ToR	<ul style="list-style-type: none"> • Context • Purpose and how it will be used • Objectives • Criteria • Scope • Audience • Roles and responsibilities • Milestones • Deliverables • Contents of the inception report
Items that may be included in the ToR or inception report	<ul style="list-style-type: none"> • Evaluation frame • Evaluation questions • Sources to be used • Evaluation matrix • Evaluation designs • Data-collection methods • Indicators to be measured • Data analysis methods • Contents of the evaluation report
Items typically included in the inception report	<ul style="list-style-type: none"> • Work plan • Allocation of work within the team • Data-collection tools

The inception report should present proposed methodologies, including an initial priority interview plan for further interviews. It should acknowledge, and where possible specify the roles of, any advisory groups. An annex should include an interview guide and focus group topic list, as appropriate.

It is also useful to present any formal dissemination plan in the inception report if the evaluation team is expected to engage in dissemination. For example, if the evaluation team is expected to produce an academic paper, the inception report should identify which team members will be responsible for this.

Table 8.6: The content of your inception report

Element	Comments
Background	<p>This section should summarise the context in which the evaluation is taking place.</p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;">  Tip Include a chronology as it can be useful for setting out the background, and can be expanded for the main report with data gathered during fieldwork. </div>
Action to be evaluated	<p>This section should show that team members understand what the action to be evaluated comprises. It may consist of a chapter describing the intervention with basic data collected by the team during the desk study, and may include a ToC or some other logic model. Data tables may be included in an annex.</p>
Purpose of the evaluation	<p>This section summarises the team's understanding of the purpose and objectives of the evaluation and the use to which it will be put, and explain how this has influenced the choice of designs and methods.</p>
Stakeholder analysis	<p>A stakeholder analysis can help the evaluation team to plan the fieldwork to maximise use of the evaluation results. See Section 3: Think early and often about utilisation. The stakeholder analysis provided for the inception report of the food security cluster is an example (Steets et al., 2013) as is the power and interest table provided in the CAR inter-agency evaluation inception report (Lawday et al., 2015: 14).</p>
Evaluation questions	<p>This section should include any drafting or redrafting of the evaluation questions that the team proposes. For example, it may propose reworked questions to reduce the total to a manageable number, or to focus on the key issues. These are summarised in the evaluation matrix. See Section 6: Choosing evaluation questions.</p>
Evaluation design	<p>While these are summarised in the evaluation matrix, this section presents the reasoning behind the choice of design(s). Currently relatively few EHA evaluations identify their designs. The evaluation of food assistance to protracted refugees in Bangladesh (Nielsen, 2012: 5) is an exception. It identifies itself as a <i>Post-facto non-equivalent comparison group design</i>. See Section 11: Evaluation designs for answering evaluation questions.</p>
Methods	<p>These are summarised in the evaluation matrix, but the methods the team proposes to use to collect and analyse data to answer the evaluation questions are presented in greater detail here. See Section 10: Desk methods and Section 13: Field methods. The details for each method should indicate the intended sampling strategy and the proposed approach to analysis. It should clearly state the limitations of the proposed data-collection methods, including the sampling strategy, and any limitations related to the resources available. See Section 12: Sampling.</p>



Table 8.6: The content of your inception report

Element	Comments
Data-collection and analysis tools	<p>These are usually annexed to the inception report and include any interview guides, survey forms, rubrics, or other data-collection instruments to be used. The data-analysis instruments should also be presented, including any initial lists of codes.</p> <p>See Section 16: Analysis.</p>
Evaluation matrix	<p>This shows how the evaluators plan to answer each of the evaluation questions. It should reflect the designs and methods set out in the inception report. It may be presented in an annex. While the text of the report details the designs and methods, the matrix shows which the team plans to use for each question.</p>
Detailed work plan	<p>This specifies where team members plan to visit and when, and the days proposed for head office visits. It should also indicate the responsibilities of each team member.</p> <p>See Section 7: Terms of Reference and budgeting for an evaluation and the discussion on evaluation timelines in Section 9.</p>
 Tip	<p>Avoid specific plans in more insecure environments. This is because it can be a security risk to indicate specific travel plans in advance, and last-minute flexibility is often required in such environments.</p>
Main report layout, and format for other products	<p>This usually takes the form of a table of contents and may also include details of other evaluation products, such as the rough outlines for dissemination workshops, and any other evaluation products. It may also include a detailed dissemination plan.</p> <p>See Section 17: Reporting and communicating evaluation findings with a utilisation focus.</p>
Interview targets	<p>This provides a preliminary list of the people the team intends to interview, or at least the types of people to be interviewed.</p>
Outstanding questions and issues	<p>This is an opportunity to highlight ambiguities, areas of concern, or contradictions that the evaluation team would like the commissioning agency to address and clarify before the next stage.</p>
Risks and mitigation measures	<p>The team identifies what risks it foresees and how it plans to take to minimise them.</p>
Ethical issues	<p>The team sets out how it will approach any ethical issues in the evaluation.</p> <p>See Section 14: Engaging with the affected population in your evaluation.</p>

Assessing the inception report

The inception report allows the evaluation manager to see how the team understands the evaluation task and plans to approach it. It also allows stakeholders in the field to see how the team plans to do the work, so they can identify any relevant issues or challenges.

The report also allows the evaluation manager to address issues with the team's understanding and approach before they become major problems. It should demonstrate the team's understanding of the context of the humanitarian crisis, the context of the response and of the actions to be evaluated, the purpose and intent of the evaluation, and the concerns of stakeholders.

An inception report could be assessed against [Table 8.6](#) of possible contents and consider:

- Should this inception report contain such a section?
- Does the section clearly set out what is planned?
- Is what is set out adequate and appropriate?

The answers to these questions will indicate the areas in which the evaluation manager needs to engage in further discussion with the evaluation team.



Tip

Aim to specify what you want. If you are expecting the team to cover particular aspects in the inception report, be sure to specify these in the ToR.

9 / Planning and managing your evaluation

This section sets out various management and governance arrangements for an evaluation. It also discusses evaluation timelines, team recruitment and leadership issues.

The scale and complexity of the evaluation management and governance arrangements should be in proportion to the scale and complexity of the evaluation. Large joint evaluations tend to require the most complex governance arrangements.

9.1 The decision to evaluate

The evaluation manager's task begins when the decision is made to conduct or consider conducting an evaluation see Section 2: Deciding to do an evaluation.

This decision may be taken at a project's design stage, during implementation, or be determined by the agency's evaluation policy. An evaluation policy may make it mandatory to evaluate programmes of a certain size or complexity.

In other cases the programme manager, senior management, or the evaluation manager may propose an evaluation, either in line with a broader evaluation policy or on an ad hoc basis.

The positive bias of evaluation

Programme managers tend to favour evaluations of programmes that are considered to have been successful. There may be some resistance to evaluating those that have been seen as unsuccessful as managers may fear that such evaluations may have a negative impact on their career. This introduces a bias towards positive results into evaluation activity.

Evaluation managers, if they are responsible for broader corporate learning, may favour evaluations of programmes seen as relatively unsuccessful because of their potential to generate lessons. In such cases evaluation managers may have to lobby for support for the evaluation.



Tip

A field visit by the evaluation manager prior to the evaluation can be useful for fine-tuning the ToR and for easing any concerns about the evaluation from field-based stakeholders (see discussion on evaluability assessments).

9.2 Advisory groups

The best arrangements for managing an evaluation ensure that primary stakeholders remain engaged in its decisions. Advisory groups are often formed for this purpose. The most common types are discussed below. While a large evaluation may use all of these advisory groups, a small evaluation may use a far simpler management structure, with an advisory group consisting of a few evaluation colleagues from similar agencies for the steering and peer reference functions, and the evaluation manager for the steering and management functions.



Tip

Rather than setting up an advisory group for a single evaluation consider establishing a group of evaluation colleagues from similar agencies to serve as a standing advisory group for each other's evaluations. This may also be an opportunity to establish partnerships with academic institutions.

Steering group



Definition: Steering group

A group established to steer an evaluation through key stages such as establishing the ToR, writing the inception report, and drafting the final report.

In a large evaluation, and especially in a joint evaluation, a steering group typically consists of evaluation managers drawn from a number of agencies. Strong leadership is essential to enable the group to work effectively and to its full potential, especially if membership is diverse. In a small evaluation the evaluation steering committee may consist of the evaluation manager and one or two colleagues. The steering group meets only at the key stages of the evaluation.

The evaluation of the response to the 2002-2003 emergency in Ethiopia was conducted by a steering committee with representatives of the government, donors, the UN, and NGOs (Simpkin et al., 2004: 4). For the evaluation of UNHCR's age, gender and diversity mainstreaming policy, the steering committee of UNHCR staff, governments and NGOs 'met three times in Geneva. It reviewed and validated the evaluation methods, reviewed interim reports, and provided feedback on conclusions and recommendations' (Thomas and Beck, 2010: 10).

Steering committees are also used for small evaluations. The CARE-Save the Children evaluation of the Haiti Earthquake Response had a steering committee comprising one representative of each agency (O'Hagan et al., 2011 37). In this case the steering committee selected the sites for field visits (*Ibid.*: 13). While most common in joint evaluations, steering committees can also be used for single-agency evaluations. The NRC's evaluation of Information, Legal Counselling, and Advice in Afghanistan and Pakistan has a steering committee of five, comprising the Regional Director and four staff from the head office (Pierce, 2009: 10).

Management group



Definition: Management group

A group that manages the evaluation on a day-to-day basis, including drafting the ToR, contracting and managing the evaluation team, and managing the review and finalisation of the evaluation report.

For a large steering group, it is a good idea to establish a smaller management group that can take decisions quickly when necessary without needing to convene a full meeting of the steering group. Typically the management group meets more often than the full steering group. Smaller evaluations may use a management group instead of a steering committee. A small evaluation may be conducted by the evaluation manager without the need for a steering committee or group.

The three-month RTE of the response to the 2010 Haiti earthquake had a management group with representatives of OCHA, UNICEF, and an NGO (Grünewald et al., 2010: 19). In the UNICEF RTE of the response to the Mali Crisis, the ToR required the management group to meet weekly to review progress (Leonardi and Arqués, 2013: 90). The structure of the evaluation report was agreed between the team and the management group (*Ibid*: 24).

In small evaluations, the evaluation manager may perform all of these tasks.

Reference group



Definition: Reference group

A group made up of primary stakeholders familiar with the local environment who can advise on practical issues associated with the evaluation and on the feasibility of the resulting recommendations.

While the terms ‘steering committee’ and ‘management group’ are used fairly consistently, this is less true of the terms ‘reference group’ and ‘peer reviewers’. One agency’s reference group is another’s peer-review group. The UNEG evaluation standards refer to ‘a peer review or reference group, composed of external experts’ (UNEG, 2005: 15). WFP makes the distinction between an internal reference group and external peer reviewers. The Strategic Evaluation of WFP’s contingency planning had an internal reference group of six and an external peer-review group of three (Ressler et al., 2009: ii). The report for the

evaluation of Danish engagement in and around Somalia (Gardner et al., 2012: 18) was revised after comments from the reference group.

Establishing a reference group is a good way to involve the primary stakeholders when an evaluation has been commissioned by the agency's head office. The UNICEF evaluation of the East Timor education programme had a reference group of key stakeholders identified by the country office (Tolani-Brown et al., 2010: 93)



Good practice example: Making the most of involving the country office in evaluation

To make the exercise as useful as possible at the country level, the UNICEF Evaluation Office sought to involve the country office at all stages of the evaluation through the formation of a local reference group. The evaluation manager went to Timor-Leste and involved the country office M&E officer in discussions regarding sampling strategy and survey design, and offered support with evaluation quality assurance to boost the overall quality of the evaluation. This element of capacity development served as an incentive for the M&E staff actively to support the education programme and ensured they had a stake in its success. The result was a high-quality, relevant evaluation that led to a strong country-office-led management response and follow-up process. As a consequence of this independent yet collaborative approach, the donor committed a significant second tranche of multi-year funding for the programme. This process also changed the culture in the country office, which became much more supportive of evaluation.

Source: UNICEF (2010)

Peer-review group



Definition: Peer-review group

A group that advises on quality issues, usually made up of evaluators and other specialists chosen for their knowledge of evaluation, the region, or the type of intervention being evaluated.

A peer-review group does not need to be large. It may advise on the contextual analysis, design and methods and comment on the draft report. It can be especially useful in evaluations that are managed by a general manager rather than an evaluator and when the evaluation team is unfamiliar with the region. The DEC review of the response to the 1999 Kosovo Crisis used a peer-review team of regional and humanitarian experts who briefed the evaluation team before travelling and later discussed the draft report with them (Wiles et al., 2000: 4). The evaluation of FAO's cooperation in Somalia noted that the two peer reviewers had provided 'constructive and insightful inputs and feedback' (Buchanan-Smith et al., 2013: v). The evaluation was also subject to a peer-review process within the evaluation department (*Ibid*: 4).

The number of peer reviewers varies. The Oxfam GB evaluation of urban food security and livelihoods in Nairobi (Macauslan and Phelps, 2012) had one external peer reviewer. The DEC evaluation of the 2000 Mozambique floods had three peer reviewers (Cosgrave et al., 2001 p.1) to whom the team leader circulated the draft evaluation report as well as to the other team members (*Ibid*. p.107). The peer reviewers also reviewed the final draft.



Tip

Keep the peer-review and reference groups separate so that each can focus on its assigned task and avoid distraction. While their responsibilities may appear to overlap, they function better separately.



Tip

Include an allowance in the evaluation budget for paying for peer reviewers who are working in their own time rather than as part of any paid employment.

Large joint evaluations will benefit from having a full range of advisory groups, but simpler arrangements are more suitable for smaller evaluations. Even the smallest evaluation can benefit from having a reference group of stakeholders and even a single peer reviewer.

9.3 Evaluation timeline

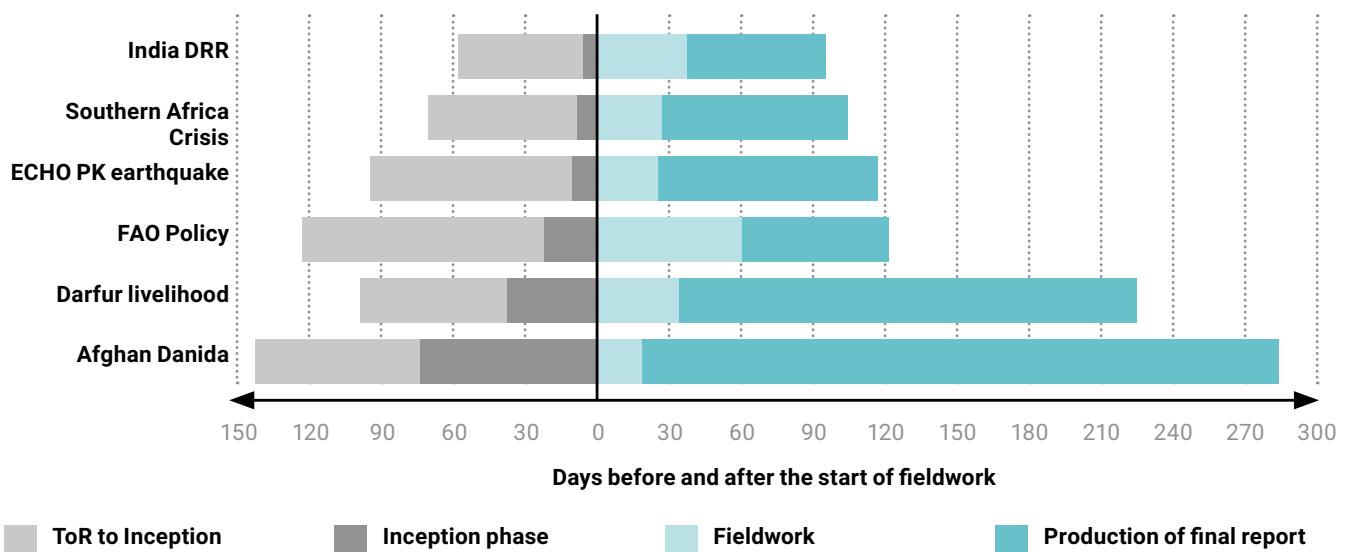
How long an evaluation should take depends on the context.
The following table gives some estimates.

Table 9.1 : Estimated timeline of an evaluation

Activity	
Developing the ToR	One week to over a year depending on how much consultation and agreement is needed. The IASC has developed a ToR template that can quickly be adapted (IASC, 2010). This approach is particularly appropriate when a series of similar evaluations is planned, or for an agency with a particular focus (e.g. a child-centred focus or agency benchmarks).
Publication of the ToR to decision to award	This can be quite fast in an internal evaluation if staff is already available. It takes longer when external consultants are contracted. Graig (2013) suggests that three to four weeks is enough time for evaluators to submit an expression of interest, but that it may take longer if there is a broad request for expressions of interest. It is more common to see this and the contract award taking two months.
Contract formalities	One day to several months, depending on the agency.
Inception phase	One week to several months or more. Longer inception phases are needed when there is a need for an inception visit and/or an extensive document review.
Fieldwork	One week (for the smallest evaluations) to several months, depending on the scale and complexity. Two to four weeks is the most common for EHAs.
Data analysis and report writing	One week to several months, depending on the complexity. An analysis of survey data can take several months if there is a need to translate collected data or to code data.
Revision cycle	Each revision cycle consists of a period for comments, typically two weeks for a small evaluation to four weeks or more for a complex evaluation where agencies need to consult their field staff. This is followed by one or two weeks to revise the draft. Almost every evaluation plan calls for one revision cycle, but there can be many more, especially when the evaluation results are less welcome, or there is a need to build consensus.
Dissemination	This may consist of workshops or field visits to disseminate the results, usually within one month of the publication of the report.

The following chart gives the timelines for eight evaluations of different types in which the authors were involved. Most of these were medium-sized evaluations. Smaller evaluations were faster than larger evaluations.

Figure 9.1: Timeliness of evaluations with periods in days before and after the start of fieldwork



There were particular reasons why the different phases took so long in each case. The illustration is therefore to show the wide variation and not to present norms.

9.4 Internal and external evaluation

An internal evaluation is carried out by the staff of the agency being evaluated. Some agencies use the term internal and external to refer not to the organisation but to the programme, for instance when an evaluation that is conducted by agency staff who are not involved in the programme is referred to as an external evaluation. An external evaluation is conducted by an evaluation team that is external to the agency being evaluated.

Some agencies have evaluation departments whose staff have unique career paths and report directly to the board rather than to the programme officers. Such agencies would argue that in such circumstances even an internal evaluation can be independent.

In some cases an evaluation may be conducted by a joint team of external consultants and internal staff. This is a mixed evaluation. If the person leading the evaluation is external, it is regarded as an independent evaluation. The combination of an external lead with internal staff can facilitate learning.

The decision to use an external, internal, or mixed evaluation team depends on the purpose of the evaluation (see [Section 2: Deciding to do an evaluation](#)). If the evaluation is principally for accountability, the evaluators should be external, because internal evaluators may not be seen as sufficiently independent. Depending on the organisational context, it may be desirable to include one or two staff members on the team – perhaps from the agency's evaluation department if there is one, or, if appropriate, from the operational departments. These members should never be in the majority or take the leadership role, since including internal staff on the team can jeopardise the perceived independence of the evaluation.

If the evaluation is principally for learning, the evaluation team should include either a majority of internal staff who are expected to do the learning, or a team of external evaluators whose primary role is to facilitate staff learning. Mixed teams can help to close the learning loop as the agency's own staff will have insights into how to best to implement its organisational learning. For example, they may be aware of key stakeholders that should be involved and of organisational areas where there is openness to change (Riccardo Polastro, personal communication, June 2013).



Key question

External contractors are seen as being more independent than agency staff of management pressures to present a programme in a favourable light. This is true as a general rule, but where a particular evaluator or evaluation firm expects to do further business with a particular client, what impact might this have on their independence?

The pros and cons of using internal and external evaluators are presented in [Table 9.2](#) on the next page.

Table 9.2: Advantages and disadvantages of internal and external evaluators

Internal evaluators		External evaluators	
+ Benefits they derive from learning and reflection during the evaluation process remain within the organisation		+ They are often more objective	
+ They know the organisation and its culture		+ They are less likely to have organisational bias	
+ They are known to staff		+ They bring fresh perspectives	
+ They may be less threatening and more trusted		+ They may have broader experience to draw on	
+ Findings and recommendations may be more appropriate for the organisation		+ They may be able to commit more time to the evaluation	
+ Recommendations often have a greater chance of being adopted		+ They can serve as outside experts or facilitators	
+ They are less expensive		+ They are not part of the organisation's power structure	
+ Builds internal evaluation capability and generally contributes to programme capacity		+ They can bring in additional resources	
- Their objectivity may be questioned		+ They are likely to be trained and experienced in evaluation	
- Organisational structure may constrain participation		+ They are regarded as experts	
- Work commitments may constrain participation		- They may not know the organisation	
- Their motivation may be questioned		- They may not know the constraints that will affect uptake of recommendations	
- They may too easily accept the organisation's assumptions		- The benefits they derive from reflection and learning during the evaluation process do not remain within the organisation	
- They may not be trained in evaluation methods		- They may be perceived as adversaries	
- They may reduce the evaluation's credibility outside the organisation		- They are more expensive	
- They may have difficulty avoiding bias		- Hiring them may require time-consuming contract negotiations	
- They may lack specialist technical expertise		- They may make follow-up on recommendations less likely	
		- They may be unfamiliar with the environment	
		- They may be influenced by the need to secure future contracts and thus be less independent than they appear	

Authors' compilation. See Gosling and Edwards (2003) and CRS and American Red Cross' Short Cut on Preparing for an Evaluation (2008).

Stakeholder learning in evaluations

Carrying out an evaluation provides the evaluation team with great learning opportunities. If stakeholder learning is the priority, an external evaluation may not be the best approach. Internal evaluations allow agency staff to capture the learning – making it more likely that the lessons will be used. Mixed evaluations, possibly led by an experienced external evaluator with some agency staff, provide a good opportunity for learning while maintaining quality and credibility.

Another approach is the peer-evaluation model. This can be done with either:

- Other partner agencies participating in an evaluation of one partner.
- Programmes from different countries participating in the evaluation of a similar programme in another country.

This approach ensures that those learning from the evaluation are people in a position to apply the learning in other contexts. Again, employing an experienced external evaluator helps to ensure the quality and credibility of the evaluation.

9.5 Contracting evaluators

External evaluators must be contracted. Evaluation managers have a number of options for doing this:

- Issuing a formal request for proposals with the evaluation details as an attached ToR or as part of the request for proposals (RfP).
- Issuing a ToR that contains the contract details.
- Issuing a letter of intent stating the agency's interest in commissioning an evaluation.

Some agencies use the term 'letter of intent' to mean an expression of interest. Here, we use it to refer to the evaluators' submission to an agency stating their interest in carrying out an evaluation.

Davidson (2012) and Graig (2011; 2013) argue for the use of an expression of interest process to select the evaluator or identify potential contractors. Using an expression of interest to select evaluators is particularly appropriate for smaller evaluations. Davidson (2010b) suggests that those submitting an expression of interest should be asked:

- What or who is the entity/person/team planning to bid on the evaluation? (maximum one page)
- Why they are interested in this evaluation? (maximum half a page)
- What they think you can do? What relevant expertise, experience, and capacity do they have? (maximum one page)
- Who they are – an evaluation firm, an evaluator, or an evaluation team? What are their distinctive values, practices, and areas of expertise and specialisation? (maximum half a page)
- What are the daily rates of the proposed team members, and what other overheads or incidentals would they require?
- If possible, provide two or three executive summaries from recent evaluation reports the lead evaluator(s) have conducted/led.

Davidson further suggests that the evaluators should not be asked for CVs or other extras at the screening phase. CVs may be appropriate where a contract is going to be based exclusively on expressions of interest, as may be the case for smaller evaluations.

Davidson notes that the standard request for proposals ends up with detailed bids full of mundane details, with little to distinguish one from another. She suggests that evaluators should be selected not on their evaluation plan but on their capability to handle the key challenges. The possible areas of questioning include:

- If this is an evaluation as opposed to a research project about the topic, ask how they define the difference. What knowledge and skills do they think they need to apply to an evaluation rather than a research project?
- How would they go about answering a question like 'how valuable are the outcomes for [a specific recipient group]?'
- How would they manage differing perspectives on the most valuable outcomes and know what quality looks like? How would they apply them appropriately to drawing conclusions?
- How would they handle any challenges anticipated in the evaluation? Ask for examples of how they have done so in the past.

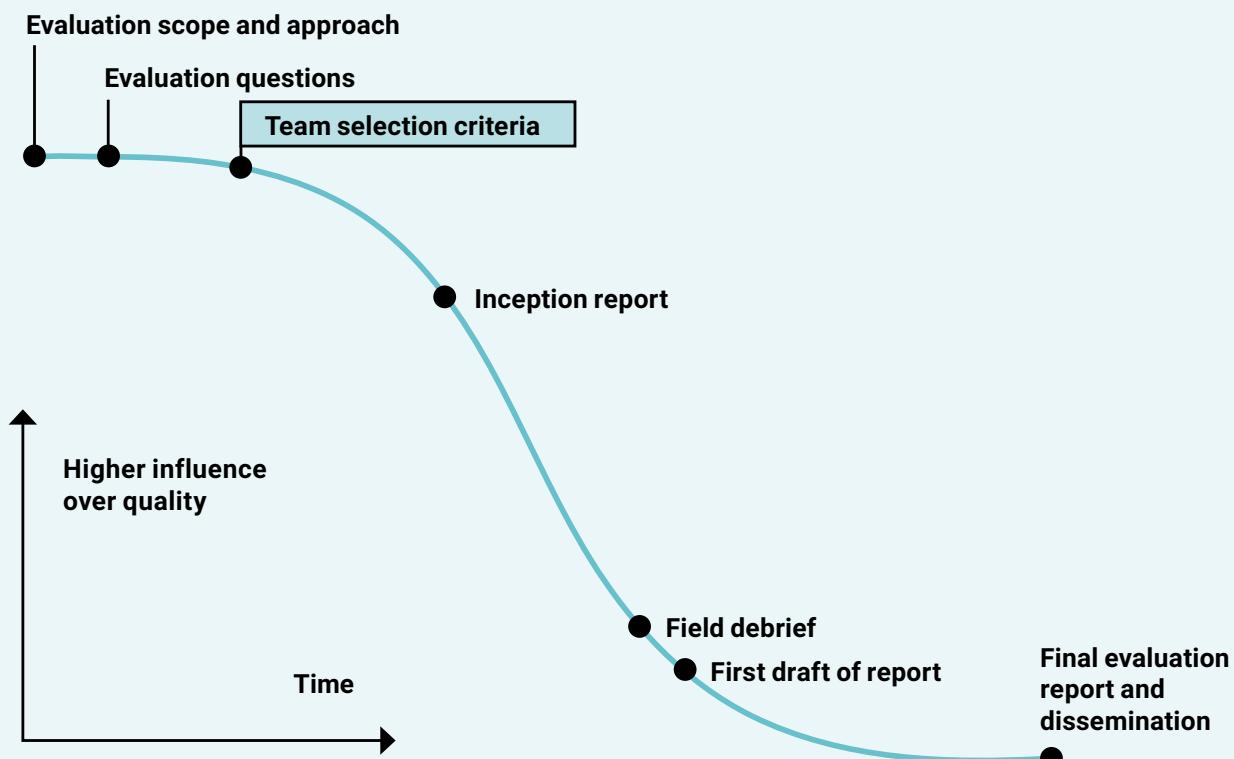
- How would they describe themselves as professional evaluators? What is their ‘brand’? What is their ‘signature approach’, for which they are best known? What kinds of project play to their strengths? What kinds of project do they avoid because they are outside their areas of strength? Probe their evaluation expertise and approach; don’t let them get away with an answer about content expertise.
- Ask for three recent executive summaries for evaluations they have completed. These speak volumes about how well the team can write, get to the point, be succinct, answer important questions evaluatively (not just associate them with data), and how well they truly understand intended users and what will make sense for them (Davidson, 2012b).

Davidson (2010c) suggests that these questions could be answered in writing or in a presentation. These or similar questions could form the basis for an interview (in person or electronically) with the team leader. In Davidson’s approach, this interview would command a good proportion of the technical points.

Getting the right external team

Defining the selection criteria for the winning bid will have a strong influence on the quality of the evaluation. These may be set out in the ToR or in a separate RfP.

Quality checkpoint: Team selection criteria



The team selection criteria are critical for ensuring the selected team has the skills necessary to conduct the evaluation.

The selection criteria should include:

- The necessary context knowledge
- The necessary communication skills
- The skills to conduct high-quality fieldwork
- The analytical skills to turn the field-based evidence into convincing answers to the evaluation questions.

Qualitative award criteria	Max. points
Understanding of the ToR and the aim of the services to be provided	10
Methods for structuring, data collection and analysis	40
Organisation of tasks and team, timetable	10
Appropriateness of the team on the basis of the expertise proposed	40
Total technical points	100
Price factor: cheapest tender/this tender	
Total score: total technical points multiplied by price factor	

Based on Maunder et al., 2015

The following is an example of the bid award criteria commonly used in European Commission evaluations, in which 40% of the technical points are for the team composition.

Without a further breakdown on the specific meaning of each criterion, the selection is still somewhat subjective. The following example (loosely based on a Swiss Solidarity evaluation in Haiti, see Groupe URD, 2013) is a more structured approach.

Aspect	Criteria	Points
Team leader	Significant experience of humanitarian evaluation, using a range of methods	15
	Conversant in French or Haitian Creole	5
	Excellent writing skills in French or English	10
	Extensive experience in Haiti an asset	5
Whole team	Experience of evaluating recovery or development interventions	5
	Strong Haiti expertise including a thorough understanding of the economic structures, institutional set-up and the social fabric of the country	10
	Strong French language skills	10
	Ability to analyse quantitative survey data (this requirement may be met by a team member who does not travel)	5
	Experience in capacity-building interventions or in evaluating such programmes	5
	Experience in livelihood programmes or in evaluating livelihood interventions	5
	Experience in DRR programmes	5
	Substantial Haitian Creole language skills an asset	5
Approach	Extent to which the proposed approach demonstrates a good understanding of the risks and constraints	10
Bidder	Assurance of stable team composition over entire evaluation period	10
	Experience of similar evaluations	10
Price	Score = $15 \times (\text{cost of cheapest qualifying bid}) / (\text{cost of this bid})$	15
Maximum possible score		130

This is a much more detailed set of criteria and makes bids much easier to score. Most of the technical points (85/115, or 74%) are for the team skills. While you might not want this much detail, it is a good idea to specify what skills required both of the team leader and of the team. This example could be further improved by providing examples of high scores and low scores such as:

- Ability to greet in Creole, 1 point, ability to interview in Creole without assistance or misunderstandings, 5 points.
- Experience of Haiti, 1 point per year of experience.
- French language skills: ranging from ability to 'get by', 1 point; to mother tongue, 10 points.

Where a team includes a number of people, they can be scored individually and the score can be averaged, unless it is sufficient for only one member to have the requisite skill.

Aspect	Criteria	Total Points	Min to qualify
Team leader	Experience and skills for the team leader, broken down into key elements such as evaluation leadership experience, geographic experience, sectoral or organisational experience etc. For greater transparency it might be worth including a scoring rubric for the individual elements, (e.g. experience of leading humanitarian evaluations – 1 point for every two years to a maximum of 7 points, or language skills (e.g. mother tongue 4, fluent 3, semi-fluent 2, basic 1).	20	15
Team leader interview	Interview to establish team leadership and interpersonal skills.	20	15
Other members	Experience and skills for the other team members, broken down into key areas of experience and skills.	30	20
Context	Demonstrated understanding of the context and of the key problems involved in the evaluation.	10	5
Approach	Proposed approach to the evaluation task (this could be included in the team leader interview instead).	20	10
Total technical score/minimum score to qualify		100	65



Tip

Not all criteria are equally important. Consider which of your selection criteria are essential to ensure a good evaluation.

It is possible to apply essential criteria by applying a minimum score to each relevant criterion. For example, it may be decided to consider only team leaders who score at least 10 out of 15 points for previous evaluation experience. Where there is a selection interview, a high score for the interview (reflecting interpersonal skills) might be an essential criterion. Favourable references from previous clients might also be an essential criterion.

Points for context and approach could be split between the expression of interest and the interview with the team leader where the award is made on this basis rather than on a full bid.

See the following sub-section on [contracting procedures](#) for a discussion on the impact of the issue of cost and why the least expensive tender should not automatically be selected.¹



Tip

Getting the right consultant(s) for an evaluation is important and takes time. The further ahead you recruit, the more choices you will have.

The type of consultant varies with the evaluation task. In order to deal with sensitive issues of partnership consultants will need an understanding of institutional relationships in general and partnerships in particular. A RTE requires consultants with sufficient operational experience to be credible with the field staff. Sector-specific evaluation may call for evaluators with experience of working in that sector. These requirements should be reflected in the selection criteria for the evaluation team.

Considerations in choosing consultants include:

- **How well the consultant knows the organisation.** A consultant who already knows the organisation will need to spend less time learning about it but may be less inclined to challenge its dominant beliefs.
- **How well the consultant knows the context.** This is especially important for complex crises and conflicts. Depending on the evaluation, it might be more important to understand the context or the type of activity to be evaluated.

- **Whether the consultant has the requisite skills in and experience of the proposed evaluation designs and methods.** This can include knowledge of sectoral areas (such as WASH) and appropriate theory (such as organisational theory and process engineering for evaluations of organisational processes).
- **The overall team composition.** Larger teams may offer more skills but may also cost more, require greater logistical support in the field, entail a bigger risk of internal disagreement, and take longer to complete a report.

Another issue is whether to recruit consultants individually or as a team via a consulting company. The pros and cons of each choice are summarised in Table 9.3. In many cases the choice will depend on organisational policies, or their implications for timeliness.²

Table 9.3: Advantages and disadvantages of using individuals or a contractor's team

Individuals		Contractor's team	
+ You select the people and skills that you want	+ There is only one contract to manage		
+ You can ensure that all team members have something to offer	+ You have more leverage on quality issues		
+ May be less complex than a contract	+ Resolving team conflicts is the contractor's responsibility		
+ You have a direct relationship with everyone on the team	+ You have to deal with only a few payments		
- This option is usually less expensive	- The contractor assumes some of the risks		
- Financial limits may prevent you from hiring the most skilled consultants	- This option is usually more expensive		
- Resolving team conflicts is your responsibility	- The team may have only one senior expert supported by junior staff		
- You have to provide human resources support to the team	- Some team members may never have worked together before		
- You have to organise logistics in the field	- You may still have to deal with field logistics and other tasks if the contractor is not competent		
- You have to deal with many different payments	- Financial limits may preclude hiring a team in a timely manner		
- You bear all the risks			

Source: Authors' compilation.

It is sometimes possible to select consultants on the basis of prior experience, but this luxury is not always an option. It is possible to ask for short notes on the approach they would take, but it is preferable to select consultants on the basis of their previous work. Who wrote the evaluation reports that you consider of good quality?



Tip

Don't rely completely on a potential consultant's authorship of a previous evaluation – contact the evaluation manager and ask about the consultant's work record. The final report may have been the 19th draft and may have been completely rewritten by another team member.

The contractual relationship will depend on the agency's recruitment policy. It is common practice to link payments to stages in the evaluation, such as signing the contract, acceptance of the inception report, completion of the fieldwork, and acceptance of various drafts and the final report. Davidson (2012a) notes that the terms of the contract should allow the evaluation manager to terminate the contract if the evaluation team prove to be incapable of meeting their obligations.



Tip

Specifically include in the terms of the contract the right to end the evaluation in the inception phase if the inception report is unsatisfactory.

Contracting procedures

If evaluators are asked to submit bids the standard practice is to ask for separate sealed technical and financial tenders. The technical bids are opened first, and points are awarded for each bid in this regard (see sub section on getting the right external team above for an example of criteria that might be specified). The evaluators may then be awarded points for the interview. After technical points have been assigned, the financial bids are opened and the total score is calculated.

If the evaluation is awarded on the basis of an expression of interest and an interview with the team leader, candidates can be asked to submit a sealed financial bid, to be opened after the interview. This approach is particularly suitable for smaller evaluations.

There are three main ways in which contracts can be awarded:

- On the lowest price for any technically qualified evaluation. Bids that obtain a defined minimum level of technical points and are not disqualified by any key criteria (e.g. obtaining no points for an essential criterion such as evaluation experience) are regarded as technically qualified. This is a good approach for procurement or simple services, but not for selecting a service that demands a high level of complex technical skills, as evaluation does.
- On the technical score plus a financial score to differentiate between bid A and the lowest qualifying bid. This is probably the best approach as it gives the greatest emphasis to the skills of the evaluation team.
- On the technical score multiplied by a financial score for the difference between bid A and the lowest qualifying bid.

Multiplying the full technical score by the price makes the selection price sensitive. In this case using the first option makes Bid 2 the winner even though it scored only 73% on technical points versus 90% for Bid 1.

Restricting price to only some of the points is a better approach and is used by UNICEF and the Scandinavian donors.

Element		Bid 1	Bid 2
Percentage of max tech points	A	90%	73%
Relative price (to cheapest)	B	1.25	1
Price factor (1 / B)	C	0.8	1
EU formula (A x C)		72%	73%
Tech score (80% by A)	D	72%	58%
Price score (20% x C)	E	16%	20%
UNICEF formula (D + E)		88%	78%

9.6

Leadership and teamwork challenges

Like any other project, an EHA needs to be well managed to be successful.

This is why Grieg (2011: 3) includes strong project-management skills as a key skill for evaluators, as follows:

1. Technical competence
2. Strong project-management skills
3. Excellent thinking skills
4. Excellent ability to communicate with stakeholders
5. Flexibility
6. An orientation towards collaboration.

The last three are all interpersonal skills. For EHA, technical competence includes knowledge of humanitarian principles and of the humanitarian system, as well as the ability to work in stressful or insecure environments.

There are potentially three levels of management: management by the evaluation department of the commissioning agency (the evaluation manager), management by the contract manager when a consultancy company is used, and management of the process and the team by the team leader. Each of these levels has different concerns. This sub-section focuses on the team leader's management of the EHA process and of the evaluation team.

Even when evaluators are competent, resources are adequate, and objectives are clear, good planning is still essential to the success of an evaluation. Almost all EHA faces time constraints. Agreeing the ToR, recruiting a team, writing and circulating the draft report, and incorporating reviewers' comments almost always take longer than expected; fieldwork often suffers most when there are delays or the evaluation goes over budget. Careful planning can help to avoid this.



Tip

Clarify from the start what support the commissioning agency will provide to the evaluation team. Will the team have access to files, working space in the head office and in the field, help with appointments at the head office and in the field, security services, transport and assistance with booking accommodation in the field?

The scale of the evaluation task determines the size of the team. As noted earlier, however, larger teams bring problems and risks that may outweigh the benefits of their wider range of expertise. Larger teams mean more work for the team leader, for example in managing assignments and collating inputs. They also pose a problem in insecure environments. Comfortable accommodation may be hard to find in some settings, and larger teams may need to split up, making it difficult to have informal discussions outside working hours.



Tip

Establish templates for documentation like persons-met lists, bibliographies, and itineraries, formatted in the way you want to receive the information. Getting inputs in a standard format minimises the work of converting and collating them.

EHA teams are usually assembled either by a consultancy company or by direct hire. In either case, the team leader may not have worked with all the other team members before.



Tip

If you are working with some evaluators for the first time, and don't know them by reputation, organise the fieldwork so that you spend a day together at the start of the evaluation and can get a sense of their strengths and weaknesses, including any potential biases.

Even the best evaluators have 'off days', and previous good performance is not a guarantee of good performance in the current evaluation. A team member may be affected by personal issues, such as the illness of a close relative. Team members who are suffering personal difficulties may need sensitive support from the team leader. The team leader should, however, set deadlines for tasks and follow up immediately if these are not met. This can help to identify and address any performance issues quickly.

Sometimes problems occur due to personality conflicts, performance issues, or differing values. Some EHA environments can be stressful. Of these problems, performance issues are the easiest to deal with – although by the time this becomes clear, it may be too late to do without or replace that person without abandoning the evaluation.



Tip

If a poor performer cannot be dropped from the team, pair that person with a stronger team member to help minimise quality problems.

Large teams are more likely to experience conflict regarding the findings, whether because of differences in perspective, in perceptions of the context, or between the areas on which team members have focused. Tried and tested ways to manage large evaluation teams include the following:

- Take time at the beginning of the fieldwork for the team to get to know each other and to clarify roles and expectations. Consider ways for everyone to share their area of expertise. Even when time is at a premium, it can pay dividends to invest in the team culture and communication at the outset.
- Ensure that all team members share the same accommodation to enable informal discussions and socialising.
- Use an evidence table (see Section 16: Analysis) to keep track of evidence and emerging findings, and share this to keep team members informed and build a common view.
- If there is a large team in a single country, incorporate some time for reflection into the work programme so that the team can discuss emerging findings.
- Plan for team meetings after fieldwork and before the completion of the first draft of the report to discuss the evidence, findings, conclusions, and recommendations.

Sometimes a team member may be especially concerned about a particular issue that, though relevant, is peripheral to the main focus of the evaluation or is too narrow or too complex for inclusion in detail in the main report. In these circumstances, the team member might be asked to write an annex on the issue.

9.7 Managing conflict

Disparity between the scale of the evaluation task and available resources may lead to conflict between the evaluation manager and the evaluation team. Many such disagreements stem from differences in understanding. An inception report can reduce the risk of misunderstandings at an early stage, but conflict may also emerge at the report stage, leading to multiple revisions and increasing frustration on both sides.



Tip

One way to minimise disputes in contentious evaluations is to use a steering group to advise on the acceptability of draft reports (although the final decision rests with the evaluation manager). Steering groups can also help ensure the quality of the whole evaluation process.

Personality clashes are another potential source of conflict. It may be useful to have a formal dispute-resolution policy specifying what steps would be taken in the event of a conflict and identifying someone who would arbitrate if necessary.



Good practice example: Establishing a dispute-resolution policy

The Tsunami Evaluation Coalition adopted a policy on resolving disputes in the evaluation teams, between a team and the steering committee, in the synthesis team, and between the synthesis team and the core management group. This policy stipulated that the main author was responsible for managing relations within the team. When there was a serious dispute about a substantive issue between team members, or between a team member and the team leader, that the main author was not able to resolve, the core management group was to ask the head of the ALNAP Secretariat to prepare a report giving both sides of the issue. The head of the ALNAP Secretariat could prepare the report or contract an experienced evaluator to do so.

The management group might then decide to ask the main author to do one of the following:

- Present only one of the interpretations
- Include both interpretations in the report
- Include only one interpretation, but note that this was not unanimously held.

If members of the synthesis team were unhappy with the resolution of the problem, they had the right to have their name removed from the report.



Tip

Formal dispute-resolution policies are appropriate not only for large-scale evaluations but for any evaluation that needs them.



10 / Desk methods

This section focuses on the use of desk methods in an evaluation, and discusses how much time should be spent on the initial desk review. It also presents several ways in which to present the data from a desk review or desk study in a succinct manner for the evaluation team and also for evaluation briefings and reports.

Desk reviews and desk studies use desk methods in order to summarise documents.

The evaluation questions determine the desk methods

As with field methods, the breadth and depth of a review, and the choice of desk methods are determined by the evaluation questions. Some evaluation questions can be answered only by desk methods and others can best be answered by them. For example, an evaluation question about whether an intervention is meeting targets might best be answered by reviewing monitoring reports and then triangulating the data with field interviews.

See [Section 6](#) for more detail on evaluation questions.

10.1 What is a desk review?

A desk review is a review of one or more documents. It can take place:

- As part of the inception phase to clarify the evaluation task or to answer specific questions
- As part of the evaluation scoping exercise for the preparation of the ToR
- During the fieldwork.

In some cases, the desk review constitutes the evaluation, as was the case of the review of the performance of the emergency response fund of the Consortium of British Humanitarian Agencies (Stoddard, 2011) or the desk review of unintended consequences of Norwegian aid (Wiig and Holm-Hansen, 2014).³

Desk reviews are also commonly used as the basis for developing studies of lessons learned, such as the lessons learned studies from ALNAP on responding to earthquakes (Cosgrave, 2008), urban disasters (Sanderson et al., 2012) and floods (Cosgrave, 2014). Such studies can be based on evaluations,

such as the World Bank review of earlier evaluations to inform the response to floods in West Africa (IEG, 2010) or on wider document sets such as Refugee Studies Centre review of lessons from a decade of discussions on protracted refugee situations (Milner and Loescher, 2011).

The desk review can be structured or unstructured. Structured desk reviews use a formal structure for document analysis, whereas unstructured reviews are basically background reading.



Tip

Written does not automatically mean accurate or reliable. Written data is sometimes regarded as more accurate or reliable than interview data. This is not necessarily the case, and all data from documents or interviews should be tested through triangulation.

A desk review entails:

- Identifying the documents. Generally the commissioning agency identifies an initial set of documents, but it is up to the evaluation team to identify further materials.
- Reading or otherwise analysing the documents.

The desk review can be carried out:

- By the evaluation team, whether as a minor initial task or as a major part of the evaluation. The literature review for the second phase of the Indian Ocean earthquake and tsunamis Linking Relief Rehabilitation and Development (LRRD) evaluation was an example of this type (Cosgrave et al., 2009).
- By consultants specifically contracted by the commissioning agency. This was done for evaluation of the LRRD theme of the evaluation of the 2004 Indian Ocean earthquake and tsunamis (Buchanan-Smith and Fabbri, 2005). In this example, the consultants both identified and analysed the document set.
- By the commissioning agency, although this is rare unless the review is part of an evaluation scoping exercise. Usually the involvement of the commissioning agency stops at identifying key documents for the evaluation team to consider.

10.2 Why do a desk review?

Desk reviews offer a cost-effective way for the evaluation to:

- Draw on the knowledge gained from previous evaluations and other research. For example, the Swiss Solidarity evaluation of its response to the 2004 Indian Ocean tsunamis based its initial draft survey questionnaire on an analysis of previous impact studies (Ferf and Fabbri, 2014: 17).
- Draw on the knowledge captured in project monitoring documents.
- Allow the evaluation team quickly to gain an understanding of the context.
- Identify potentially key issues for later fieldwork. The WFP thematic review of mother and child nutrition was informed by a 2002 desk review that identified key differences in WFP operations in different regions (Hoogendoorn et al., 2005: 10-11).
- Identify potential judgement criteria, sources, and methods for the evaluation matrix.

For these reasons, a desk review should form part of every evaluation. Even a low-budget evaluation will need the team at least to review the project documents and any relevant prior evaluations.

Desk reviews in EHA are particularly useful for:

- Establishing a chronology of what happened and when. They are essential if the evaluation is to consider questions of timeliness or effectiveness.
- Showing how intervention priorities have changed over time. Humanitarian crises are fluid and the context can change swiftly. The rapid turnover of key personnel in humanitarian crises means that the current management team may not know why particular approaches had been adopted. A desk review can help the evaluation team to understand the evolution of the response.
- Developing an understanding of different views. This is particularly the case in complex emergencies, as humanitarian workers may identify with the population they are assisting. A desk review can help the evaluation team to gain a broader understanding of the context.
- Identifying lessons learned from previous operations.

10.3 How long should a desk review take?

The time needed for the desk review depends on:

- The evaluation question(s) and the importance of documents are likely to be as a source of useful information. If documents are expected to be a key resource for answering the questions, then more time needs to be allocated to the review.
- The nature and scope of the evaluation. If the evaluation is examining a long period of intervention, documents are likely to be important sources of data and more time will be needed for the review.
- The richness of the available document set. New crises may have few documents beyond situation reports and needs assessments. Protracted crises may spawn thousands of documents. If the document set is extensive, then more time needs to be allocated to the review.
- The availability of summaries of any analytical literature. If these already exist, then less time may be needed for the desk review.

10.4 Conducting a desk review

The CDC's evaluation research team offers succinct advice on document reviews for evaluation (Evaluation Research Team, 2009). Hagen-Zanker and Mallett (2013) provide guidance on rigorous literature reviews in international development.

Step 1: Identify possible sources

The sources for documents depend on the evaluation questions and the time available for the review. The document set sources may include:

- Key documents referred to in the ToR. These may include project- or programme-specific documents as well as broader strategy documents.
- Agency-specific sources. These may include the key documents related to the evaluation (some commissioning agencies supply these as a CD-ROM or place them in an online folder). Agency websites or intranets can be a rich source of needs assessments, situation reports, project proposals, and monitoring reports.
- Relief Web, crisis-specific portals (such as the ALNAP Haiti Portal)⁴ and thematic portals (such as the Urban Humanitarian Response Portal).⁵

- The ALNAP Humanitarian Evaluation and Learning Portal (HELP).⁶
- Existing references and bibliographies.
- Web searches.
- Google scholar searches.
- Academic database searches.



Tip

Draw up a list of characteristics for the documents that you are looking for. You may decide initially to focus on evaluations and reviews, or material after a certain date – doing so can help to make the review faster.

Step 2: Categorise your documents

Not all documents should be given equal weight or attention. It can be useful to categorise the available documents into tiers:

Tier 1

Key documents that frame the subject of the evaluation are often listed in the ToR. They include the main programme planning and strategy documents and previous evaluations. Typically there may be between five and 20 key documents, although evaluations with a large scope that spans several countries, sectors, and long time periods may have far more.

Tier 2

Documents specifically on the subject of the evaluation, such as situation reports, progress reports, monitoring reports, and project proposals (where the evaluation is of a wider programme). The number of documents can range from 20 to several thousand, depending on the scope of the evaluation.

Tier 3

Background documents including media coverage and other agency reports. ReliefWeb postings are an example of this type of document, and may have over 700 postings a week on a new crisis, meaning that there may be anything from fewer than a hundred to tens of thousands of documents, depending on the scope of the evaluation.

Step 3: Decide if you need to take a structured approach

Many evaluations limit themselves to the first tier of documents, and an unstructured approach is therefore suitable. A desk review that includes documents from the second and third tiers requires a structured approach, such as:

- A structured observation tool to record comments (see below for an example)
- Using rubrics to rate aspects of the documents (see below for an example)
- Indexing and searching documents for content analysis

Step 4: Match the document study strategy to the tier

The evaluation team needs to read the first-tier documents at the earliest opportunity. Strategies for the second tier can include using a rubric and a tool for recording the reviewer's observations. Third-tier documents can be subject to content analysis.



Tip

Use bibliographic software, whether commercial products such as Endnote, or initially free products such as Mendeley or free products like Zotero, which is essential if there are many documents. They allow reference libraries to be shared within the team and ensure that documents are cited in a consistent way.



Tip

Use a cloud application to share documents such Dropbox, Box, One Drive or Google Drive to store a copy of the document set and make it available to all the team members.

10.5 Tools for desk reviews

Key tools for desk reviews include forms for recording review comments in a structured way. These can be combined with rubrics to assign scores to documents for particular aspects.

Rubrics

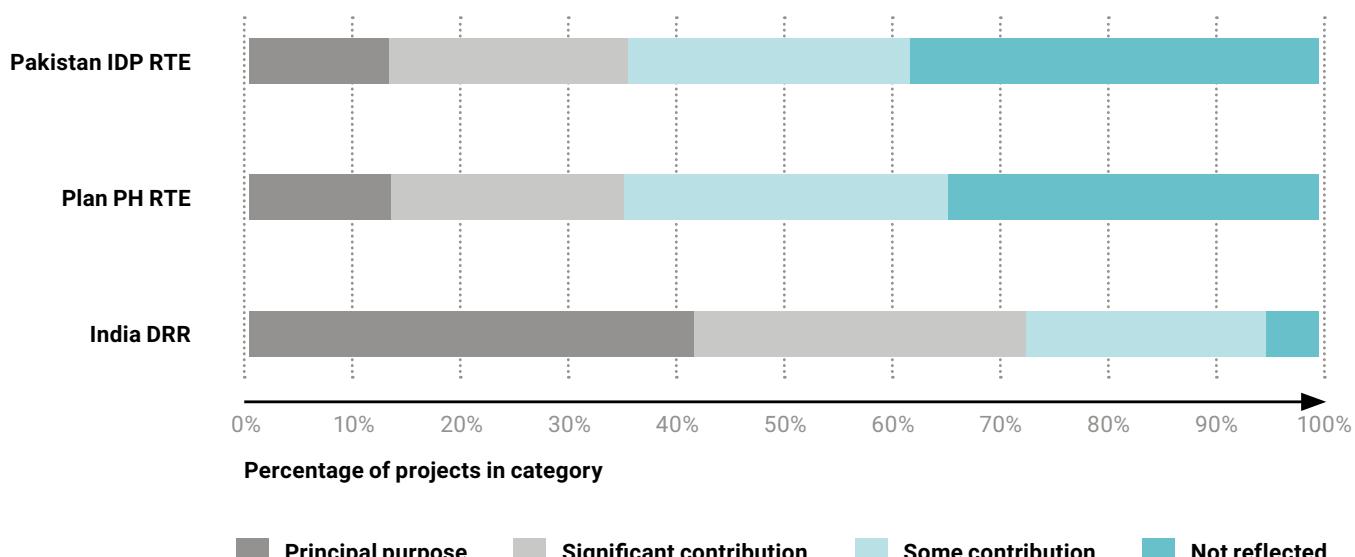


Definition: Rubric

A rubric is a scoring tool used to assess a document against a set of criteria in a consistent way.

The assigned scores can be used to categorise documents or to compare changes in emphasis over time. For example, the five-year evaluation of the Central Emergency Response Fund (CERF) used rubrics for gender, vulnerability, and other cross-cutting issues to find that while 95% of projects planned to make least some contribution to reducing vulnerability, less than two-thirds planned to make some contribution to promoting gender equality or other cross-cutting issues (Channel Research, 2011).

Figure 10.1: Gender, vulnerability and cross-cutting markers



*Gender, Vulnerability and Cross-cutting markers for a random sample of 258 CERF-Funded projects from 16 study countries. 41 projects excluded as they had insufficient information or were common services.

The Gender Marker (IASC, 2010) is an example of a rubric. It is most commonly used to assess the project design.

Level	Description
0	Gender is not addressed in any component of the project.
1	The project is designed to contribute in some limited way to gender equality. Gender dimensions are meaningfully included in only one or two of the three essential components: needs assessment, activities and outcomes.
2a	Gender mainstreaming: The project is designed to contribute significantly to gender equality. The different needs of women/girls and men/boys have been analysed and integrated well in all three essential components: needs assessment, activities and outcomes.
2b	Targeted actions: The principal purpose of the project is to advance gender equality. The entire project either (a) targets women or men, girls or boys who have special needs or suffer from discrimination or (b) focuses all activities on building gender specific services or more equal relations between women and men.

There is increasing interest in using rubrics in evaluation, not just for the desk review (Oakden, 2013a; 2013b; Rogers, 2013). See use of mini-rubrics by Davidson in workshops (2014).

Similar rubrics can be developed for any relevant aspect. A rubric may involve a more complex scale or simply use a four-point scale for the level of attention to any topic, such as disability, corruption mitigation, DRR or child rights. A simple four-point scale can divide documents into:

- Those that do not address the topic at all.
- Those that address the topic in a minor way. The rubric should provide guidance on what constitutes a minor way.
- Those that address the topic in some significant way. The rubric should provide guidance on what is considered to be significant.
- Those focused principally on the topic.

Rubrics are useful for ensuring:

- Consistency between different reviewers.
- Consistency over time for the same reviewer.

It is a good idea to have a second person to review a random selection of rated documents in order to check for bias or inconsistent ratings.

Structured review forms

These are simple forms where the reviewer can note key aspects of a document for later follow-up. For example, if the interest is in coordination within and between sectors, we might review documents with a quick initial scan and note items of interest, as shown below.

Another option for structured review is to use the evidence tool described in [Section 16: Analysis](#). Structured review forms can also be used to record rubric scoring for different aspects.

Doc	Pg	Sectoral coordination	Inter-sectoral coordination
1	6	'Regular attendance at cluster meetings.'	
	9	'We immediately implemented the WASH cluster decision to chlorinate all tankered water at sources and test chlorine levels at discharge.'	
	19	'87% of the tankers met the cluster standard. Those that persistently failed to do so were removed from contract and other cluster members were informed of the plate numbers.'	'The Water and Sanitation team met with the logistics cluster core team to discuss solid waste disposal.'

Content analysis

Content analysis is a useful way to manage a large number of documents. Most of the main packages for qualitative data analysis include some facilities for content analysis but are not always easy to apply. A cheaper and simpler alternative is to use an indexing tool such as dtSearch⁷ to search documents for specific keywords or perform other types of content analysis.



Definition: Content analysis

Content analysis is analysis of textual information in a standardised way that allows evaluators to make inferences about the information.

Content analysis typically takes the form of coding documents to reduce the text to a set of categories. One form of content analysis used in a number of EHAs, including the five-year CERF evaluation (Channel Research, 2011), is keyword analysis.



Definition: Keyword analysis

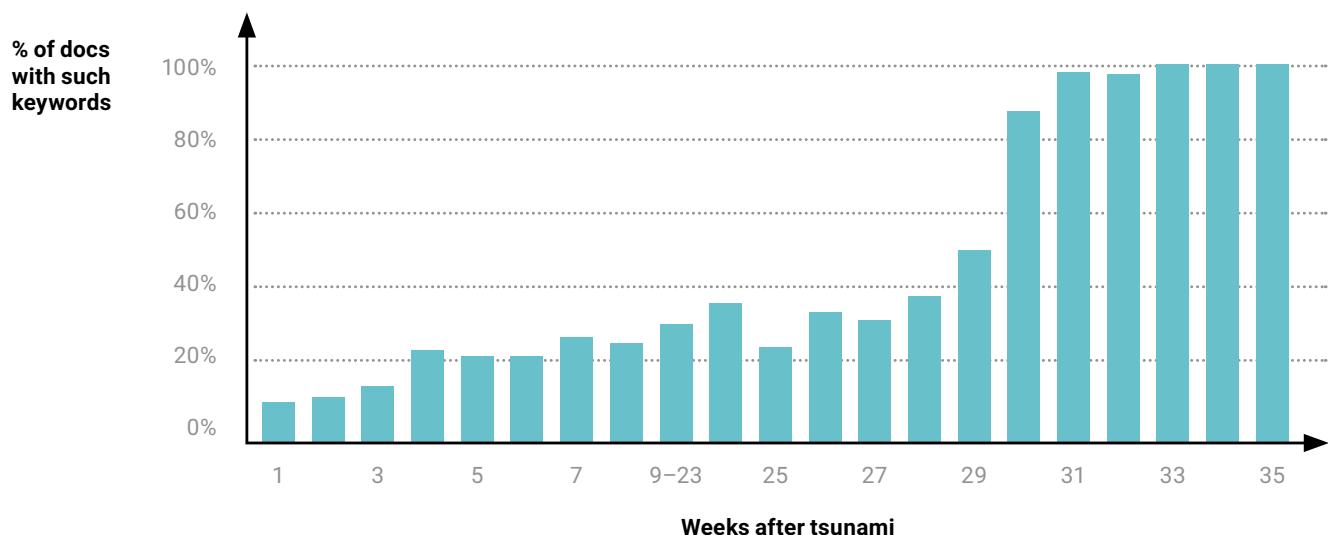
Keyword analysis is a form of content analysis that examines the frequency of occurrence of categories to highlight trends over time in a single document set or to compare two document sets.

A keyword list might have up to 500 terms. Some categories will be represented by several keywords. For example, keywords for the water, sanitation and hygiene (WASH) category include the following: borehole, chlorine, chlorination, defecation, faeces, hygiene, latrine, night soil, sanitation, soap, spring, toilet, wash, water, water supply, watsan, wells.

Other forms of content analysis are rarely used in EHA. One exception was the impact evaluation of assistance for returnee integration in Burundi where transcribed interviews were subject to a content analysis (Telyukov et al., 2009).

Three useful guides to textual analysis are Krippendorff (2004), a standard text; Benini (2009), a short guide for humanitarian and development workers in how to use three textual analysis tools; and the US General Accounting Office guide (GAO, 1996). Content analysis can be used to show the emergence and evolution of different concepts over time. The following example shows the prevalence of terms related to training in the ReliefWeb document set for the Tsunami Evaluation Coalition evaluation of the 2004 Indian Ocean earthquake and tsunamis. It can be seen that after 30 weeks nearly all documents referred to training in some way, but that initially there was little attention to this.

Figure 10.2: Prevalence of the keywords related to training in 14,560 tsunami documents



10.6 Summarising data through desk review

The findings from desk reviews undertaken by the evaluation team can be included in briefings, the inception report, and the main evaluation report. A particular challenge is to compress a large amount of data from a desk review into a readily accessible format. This can be achieved in the form of tables and graphs. Summarising data in this way can also help to inform the whole evaluation team.

Chronologies

Chronologies are very useful in EHA as they can help to address the sequencing of assistance.



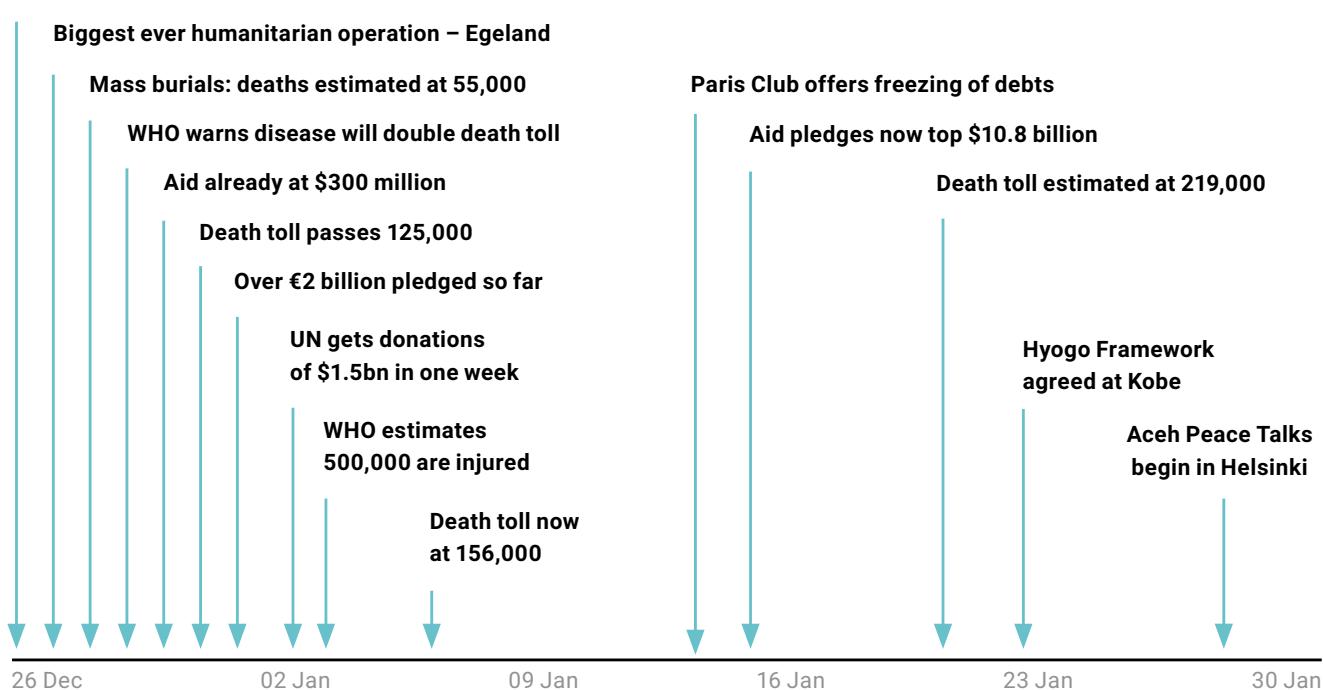
Tip

Start gathering chronology data as early as possible. Getting an early start on the chronology can help to put other reading into context as you begin the desk review.

Chronologies can be presented in tabular or graphic form. The following partial chronology illustrates the start of the response to the 2004 Indian Ocean tsunamis, and illustrates that funds were committed long before there were any detailed needs assessments.

Figure 10.3: Partial chronology for the Asian earthquake and tsunamis of 26 December 2004

Earthquake followed by a train of tsunamis



Chronologies can be developed from dated documents in the document set.



Tip

Data stamp documents by adding a prefix to every file name in the data set in the form YYYYmmDD. This allows the rapid sorting of documents and the checking of what documents were published on what date.

The indexing or content analysis software, such as the dtSearch software mentioned above, may be able to search for dates in documents and this can save time in building a chronology.



Tip

Involve the whole team in building the chronology because it helps everyone to gain a good understanding of how the response evolved.

Other time-ordered data

Chronologies are only one form of time-ordered data, which can help the team to understand how a crisis developed. It is possible, for instance, to contrast events with an epidemic curve, as was the case for CERF funding and the 2009 Dengue outbreak in Cape Verde.

Figure 10.4: Timeline for the Dengue outbreak in Cape Verde and the CERF application process

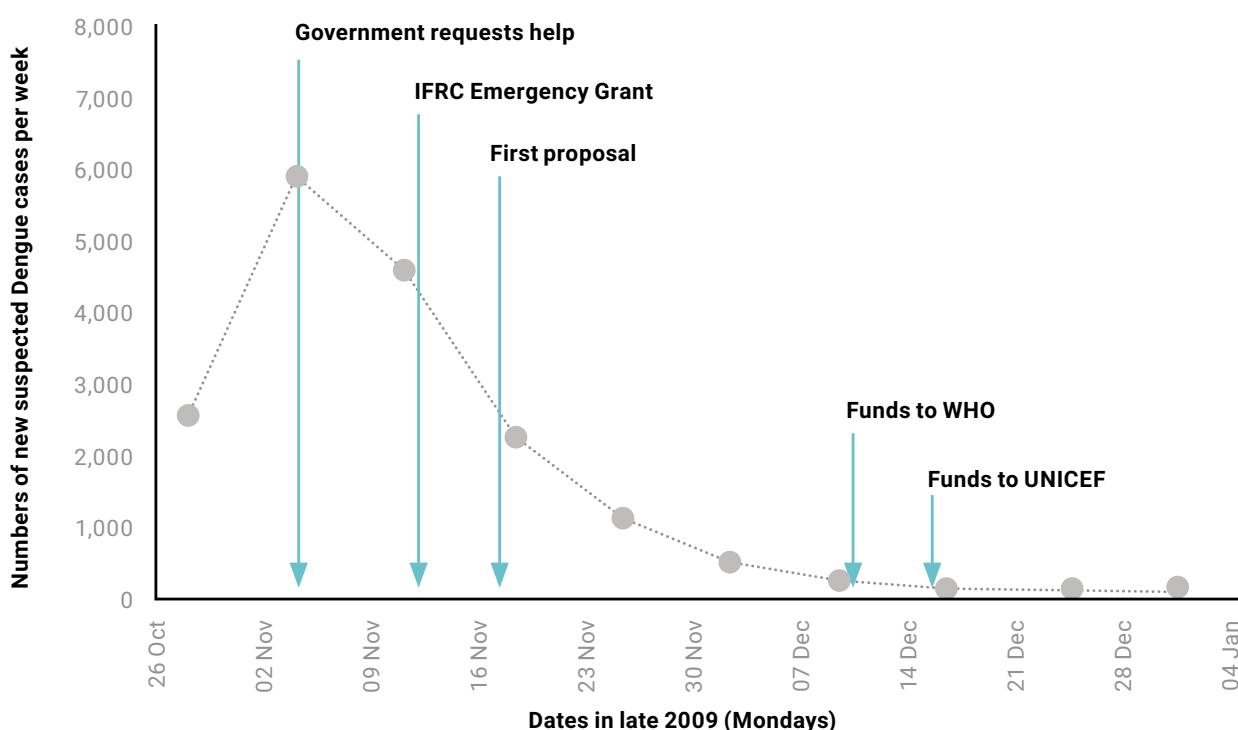
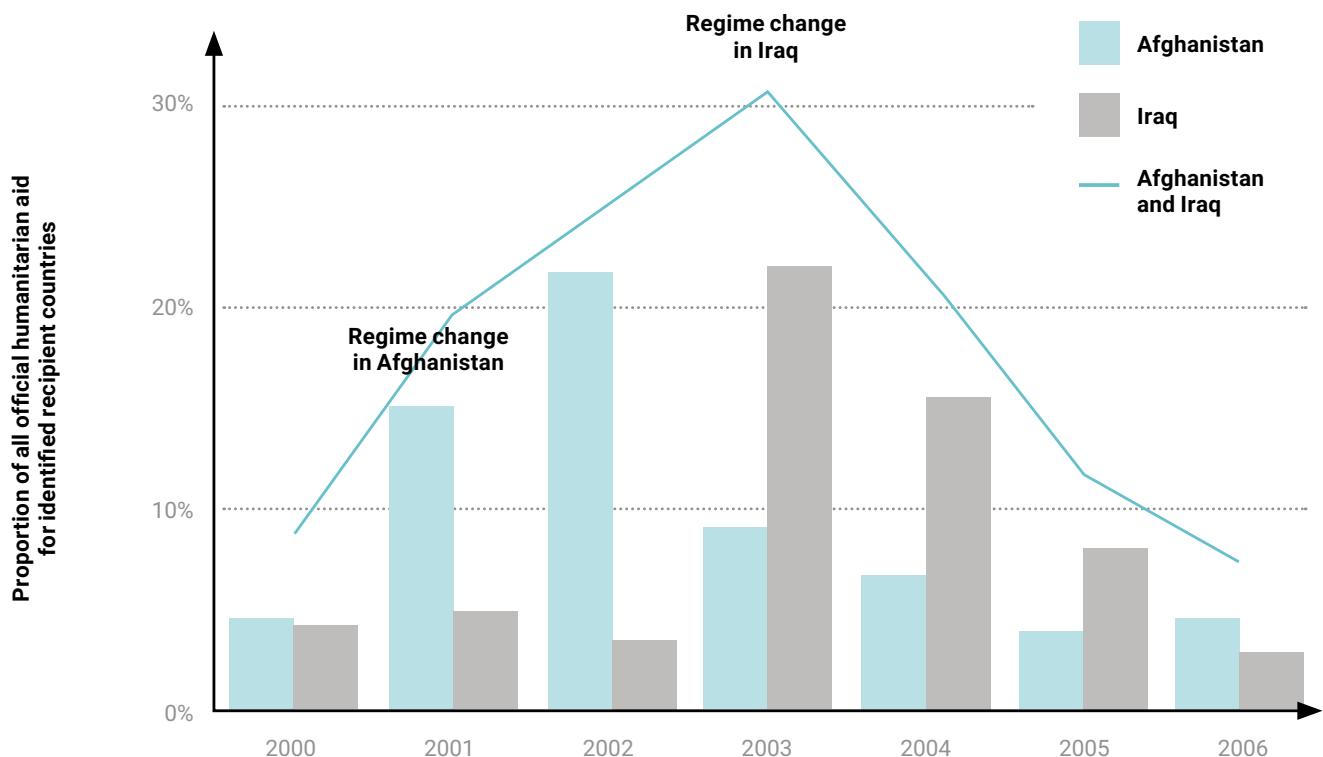


Figure 10.5: Humanitarian aid for Afghanistan and Iraq 2000-2006

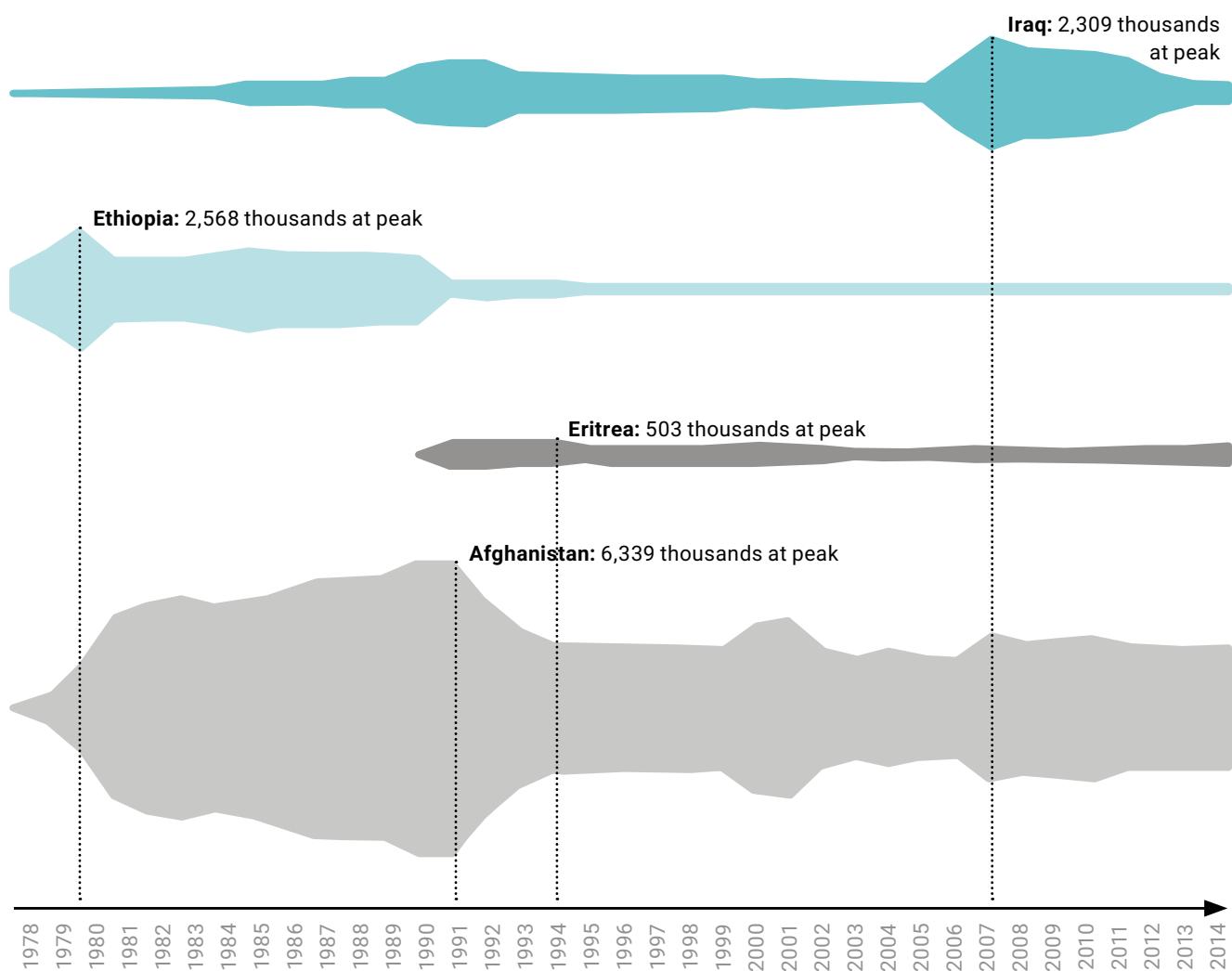


Source: OECD DAC Table 2a, 10 April 2008. Two thirds of humanitarian aid was for identified recipient countries during this period.

Time-ordered data are also useful for policy evaluations. [Figure 10.5](#) shows how regime change in Afghanistan and Iraq was associated with a large increase in official humanitarian assistance.

Refugee flows naturally lend themselves to being presented in graphic form. [Figure 10.6](#), for a study on protracted displacement (Crawford et al., 2015), shows the evolution of the refugee caseload from ten countries from 1978 to 2014. The width of each trace is proportionate to the number of refugees at the time. This graphic helps to illustrate the argument that there is no overriding pattern for protracted displacement, but that each crisis follows its own pattern.

Figure 10.6: A selection of refugee crises generating more than 400,000 refugees from 1978 to 2014



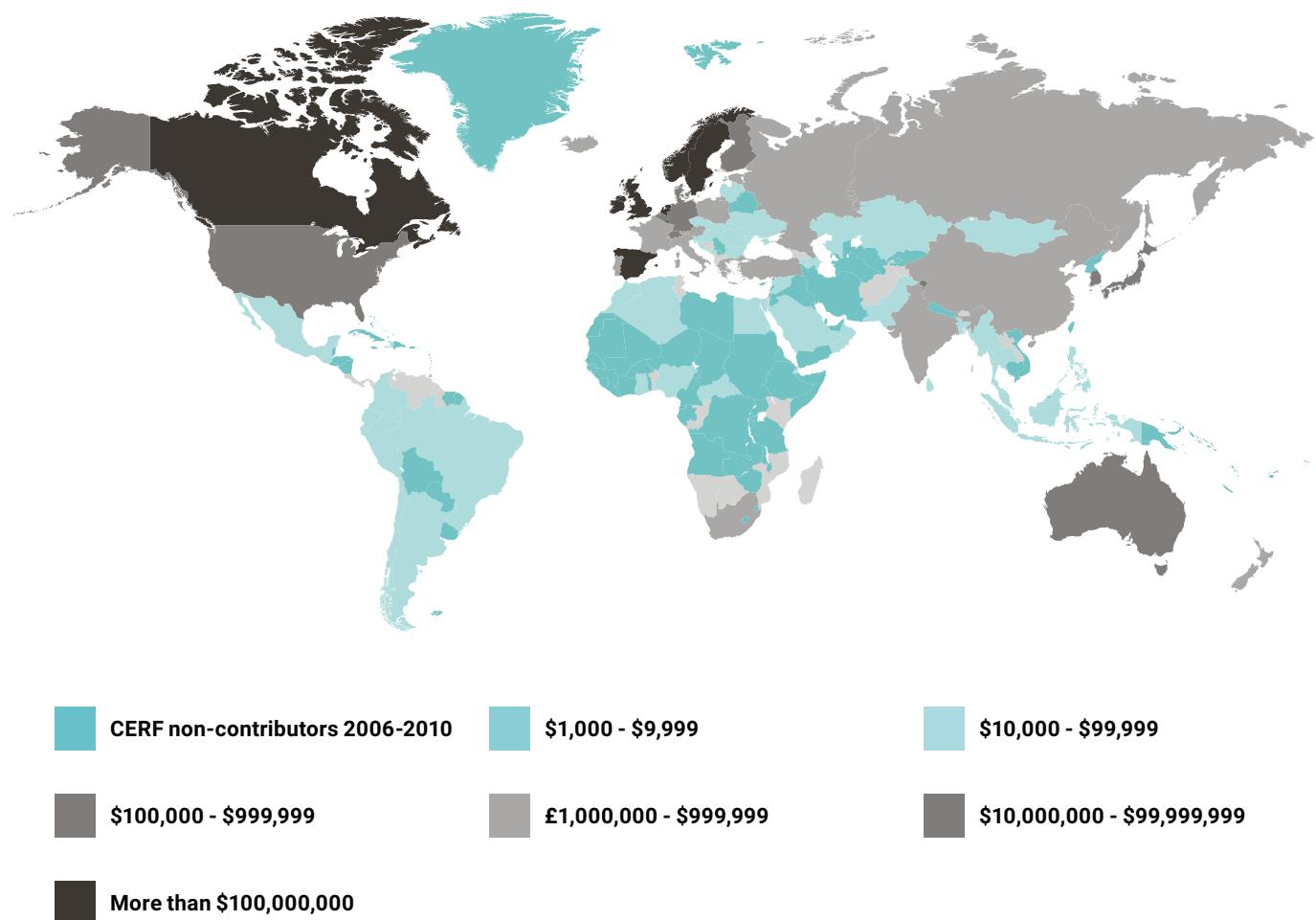
The width of each plot is proportional to the caseload from that country in that year. Based on UNHCR Data.

Geographical data

Geographical data can be summarised using Graphical Information Software (GIS) to highlight key issues. There are inexpensive and open-source GIS packages available but they are not always easy to master. The following map shows which countries contributed to the CERF in its first five years, with darker colours corresponding to higher contributions. This makes clear that the CERF enjoys wide support, including from developing countries.

Again the content analysis software can highlight the extent to which geographical names are used in a particular data set and this can be used to show their geographical focus.

Figure 10.7: Countries contributions to CERF between 2006-10

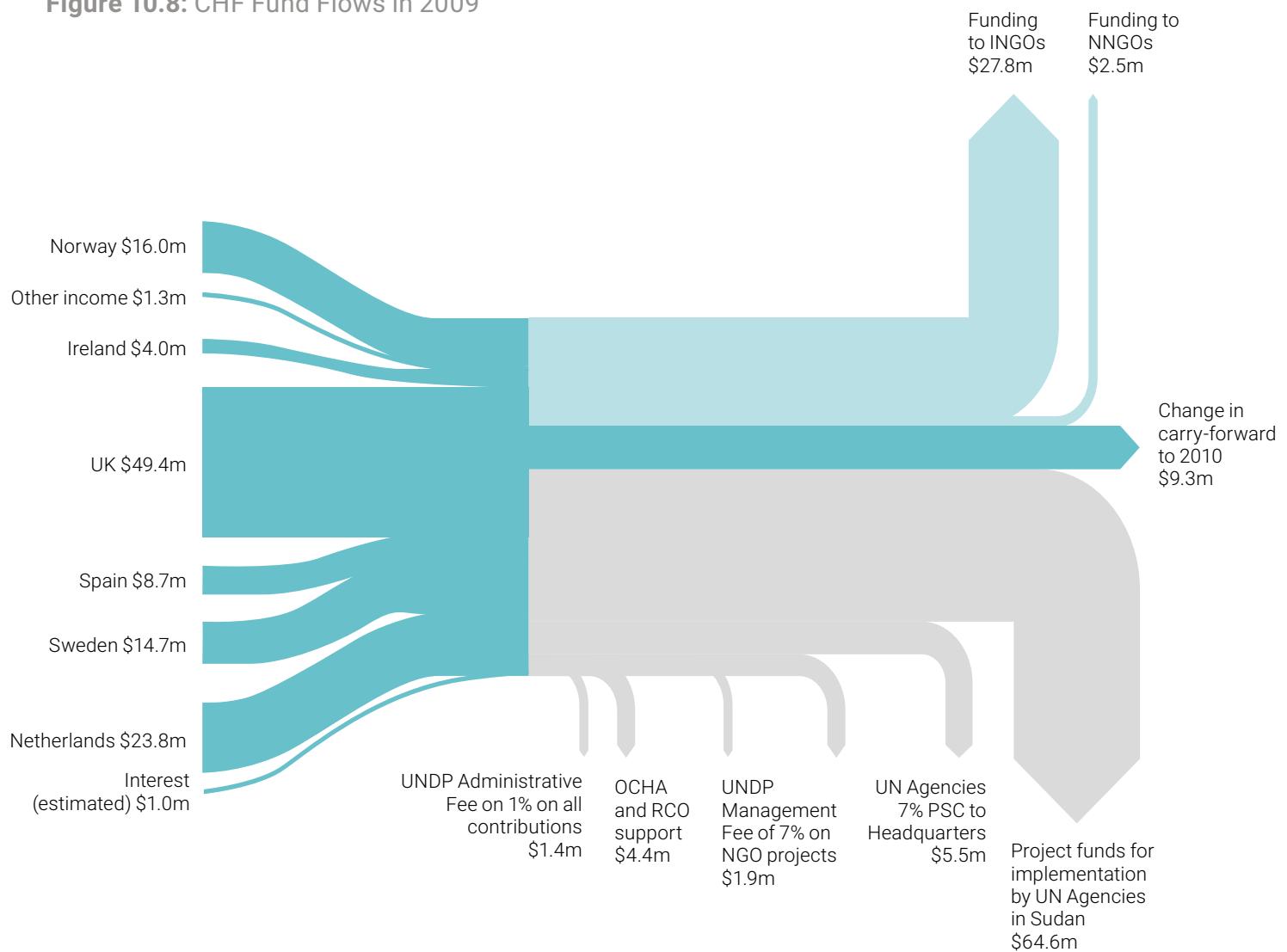


Flow analysis

Complex flows can be easier to understand if they are presented in a Sankey diagram. Figure 10.8, from the evaluation of the Common Humanitarian Fund in Sudan in 2009 showed that relatively little funding was being channelled to national NGOs, and that a considerable amount was dedicated to administrative overheads.

Section 16 will address numerical analysis.

Figure 10.8: CHF Fund Flows in 2009



Source: www.unsudanid.org, UNDP MDTF gateway and OCHA fund management section

Endnotes

9 / Planning and managing your evaluation

1. Jane Davidson (2010a) includes 'opt for the contractor with the lowest daily rate' in her tongue-in-cheek list of 9 golden rules for commissioning a waste-of-money evaluation.
2. For some UN organisations, for example, it is far faster to recruit evaluators directly than to work with a consulting company.

10 / Desk methods

3. This study found that all of the sampled Norwegian evaluations of humanitarian assistance found unintended consequences.
4. See www.alnap.org/current/haitilearningportal.
5. See www.urban-response.org.
6. See www.alnap.org/resources.
7. See <http://dtsearch.com>.

Notes

Notes

Carrying out the evaluation



11 / Evaluation designs for answering evaluation questions

This section deals with how to structure the evaluation in order to answer the evaluation question, which in turn can help evaluators and evaluation managers consider the possible evaluation designs.

The evaluation design is the overarching logic of how you organise your research to answer an evaluation question. Some evaluation designs are better at answering particular types of question. The objective of this section is help you consider which evaluation design or designs will allow the evaluation team to best answer the evaluation question(s) given the nature of the programme, the context in which it was implemented, and the constraints of the evaluation including access, budget, data and other factors.

Traditional research typically has a single hypothesis or question to consider, or a group of closely linked hypotheses or questions. Evaluations often include a range of disparate questions, however, and this may mean either using more than one type of evaluation design, or using designs that can answer all the questions but are not ideal for some questions.



Keep in mind

It is important not to confuse designs and methods. Design refers to the structuring of the data gathering and analysis, and method refers to how the data is gathered.

As noted in Section 6: Choosing evaluation questions the evaluation questions determine the evaluation design, the data-collection and analysis methods, and the sampling approaches.

This section also addresses the problem of bias, as this can vary depending on the evaluation design, methods, and sampling approaches.

11.1 Qualitative and quantitative

Some evaluation designs (such as case studies) are commonly associated with qualitative approaches and some (such as experiments) with quantitative approaches.



In depth: The problem with qual vs. quant

In principle, quantitative methods collect numerical data and qualitative methods collect non-numerical data. In practice, however, these labels are much more complex than this and Goetz and Mahoney (2012) describe them as representing two different cultures of research. Mahoney and Goetz (2006: 245) note that 'the labels quantitative and qualitative do a poor job capturing the real differences between the traditions. Quantitative analysis inherently involves the use of numbers, but all statistical analyses also rely heavily on words for interpretation. Qualitative studies quite frequently employ numerical data; many qualitative techniques in fact require quantitative information' (2006: 245).

The weaknesses of the labels have led to several others being proposed including, small-n, explanatory, and case-based for qualitative research and large-n, statistical, estimative, and population-based for quantitative research. Mahoney and Goetz (2006: 245) list 10 areas of difference between them. Patton (2014, exhibit 3.1) uses ten factors to contrast quantitative, qualitative, and mixed-method approaches.

This is a somewhat academic discussion but it matters to humanitarian evaluators because the two cultures have different theories of knowledge. In the case of quantitative research, knowledge springs from experimental methods, as in the physical sciences. This model of developing knowledge is called positivism. In the case of qualitative research there are several different theories of knowledge formation including ethnography and grounded theory. Patton (2014, ch.3) lists 16 different theoretical perspectives used in qualitative research.

The problem for evaluators is that there has been a recent trend towards treating only knowledge generated by experimental means as rigorous. Both the American Evaluation Association (2003) and the European Evaluation Society (2007) have criticised this approach. Knox-Clarke and Darcy (2014) state that using qualitative or quantitative methods does not automatically lead to stronger or weaker evidence.



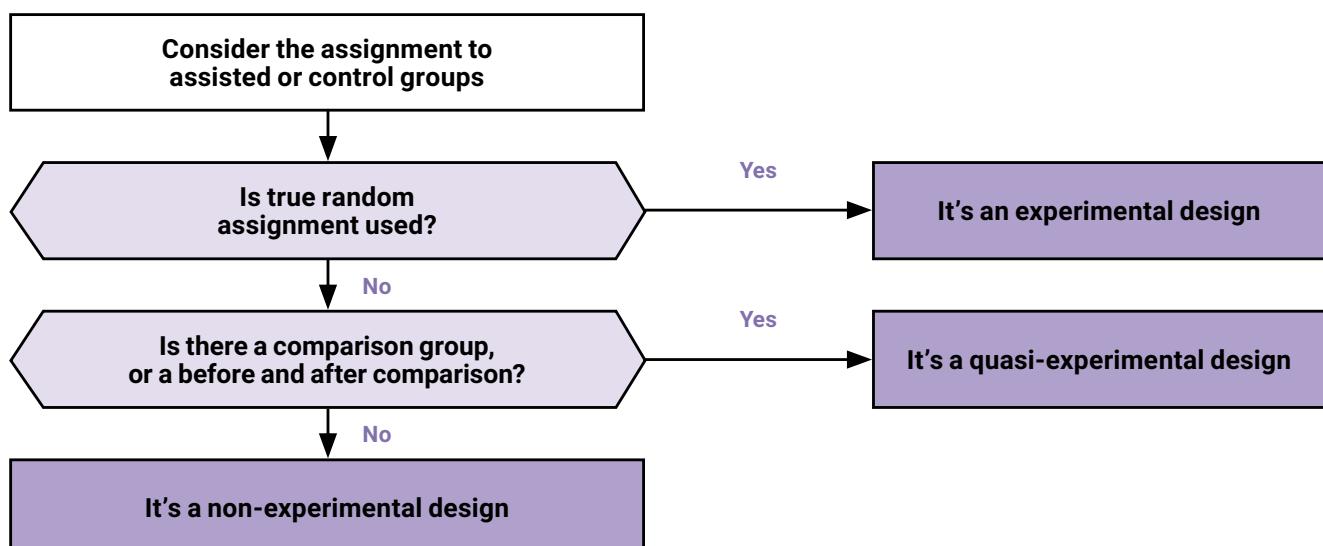
One further issue for evaluators is that commissioning agencies privilege experiments or apply measures that are appropriate to one type of research to another. Qualitative research cannot be assessed by uncritically applying the terminology and concepts used in evaluating quantitative research (Anastas, 2004: 57). The ACAPS paper distinguishes between the indicators of quantitative and qualitative research as: internal validity/accuracy, external validity/generalisability, reliability/consistency/precision, and objectivity for quantitative research against credibility, transferability, dependability, and confirmability, for qualitative research (ACAPS, 2013).

11.2 Families of design

There are three broad families of research design (examples are given below):

- Non-experimental designs with neither comparison nor control groups. Non-experimental design is the most common type of design in EHA.
- Experimental designs where assignment to the assisted or control groups is done before the start of the assistance.
- Quasi-experimental designs where comparisons are made either of the assisted group over time, or between the assisted and a comparison group selected after the start of the assistance.

The category an evaluation design fits into depends on whether a control or comparison group is used. Comparisons can be between groups or over time for the same group. The following flowchart summarises the distinctions between the families of EHA design.



If a research design uses neither control nor comparison groups, it is a non-experimental design, and is the most common type of evaluation design in EHA.

A single evaluation may require different designs in sequence in order to answer questions properly. This may be the case with mixed-method designs, which can consist of intervals of quantitative and qualitative methods that use different designs.

The ToR of Swiss Solidarity's major impact evaluation 10 years after the Indian Ocean earthquake and tsunamis originally asked that a quantitative survey be used prior to other qualitative methods. Ferf and Fabbri (2014: 16) advocated reversing this order to 'create an interview that was as open as possible, avoiding pre-conceived expectations on results, and affording opportunities to capture the unexpected as much as possible. Thereafter, the focus of the research narrowed to the issues that were identified as most relevant to the recovery of the beneficiaries.'

This can be still further improved with follow-up non-experimental examination of issues identified in the survey. This model, with a non-experimental scoping study defining the focus of a quasi-experimental survey, followed by a final non-experimental in-depth examination of emerging issues, is the cycle of research model offered by Remler and Ryzin (2015), and is similar to the sandwiched mixed-method examples given by Bamberger (2012).

Non-experimental designs



Definition: Non-experimental designs

Non-experimental designs are designs where there is no comparison, either between assisted and non-assisted populations, or for those who have received assistance over time.

Most EHAs use one-shot non-experimental designs because, among other reasons:

- They are the least demanding, in terms of the conditions that need to be met
- They are the most flexible
- They are relatively low cost
- They are able to answer all types of evaluation question
- They are a good fit with the skills of humanitarian evaluators, typically developed through needs assessment and programme planning

Non-experimental designs can include:

- Case studies
- Process reviews
- Outcome reviews
- Participatory designs.

Theory-based evaluation is sometimes presented as a type of non-experimental design, although it is more complex than that. A theory-based evaluation can use a range of designs to test the underlying theory of change (ToC), including experimental, quasi-experimental and non-experimental designs.

Case studies, where the evaluation examines a series of different units of analysis to draw general conclusions about the intervention, are probably the most common EHA design. True participatory designs (where the stakeholders define the evaluation) are not used in EHA, as further explained in [Section 14: Engaging with the affected population in your evaluation](#).

Humanitarian evaluators need to move beyond using only non-experimental designs because:

- Other designs are better able to answer some types of evaluation question
- Some donors are exerting pressure to move towards what are sometimes seen as more rigorous designs
- Other designs may be able to answer some evaluation questions at lower cost than non-experimental designs (this can be especially true of quasi-experimental designs using secondary data).

The [sub-section on selecting your design on pg 200](#) suggests when particular designs are most useful.

Quasi-experimental designs



Definition: Quasi-experimental designs

Quasi-experimental designs are designs using a comparison where the comparison group is not randomly selected.

Quasi-experiments cover a wide range of designs including:

- Designs where the assisted group is compared with a comparison group (this can be at one point, or over a period of time)
- Designs where a single group is compared over time.

A range of quasi-experimental designs is presented in the sub-section 'selecting a design'. The main strength of such designs is that they can provide rigorous evidence for whether an intervention has been effective, while avoiding the ethical problems of experimental designs.

The use of comparison groups

The use of comparison groups helps to reduce the risk that any changes seen in the assisted group are due to broader background changes, rather than the assistance given. Designs with comparison groups are more robust than designs without.

However, comparison groups in humanitarian contexts are very susceptible to experimental contamination due to the large number of actors and other support networks (including families).

Experimental contamination occurs when we cannot control all aspects of the assistance to the assisted and control or comparison groups.

Contamination in experiments and quasi-experiments in humanitarian settings can occur in many ways including when:

- Other agencies or family members provide assistance to members of the assisted control group
- Members of the control group learn from the assisted group (about the benefits of washing hands before touching food, for example) and adopt the improved practice.

Problems with comparison groups, some examples

An evaluation of cash grants for returning refugees in Burundi had planned to compare recipients and non-recipients, but soon noted that the two groups were not comparable since they had returned at different times (Haver and Tennant, 2009).

The Mozambique Child Soldier Study had planned to use a comparison group but found that many of those they selected for interview would not fully answer their questions (Boothby et al., 2006: 106). The Inter-Agency Guide to the Evaluation of Psychosocial Programming in Humanitarian Crises notes that this led the study to used 'local norms' as a comparison rather than other child soldiers (Ager et al., 135).

An impact evaluation of food assistance for refugees in Bangladesh noted that the difference between the assisted and comparison groups jeopardised the internal validity of the study (Nielsen et al., 2012).

The Milk Matters study in Ethiopia used a comparison group that was already significantly different from the assisted group even before the treatment (Sadler et al., 2012). When there is a big pre-intervention difference between the two groups it is impossible to attribute any subsequent difference to the intervention.

Comparison groups can be established through a number of methods. The most rigorous is probably propensity matching through statistical methods.



Definition: Propensity score matching

Propensity score matching is a statistical matching technique that attempts to match the comparison group to the control group through selecting one with the same probability of being assisted based on the group's characteristics.

Caliendo and Kopeinig (2005) provide guidance on the use of propensity score matching in evaluation. Some of the difficulties in applying it emerged in the review of mass supplementary feeding in Niger (Grellety et al., 2012).

Another way to select a comparison group is to draw on expert opinion.

Experimental designs

Experimental designs are the most demanding and many conditions need to be met in order to apply them successfully. Some people regard them as the most rigorous design.



Definition: Experimental designs

Experimental designs are where units of analysis are randomly assigned to the assisted or the control group. Each element (e.g. a person, family, or community) has an equal chance of being assigned to either the assisted or the control group.

Random assignment of assistance poses ethical problems in EHA since assistance should be given on the basis of need rather than randomly.

The randomised control trial (RCT) is the most common experimental design. Where RCTs have been conducted in EHA settings, they have been sometimes been conducted between different forms of assistance (cash or vouchers or food, such as Sandström and Tchatchua, 2010) rather than between assistance and no assistance, or in the recovery phase such as the evaluation of community-driven reconstruction in Lofa Country (Fearon et al., 2008). Some RCT studies have been conducted on health in humanitarian settings, such as the effect of supplementing food with probiotics and prebiotics in Malawi (Kerac et al., 2009) or insecticide-treated plastic sheeting in Sierra Leone (Burns et al., 2012).

11.3 Selecting a design

Ideally, the design is determined solely by the evaluation questions, and no evaluation design is perfect. The constraints imposed by timing, budget, data availability, and so on limit the options. The options chosen, and the reasons for doing so should be noted in both the inception and final reports. This is because it may be necessary to depart from the planned design in the field. This section describes the key points of a range of designs. The parentheses indicate the family of design.

Case study (non-experimental)



Definition

Case studies are an intensive description and analysis of one or more cases (which can range from individuals to states) to draw general conclusions about the intervention.

Use in EHA

Case studies are commonly used, usually relating to families and communities. Country case studies may be used in large-scale evaluations.

Guidance

Yin (2003a; 2003b) provides guidance on the use of case studies

Strong points	Weak points
<ul style="list-style-type: none">• Good for answering questions about why people have done things• Good for examining rare conditions or complex interventions• Good for testing theory• Provide rich data about the cases• Good for theory building• Good fit with EHA	<ul style="list-style-type: none">• May be difficult to generalise• Cases are usually purposively selected, making it harder to generalise from the findings• Less useful for attribution studies

Examples

The 2015 evaluation of Danish humanitarian assistance used two in-depth field-level case studies (South Sudan and Syria crisis) and one more limited desk study (Afghanistan) (Mowjee et al., 2015).

Process review (non-experimental)



Definition

A process review compares how processes function with how they were planned to function.

Use in EHA

Process reviews are used for checking whether processes conform to broader standards such as Sphere or to an agency's own standards.

Guidance

Most of the guidance for process reviews has concentrated on business process reviews. No humanitarian-specific guidance was found.

Strong points	Weak points
<ul style="list-style-type: none">• Good for answering normative questions• Good for answering questions about planning assumptions• Good for answering questions about process efficiency• Can be done at any point in an intervention	<ul style="list-style-type: none">• Generally pays little attention to impact• Good only for some questions• Inherent assumption that a correct process will ensure results

Examples

The 2013 UNHCR review of participatory assessments is an example of a process review. The Process review of the Common Humanitarian Fund in Somalia is another (Willets-King et al., 2012).

Outcome review (non-experimental)



Definition

An outcome review compares outcomes with planned outcomes.

Use in EHA

Limited formal use, but many EHA reports focus on outcomes rather than impact.

Strong points	Weak points
<ul style="list-style-type: none">• Avoids the often difficult task of assessing contribution or attribution for impacts• A good match with programme planning documents	<ul style="list-style-type: none">• Does not address impact. While outcome may be achieved, it does not necessarily lead to positive impacts• Does not indicate why things happen

Examples

No formal examples in EHA, but the Oxfam GB evaluation of the Kenya Drought Response (Turnbull, 2012) refers to an earlier outcome review of cash transfer programmes.

Participatory design (non-experimental)



Definition

Participatory design involves all stakeholders throughout all phases of the evaluation, from the initial planning to the implementation of the recommendations. (see also participatory evaluation in [Section 4](#)).

Use in EHA

No true EHA examples were found. While many evaluations strive to have participation, none were found where the whole process was controlled by the stakeholders. This approach is used mainly in the US charitable grant sector. Participatory methods are often used in evaluation (Catley, 2009; 2014), but an evaluation is classified as participatory only when the stakeholders define the evaluation questions.

Guidance

There is a brief USAID guidance note on participatory evaluation (USAID, 1996) and a more detailed CRS manual (Aubel, 1999). A good example of a participatory evaluation is the evaluation of UNHCR's Age, Gender and Diversity Mainstreaming in Colombia (Mendoza and Thomas, 2009).

Strong points	Weak points
<ul style="list-style-type: none">• Empowers participants• Supports utilisation• Promotes partnership	<ul style="list-style-type: none">• Difficult to predict the results• Subject to stakeholders' existing biases (for continued funding, for example).

Interrupted time series (quasi-experimental)



Definition

The interrupted time-series design provides an estimate of the impact of an intervention by examining a time series of data before and after an intervention.

Use in EHA

Little current use.

Guidance

Cochrane provide a short worked example of an interrupted time-series design (EPOC, 2013). Glass (1997) provides many more examples of its use. Gilmour et al. (2006) give an example where the exact intervention date is unclear. Wagner and Ross-Degnan (2002) offer guidance on the use of segmented regression analysis or interrupted time series.

Strong points	Weak points
<ul style="list-style-type: none">Inexpensive, since it tends to use existing secondary data	<ul style="list-style-type: none">Subject to contamination, especially from time-based effects such as maturation (behavioural changes over time) and so onNeeds careful examination to eliminate alternative causesNeeds strong statistical skills

Examples

There do not seem to be any evaluations of humanitarian action using interrupted time series, but it has been used for non-evaluation studies in humanitarian settings, such as suicide after the 1999 Taiwan earthquake (Xang et al., 2005) or pneumonia after the Fukushima disaster (Daito et al., 2013).

Difference in differences (quasi-experimental)



Definition

This design estimates the effect of assistance by comparing the average change over time in the outcome of interest between the assisted group and a comparison group.

Use in EHA

No examples of use in EHA. Some evaluations refer to the method but do not apply it rigorously.

Strong points	Weak points
<ul style="list-style-type: none">• Can be inexpensive if secondary data are used• Where large changes are seen, provides compelling visuals of the difference• Is much more intuitive than many other statistical methods	<ul style="list-style-type: none">• Needs strong statistical skills• Comparison group should be sufficiently similar

Before and after comparison (quasi-experimental or non-experimental)



Definition

This design compares the situation of a group before and after the intervention.

Use in EHA

Little current use.

Strong points	Weak points
<ul style="list-style-type: none">• Shows that some change has unquestionably taken place• Shows how large the change is (the effect level)• Can be inexpensive if there are secondary data available	<ul style="list-style-type: none">• No comparison group, so any changes seen may be due to other factors, such as the external economy etc

Examples

A good use of baseline data for a before and after study was the study of flood survivors in Bangladesh by Durkin et al. (1993). Using data from an unrelated study undertaken six months before, they compared the behaviour of the children measured then with their behaviour after the floods, suggesting that the children were showing some signs of post-traumatic stress.

Comparison group (quasi-experimental or non-experimental)



Definition

Comparison group designs compare the assisted group with a selected comparison group

Use in EHA

Weak comparison groups are sometimes found in EHA. Stronger comparison groups can be established.

Strong points	Weak points
<ul style="list-style-type: none">• Comparison groups reduce the risk of mistaking background changes for the impact of the intervention• Comparison groups can make the findings more compelling	<ul style="list-style-type: none">• Very difficult to avoid contamination in humanitarian settings• Strong statistical skills and good data on the assisted group needed for methods such as propensity matching

Regression discontinuity design (quasi-experimental)



Definition

A regression discontinuity design compares the regression lines for the variable of interest against the score on which the intervention was based.

Use in EHA

No true examples of regression discontinuity designs were found, although some treatment discontinuity studies described themselves as regression discontinuity designs.

Guidance

Jacob et al. (2012) provide guidance on the design as do Better Evaluation and Khander et al. (2010: ch.7).

Strong points	Weak points
<ul style="list-style-type: none">• A good match for humanitarian action as the assisted and comparison group are selected on the basis of need• Can produce convincing visuals• More robust than simple comparison of above and below cut-off	<ul style="list-style-type: none">• Needs strong statistical skills• Risk of contamination if the cut-off criteria are not strictly applied• Results are generalisable only around the cut-off

Treatment discontinuity comparison (quasi-experimental)



Definition

A treatment discontinuity design compares the group just below the cut-off point for assistance with the group just below.

Use in EHA

Some use, although some studies are falsely labelled as regression discontinuity designs.

Strong points	Weak points
<ul style="list-style-type: none">• No ethical concerns• A good match for humanitarian action as the assisted and comparison group are selected on the basis of need• Does not necessarily need strong statistical skills	<ul style="list-style-type: none">• Risk of contamination if the cut-off criteria are not strictly applied• Results are generalisable only around the cut-off

Examples

The review by Lehmann and Masterson (2014) of cash grants in Lebanon uses this design.

Longitudinal design (quasi-experimental or non-experimental)



Definition

Longitudinal studies make repeated measurements of the same population over years.

Use in EHA

Little used except for analysis of secondary data from such studies. Most common use is in cohort studies for health in developed countries.

Strong points	Weak points
<ul style="list-style-type: none">• Robust indication of gradual trends over time	<ul style="list-style-type: none">• Such studies are expensive to undertake• Identify changes and trends but not the underlying reasons

Examples

Khoo (2007) used the Longitudinal Surveys of Immigrants to Australia to investigate the health and economic participation of humanitarian immigrants using logistic regression and two comparable cohorts.

Randomised control trial (RCT) (Experimental design)



Definition

A randomised control trial (RCT) compares two randomly selected groups, one of which is given assistance while the other (the control group) receives none.

Use in EHA

Very few examples because of ethical issues, cost and complexity. Sometimes RCTs have been used in the recovery phase, or have compared different assistance packages rather than using a control group with no assistance.

Strong points	Weak points
<ul style="list-style-type: none">• Some believe it is the most rigorous design• Good for addressing causal questions about attribution• High degree of confidence that the answer is an accurate reflection of the truth in the context• Random assignment largely resolves selection bias issues	<ul style="list-style-type: none">• Many ethical concerns (including random assignment of assistance rather than on basis of need)• Most expensive design• Least flexible design• Demanding preconditions• Appropriate only for specific question types• It is rarely possible to keep the groups ignorant of whether they are the assisted or the control group – leading to contamination• Difficult to control contamination (e.g. the control group may get similar assistance from another source)• A sampling frame (a list that more or less included the whole population) can be hard to obtain in humanitarian context• Need to randomly select assisted and non-assisted groups prior to intervention• Limited generalisability as results apply only to the single experiment

Examples

Fearon et al.'s 2008 impact assessment of community-driven reconstruction in Lofa County used an RCT.

Natural experiment (quasi-experimental)



Definition

Natural experimental designs make a comparison between an assisted group and a similar group that, by chance, has not been assisted.

Use in EHA

Occasionally used when the researchers notice that conditions have provided a natural experiment.

Guidance

Natural experiments are described in chapter 15 of Remler and Ryzin (2014).¹

Strong points	Weak points
<ul style="list-style-type: none">• Good for addressing causal questions about attribution• High degree of confidence that the answer is an accurate reflection of the truth in the context• Free of the ethical concerns regarding deliberate experiments• Inexpensive	<ul style="list-style-type: none">• Concerns that the assisted and the comparison group may be subtly different, thus biasing the result• Researcher cannot set them up, only discover them if they occur

Examples

One example of a natural experiment is the study of how children in Nicaragua were affected by Hurricane Mitch in 1998 (Baez and Santos, 2007). Other examples are on the long-term impact of famine on survivors (Meng and Qian, 2009), of the impact of destocking on child nutrition (Boonstra et al., 2001), on the impact of refugees on the local economy (Maystadt, 2011), on risk-taking after disasters (Cameron and Shah, 2012; Page et al., 2012) and on security restrictions on support for violence (Longo et al., 2014).

A simpler example of a natural experiment occurred in the aftermath of the 2006 Yogyakarta earthquake. Some local government jurisdictions were already on alert for an eruption of Mount Merapi and had a disaster-preparedness and management infrastructure. The Fritz Institute found high levels of satisfaction with the emergency response in those areas (Bliss and Campbell, 2007).

11.4 Bias

The choice of design, methods and sampling approaches influences the potential types of bias. Bias is a threat to the accuracy of the evaluation findings (Knox Clarke and Darcy, 2014: 15).

There is a difference between bias in qualitative and quantitative methods. Bias is deliberately embraced in quantitative methods – samples are purposively chosen, and the researchers are required to reflect on the impact of their own biases and those generated by the research process. Even so, bias poses a significant threat to the internal validity of quantitative methods.

Internal validity is the extent to which conclusions about cause and effect are justified. Bias can arise from the evaluator's prejudices and/or from the evaluation design. Nobody is immune from bias, but a bias arising from the evaluation design is insidious as it may not be evident.

Design bias

Design bias arises through systematic errors in the evaluation design. Sackett (1979) identified 35 different types of bias that could affect analytical research, including various forms of selection and measurement bias.²

Selection bias occurs when the sample elements chosen are biased in some way. One example would be if beneficiary distribution lists were used as the sampling frame from which to select a sample. This would bias the sample if a significant number of people did not receive assistance. Using any method other than true random selection of the sample from the sample frame also introduces the risk of bias.

Measurement bias occurs when the measurement is biased in some way. Measurement bias can occur for a variety of reasons. For example, if a beam-balance weighing scale is not properly levelled, all the measurements will be inaccurate. Another example is if survey interviewees overstate the number of persons in their household in the hope of obtaining more assistance.

For experimental designs, random selection helps to prevent selection bias and concealing whether someone is in the assisted and control group from both the group members and the staff working with them (known as double blinding). This helps to eliminate measurement bias.² Quasi-experimental studies are subject to both potential selection bias and measurement bias, but these may be controlled to some extent by statistical measures.

Purposive sampling is deliberately biased in order to select the most appropriate cases for the evaluation questions, but non-experimental designs risk unintended selection bias through the exclusion of particular groups (youth, or those furthest from the road, for example).

For both quasi-experimental and non-experimental design, a transparent and rigorous approach to sampling, based on explicit criteria established in advance, can help to control design bias. It can be useful to include a sampling plan in the inception report.

Evaluand, source, and publication bias

Organisations are more likely to evaluate what are perceived to be successful programmes than failed ones. The 2009 American Evaluation Association session on Alternative to the Conventional Counterfactual noted that 'many evaluations have a systematic positive bias as most of the information is collected from people who have benefited from the project' (Bamberger et al., 2009: 2). The UN Evaluation Group guidance notes that since implementing agencies benefit from the projects this contributes to positive bias (UNEG, 2011: 22).

The 2015 State of the Humanitarian System report (2015: 29) states that: Over 50% of evaluations rate the performance of the evaluation subject as good. Most evaluations are commissioned by agencies or donors with a vested interest in the results, so it is very possible that the incentive structure tends toward more positive findings, even when external consultants are hired to carry out the evaluation. A summary reading of results led us to conclude that 'good' was also often used to indicate 'adequate' performance.

Of course the '50% of evaluations' should probably be qualified as '50% of published evaluations' since evaluations are also subject to a publication bias, which reduces the likelihood that critical evaluations will be published.

Table 11.1: Example of focus group discussions and reasoning

Group	Assisted group	Comparison group	Reasoning
Adult male	3 focus groups	3 focus groups	Previous research has shown that men and women value various elements of the assistance in different ways.
Adult female	3 focus groups	3 focus groups	
Youth	2 focus groups	2 focus groups	One group each for males and females. Some reports express concern about impact of assistance on youths.
Disabled	1 focus group	1 focus group	One group due to small number.
Elderly	1 focus group	1 focus group	One group due to small number.

Biases in data collection

The evaluation of the response to tropical cyclones in the Caribbean in 2004 (Gamarra et al., 2005: 3) included a short section on how the team addressed bias, including agency bias, location bias and memory bias.

Measures to control bias include triangulation and other approaches including:

- Interviewing people who are away from the road or who have been interviewed often. Including as many non-agency interviewees as possible, and controlling for memory bias by checking against situation reports (Gubbls and Bouquest, 2013: 18).
- Interviewing the affected population without the presence of agency staff. Avoiding asking direct questions about an agency's performance, using indirect evidence on the relationship, the understanding of the project and perceptions of ownership (Cosgrave et al., 2010: 14).
- Ensuring a gender balance among interviewees in order to limit gender bias (Barakat et al., 2006: 139).
- Routinely cross-checking information with as wide a range of people as possible (Humanitarian Initiatives et al., 2001: Vol. 3, sect. 6.1).

- Including an external consultant as the team leader with a team drawn from the evaluated agencies (Reed et al., 2005: 1-2). Avoiding anyone who has previous experience of a particular agency taking the lead in the interviews with that agency (Cosgrave et al., 2008: 21).
- Using collaborative tools to share emerging findings and reviewing them within the team. Openly identifying any potential biases in the evaluation team (Ternström et al., 2007: 90-91).
- Analysing results on a country-by-country basis to avoid any bias introduced by the over-representation of some countries in an online survey (Steets et al., 2013: 15).
- Not informing enumerators of the purpose of the treatment status of a community, and not informing the affected population of the purpose of the survey (Beath et al., 2013: 46).

Gender and age bias

Gender and age bias can arise in data collection and give a misleading picture of what is happening.

UNEG has guidance on incorporating human rights and gender equality in evaluations (2014).

Preventing gender and age bias in field data collection

The following steps will help ensure that an evaluation is gender- and age-sensitive and yields data that can be disaggregated by sex and age:

- Ensure that the fieldwork team comprises an equal number of women and men. Female interviewers and translators usually have the best access to women and girls in the affected population.
- Ensure that key informants include women who are knowledgeable about the community and about the particular needs of women and whether and how they have been met. These could include female teachers and nurses and leading market women. Ensure that some key informants are also knowledgeable about the needs of children, youths and older people.
- Ensure that women comprise at least half of the participants in focus groups and group interviews. Ideally, women should be interviewed separately from men, who tend to dominate mixed-sex discussions.





- Hold age-set focus groups and group interviews, for example with children, youths and older people. Adults tend to dominate mixed-age discussions.
- Keep a record of the gender of key informants and other interviewees in order to check for any potential gender bias in the information gathered from interviews.

Source: Based in part on Mazurana et al. (2011)

Evaluator bias

Evaluator biases may include a dislike of a particular agency or programme approach, or the temptation to repress highly critical findings for fear of losing future contracts with the client. Such biases may apply to all forms of research design, but non-experimental methods present the greatest risk. Such biases may be controlled to a limited extent by using researcher triangulation, comparing the findings from different researchers.

Approaches that can help to control for evaluator bias include:

- Recruiting evaluators who are aware of their own biases and try to allow for them
- Transparent decision-making about sampling
- Transparent data-collection and analysis strategies
- Clear links between evidence, findings, and conclusions
- External quality assurance
- Triangulation of findings between evaluators
- Having more than one evaluator code the same data and comparing the results for systematic bias.



Tip

Be transparent about bias. Include a statement in the evaluation report stating the potential biases in the evaluation team and identify what steps you have taken to control this and other forms of bias.

12 / Sampling

This section covers sampling. Sometimes it is possible to investigate every single case, but often there are too many cases or too many potential informants to consult all of them. Instead we take a sample.



Definition: Sampling

The selection of a subset of a population for inclusion in a study instead of the entire population.

Almost every evaluation uses sampling. This section sets out how even the smallest evaluation can make good sampling choices to improve the quality of the evaluation.

At its simplest, a sample is drawn by convenience from the affected population, key informants or instances. At their most complex, samples can be randomly drawn from different social strata. A good quality small evaluation would be expected to use:

- Non-random sampling for qualitative data collection
- Random sampling for quantitative data collection



Definition: Non-random sampling

Non-random sampling selects the sample based on some property of the sample.

In non-random samples the chance that any given member of a population is selected depends on their characteristics rather than being an even chance. For this reason, non-random samples are not representative of the population as a whole.



Definition: Random sampling

Random sampling draws a sample from a population where each member of the population has an equal chance of being selected.

As each member of the population has an equal chance of being selected in a random sample, such samples are said to be representative.

This section discusses the factors that evaluators need to take into account in designing their sampling strategy. Evaluators should have an explicit sampling strategy that is designed to answer the evaluation questions in a credible way.

12.1 Common sampling problems in EHA

The following sampling problems are often seen in EHA:

- Over-reliance on availability or convenience sampling in small n-studies. This is a weak approach. Morra and Rist (2009: 363) note that 'convenience samples are a very weak means of inferring any type of evidence or pattern from the data'. Convenience sampling of the affected population is very misleading if assistance was provided on a similar convenience basis as the sample will suggest that everyone was assisted.
- Insufficient random sample size to enable statistically valid generalisations from the sample.
- Inappropriate use of random sampling in small-n studies: Collier and Mahoney (1996: 57) note that for small-n samples, 'the strategy of avoiding selection bias through random sampling may create as many problems as it solves'.
- Failure to make the sampling approach clear in the evaluation reports, especially when small-n methods are used.

Daniel (2013) provides a thorough introduction to probability and non-probability samples. A good simple and basic resource for sampling is the Practical Guide to Sampling Booklet produced by the UK's National Audit Office (2000).

12.2 Non-random sampling

Random sampling is often inappropriate for the qualitative approaches that predominate in much EHA. Daniel's guide on sampling (2013, Chapter 3) suggests that non-random sampling may be appropriate if:

- The research is exploratory
- There is a need for a quick decision
- There is a need to target specific individuals
- The desire is to provide illustrative examples
- Access is difficult or the population is very dispersed
- Time and money are limited but there is access to skilled and highly trained personnel
- A sampling frame is not available
- Qualitative methods are being used
- There is a need to use easy operational procedures
- The target size is small.

The case of difficult access is common in EHA. In one evaluation in DRC, the evaluators noted that:

The dynamic context and logistics in eastern DRC meant that it was not possible to randomly select sites in advance, especially in North and South Kivus, which witnessed frequent conflicts and displacements during much of 2012. At the same time, sites were selected opportunistically at short notice so that team members could observe ongoing activities. (Backer et al., 2013: 12)

In small-n research there is seldom a need for representative samples or to interview a manager who has been there for the median amount of time, but it is useful to interview the most long-standing manager who can provide the most information. If we want to know what factors contributed to success in microcredit, and are using a qualitative approach, we bias our sample by interviewing the most successful users of credit and asking them what led to their success. We can also interview the least successful users to find out why they did not succeed. This means that we cannot generalise to the whole population, but we can say something about success factors.

Purposive sampling is probably the strongest sampling type for small-n studies because the members of the sample are deliberately chosen for the knowledge they can contribute to the research. This is why data saturation is used as the criterion for the sample size in non-random samples.



Definition: Data saturation

Data saturation occurs when new cases no longer add new knowledge.

The risk with small-n samples is not that they are unrepresentative, but that they do not include all major categories.

Types of non-random sampling strategy

Daniel (2013) categorised four types of non-probability sampling – availability, purposive, quota and respondent-assisted. The appropriate sampling strategy depends on the method and the type of question that the team is trying to answer.

Availability sampling



Definition: Availability sampling

Availability sampling is a sampling procedure where the selection is based on their availability, researcher convenience, or self-selection.

It is commonly called convenience sampling, and Daniel reserves this name for one sub-type. While Patton states that 'convenience sampling is neither strategic nor purposeful. It is lazy and largely useless' (2014: Module 39), Daniel notes that it is 'the most frequently used sampling procedure in research'. Availability sampling has the advantages of being inexpensive, simple and convenient. Its major weaknesses are that it is the least reliable non-random method, over-representing the most available, and leads to underestimating variability within the population. It should be used only when there are no feasible alternatives.

The worst possible combination is to use convenience sampling when the assistance was also provided on a convenience basis.

Purposive sampling



Definition: Purposive sampling

Purposive sampling selects the sample based purposively so that the sampled elements can provide the most information for the study.

Purposive samples can be selected based on:

- Their reputation and experience. The team may therefore favour talking to the most experienced humanitarian managers, or to the affected population with the greatest exposure to what is being evaluated.
- Being judged to be best placed to provide the information needed to answer the evaluation questions. For example, if there is an evaluation question about whether vulnerable groups were treated with dignity, the members of the vulnerable groups are best placed to answer.
- Being judged as likely to be the best at demonstrating a particular component of the theory of change. Such cases can disconfirm components of the assumed ToC.
- Whether they are thought to have had the most typical experience. This is used for descriptive questions that ask about the most common experience.
- Whether they are thought to be outliers in terms of the typical experience. These cases are often critically important for questions of causation, because outliers may be disconfirming cases of the underlying ToC. For example, an evaluation of a feeding programme might take a sample of mothers who dropped out early as well as from those who remained in the programme for the longest time.
- Their diversity, so that the group is inclusive. Thus we would try to include lone-parent families as well as nuclear families, youth as well as elderly and so on. For small-n research, samples are often chosen to be inclusive rather than representative. This is because each case has to add meaning and data to the research.
- Their ability to indicate whether the theory of change is correct or not. These types of cases are again critical for confirming or disconfirming particular theories of causation. For example, if the ToC stated that improved nutritional status depended not only on food, but also on hygiene and access to healthcare, we might draw samples that met only some of these conditions to test the ToC.

Quota sampling



Definition: Quota sampling

Quota sampling divides the population into mutually exclusive subcategories, and then collects information for a previously established sample size or proportion for each category.

For example, if we were sampling households from a population that received assistance we might set a quota with at least two of each of the following:

- Households with young children
- Households only with elderly parents
- Female-headed households
- Male-headed households
- Child-headed households
- Wealthy households
- Poor households
- Households near the road
- Households far from the road
- The samples drawn within each quota can be drawn on an availability, purposive, or even respondent-assisted basis. Quota sampling can help to ensure that the samples include sufficient women, children, elderly, or other categories of interest.



Tip

Use quotas if you have to use availability sampling.

If you have no option but to use availability sampling, introduce quotas so that the sample is at least more inclusive. You can use quotas in availability sampling by selecting every second interviewee to be female, every fifth one to be a young male, and so on.

Respondent-assisted sampling

The most common use of respondent-assisted sampling in EHA is the use of snowball or chain-referral sampling with key informant interviews. In snowball sampling, each interviewee is asked to suggest the names of other potential interviewees (Goodman, 1961) who could speak authoritatively on the topic. The chain continues until no new names emerge. See Biernacki and Waldorf (1981) for a good discussion on the technique.

Figure 12.1: Snowball sampling



Snowball sampling is open to a number of criticisms (Erickson, 1979) and other approaches are sometimes advocated, including:

- Respondent-driven sampling (where payments are made to both the person suggesting other contacts and the other contact) (Heckathorn, 1997; Wang, 2005)
- Targeted sampling, where there is some initial research using ethnographic mapping to identify the sample population (Watters and Biernacki, 1989).

Such criticisms of snowball sampling and the alternatives apply to sampling hidden populations such as intravenous drug users rather than to sampling humanitarian managers. Snowball sampling can often lead to dozens of interviews. It is a good technique for overcoming any unintended selection bias in the initial interview targets.

The research by Guest et al. (2006) on the number of qualitative interviews needed for saturation suggests:

1. Using snowball sampling for identifying interviewees may be inefficient as few new themes emerge after 6-12 interviews with a particular group.
2. The conventional final interview question, 'Who else would you suggest that we talk to about this topic?' should probably be supplemented by 'We are interviewing groups A, B, and C. What other groups do you think it would be useful to talk to?'



Tip

Categorise interviewees into groups and aim to have 6-12 per group while maintaining a check on whether new themes and data continue to emerge.

Non-random sample size

The small-n researcher still has to make many sampling decisions, not only about the sampling strategy but also about sample size. Usually the indication of adequacy for small-n samples is that of data saturation. However, there is very little operational guidance on how many interviews are needed to reach saturation.

Using an evidence-collation tool such as the evidence table described in [Section 8: Inception phase](#), where evidence is recorded against a pre-established coding scheme, can help to identify saturation. Such an evidence tool could be made more useful for this purpose by recording whether a point made in a particular interview or source is new or simply reinforces points already captured. This will help to identify data saturation more quickly.

Creswell and Clark (2011: 174) suggest that four to six cases are adequate for a case study and 20-30 interviews for an interview-based study. However, these seem to be rules of thumb based on the authors' practice.

For qualitative household interviews, Few et al. (2014: 58) suggest a minimum of 20 for a community of 500 households and a minimum of 50 for a community of 5,000 to ensure that an adequate range of opinions is captured. Few et al. also suggest that stratifying this sample to ensure coverage of different social groups and so that households can be picked at random to minimise bias. It should be noted that samples of this size are seldom generalisable to the whole population.

For focus groups, Morgan (1997) advises that few new themes emerge after the third group. A study of data saturation in key informant interviews (Guest et al., 2006) concluded that data saturation had been reached by six to 12 interviews. In both cases Morgan and Guest are referring to homogeneous groups. Most humanitarian action has a range of stakeholders, each of whom should have their voice heard. Thus if you were looking at education and youth, you might have three focus groups each with male youths, female youths, teachers and parents. Similarly you might need a series of interviews with local community leaders, women's group representatives, NGO field staff, NGO managers and so on.

Credibility is a key issue in evaluation. An evaluation can be useful only if it is deemed credible. Even though key informant interviews with a sample of 12 from a given group may achieve data saturation, this is not the only factor to be considered.

Further interviews may:

- Make the evaluation seem more credible by having sought the opinions of a larger number of respondents.
- Reinforce the themes from the first 12 interviews, providing triangulation from these themes.

Triangulation is discussed in [Section 13: Field methods](#). Triangulation is critical in qualitative and mixed methods, which means that evaluations should go beyond the number of interviews and cases needed for simple data saturation.

One approach might be to use group interviews, workshops, or after-action reviews as a more efficient way to consult a larger number of respondents, over and above the key informant interviews. In this case the evaluators should strive to achieve data saturation for each of the methods used. If different methods identify the same themes and issues this constitutes a type of triangulation for the evaluation results.

What's wrong with using random sampling for small-n methods?

At first glance, using random sampling should help prevent bias in sampling both for small-n studies and for large-n studies. Gerring (2007: 87-88) argues, however, that while a collection of small sample sizes will on average have the same mean value as the population, the individual samples may have mean values that are very different from those of the population. Thus small-n random samples may give a highly misleading estimate of the situation of the population as a whole.

With large-n methods, the fact that each new case adds only a small amount of information is not a problem because there are many cases in the sample. It is a problem in small-n studies where each case is expected to contribute significantly to the knowledge gained.

The other issue is that, because the sample was randomly selected, it may be wrongly assumed that the results can be generalised. They cannot if the sample size is too small.



Tip

If faced with a few cases to select, don't use random or pseudo-random sampling but select the cases purposively or by using quotas.

12.3 Random sampling

Random sampling is essential for large-n methods where there is a need to generalise from the sample to the whole population. Unless randomly selected samples have been used, it is impossible to make valid generalisations from the sample to the whole population because the sample is not representative. True random sampling requires a sampling frame and a list of the whole population from which the random sample can be drawn. There is seldom a full list, but there may be some approximation, such as beneficiary lists, from which the sample can be drawn.

Pseudo-random sampling

In many cases, pseudo-random methods may be used rather than true random sampling. This is common where there is no full sampling frame. Pseudo-random methods include taking a random (or pseudo-random) starting point and sampling at an interval after that.



Definition: Pseudo-random sampling

Sampling where there is no sampling frame. Typically, the first instance is randomly selected from a purposively selected starting point, and subsequent instances are selected using some rule.

Various rules can be applied, such as picking the centre point of a group on a map or aerial photograph and then starting off in the direction set by a pen that is spun on the ground. Two interval rules are applied: fixed interval rules (e.g. every tenth house is selected – skip nine houses) or variable interval rules, using a random number sheet or a random number generator.



Tip

Use a random number application on your smartphone to generate random interval values, for instance, if you wanted to interview one in every ten houses on average you would set the random number values to a range of 1 to 17 to generate random intervals with an average of 9.

It is important to be aware that statistical inference is based on true random sampling and using pseudo-random methods may introduce an unknown bias.

Cluster sampling and the design effect



Definition: Cluster sampling

Cluster sampling is sampling where a number of locations are sampled, each with a cluster of a particular number of cases.

The main advantage of cluster sampling is that it can be used without a sampling frame for individuals in the population.

A common cluster-sampling arrangement is 30 clusters of 30 individuals but other arrangements are possible. Cluster sampling needs a larger overall sample, but is an effective technique in the absence of a good sampling frame or where the population is dispersed over a wide area. In the latter case, cluster sampling allows faster and more efficient fieldwork. Cluster sampling is widely used for nutrition and immunisation surveys. Cluster sampling may use random or purposive sampling for picking the clusters, and random or pseudo-random sampling within the cluster.

Use a sample calculator to make an accurate estimate of the sample size you need, and then add a safety margin of at least 10% for non-response. If repeated surveys are planned, you should allow for attrition between surveys.



Definition: Design effect

The design effect is the factor by which you have to modify your sample size when you depart from simple random sampling.

A design effect of two (doubling the sample size) is traditionally used for cluster-sampled nutrition surveys (IFAD, 2014), but one review of cluster sample design effects (Kaiser et al., 2006) found that a design effect of 1.5 was adequate for most surveys of acute malnutrition. The same study found that cluster surveys of mortality due to conflict had design effects of four or more. In their study of mortality during the 2000 food crisis in Ethiopia, Salama et al. (2001: 564) used a design effect of four for mortality.

The design effect arises because, due to similarities between cluster members, each new cluster member provides less new information than an independent new member. Thus the design effect is higher for situations in which the cluster members are more likely to be similar.

Stratified sampling

The only sampling strategy with a design effect of less than one is stratified random sampling, where a marginally smaller sample may be needed than for simple random sampling.



Definition: Stratified sampling

Stratified sampling is a sampling approach where the population is first divided into mutually exclusive segments and a simple random sample is taken from each one.

Stratified sampling can be useful if there is a concern that particular minorities might be missed.

Estimating the needed size for random samples

Estimating the sample size need depends on whether a sample is being used to generalise about a population or to compare two populations.

It also depends on whether you are looking at the proportions in a particular category (malnourished, not malnourished; received a shelter kit, did not receive a shelter kit) or the mean values of some variable (household income, percentage of relief food sold, litres of water used per day).

It is a common fallacy that the sample size should be proportional to the population, but smaller samples can be used with small populations. Daniel suggests that it is worth correcting for the actual sample size when this represents more than 5% of the population.

Once the population drops below approximately 400 the sample size needed rises above 50% of the population (in the worst case). Taking a census becomes more attractive as the sample size grows as a proportion of a population.



Tip

Use an online calculator for sample size to estimate the sample size you need. The following online calculators have the advantage that the underlying formulae are presented with definitions of the terms used.

- Estimating the proportion for a population:
www.select-statistics.co.uk/sample-size-calculator-proportion
- Estimating the mean for a population:
www.select-statistics.co.uk/sample-size-calculator-mean
- Comparing the proportions for two groups:
www.select-statistics.co.uk/sample-size-calculator-two-proportions
- Comparing the means from two groups:
www.select-statistics.co.uk/sample-size-calculator-two-means

When we estimate the value of some parameter of a population from a sample we make an estimate that has a particular likelihood of falling with a given confidence interval.



Definition: Confidence interval

The confidence interval is the range within which we expect the value in the population as a whole to fall.

The confidence interval is usually expressed as +/- a percentage: +/- 5% is the most common. It is also called the margin of error. We cannot be absolutely sure, however, that the value in the population will fall within the confidence interval. In a small percentage of cases it may fall outside.



Definition: Confidence level

The confidence level is the probability that the value in the population as a whole will fall within the confidence interval.

A confidence level of 95% means that there is a 5% chance that the value in the population falls outside the confidence interval – that is, a 5% chance of a false positive result.

When we are estimating a population mean we have to consider the variability in the population (estimated from the sample variability).



Definition: Standard deviation

The standard deviation is a measure of the variability of a parameter.³

Two further factors are important in comparing groups: the effect size and the statistical power.



Definition: Effect size

The effect size is the proportionate difference between the variable of interest in the treated and control group.

A common problem in EHA is that planners often overestimate the effect of their intervention. Very large samples are needed when the effect level is small. Conversely, a large effect level needs only small samples.



Definition: Statistical power

Statistical power is the probability that a negative result is not a false negative (1 minus the risk of a false negative result).

A power of 80% means a 20% probability of failing to detect a significant difference, if there is one – in other words, a false negative result. In medical trials, false negative results are usually more acceptable than false positive results, hence the traditional wider margin of error. It is, however, questionable whether evaluations to determine whether interventions have had any impact can use such a low power. A power of 90% (with a 10% chance of a false negative) may be more appropriate. The sampling guide from the Food and Nutrition Assistance Project suggests using a power of 90% wherever resources permit (Magnani, 1997).

Table 12.1: Factors influencing sample size

	Estimating a summary statistic for a population		Comparing two populations (between two group or between before and after)	
	Estimating a proportion	Estimating a mean	Comparing two proportions	Comparing two means
The confidence level	Typically 95% – a 99% confidence level needs a 70% larger sample, while a 90% confidence level needs a sample that is 30% smaller			
The confidence interval	Typically +/-5% – using +/- 3% needs a sample size that is 2.8 times the size of the sample for an interval of +/-5%			
The base proportion	A proportion of 50% needs a sample nearly three times greater than a proportion of 10% or 90%		A base proportion of 10% many need a sample size that is 10 times or more greater than for a 50% proportion	
The variability		The sample size needed increases four times when the standard deviation doubles		The sample size needed increases four times when the standard deviation doubles
The effect size			The sample size needed increases four times when the effect size falls by half	
The power			A power of 90% needs a sample that is one-third bigger than a power of 80% (at a 95% confidence level)	

12.4 Sampling for mixed methods

Only large-n methods or censuses can confidently answer questions about what percentage of a population obtained a particular benefit or experienced a particular outcome. Only small-n methods can offer good explanations of why this happened. It could be argued that large-n methods with factor analysis can offer explanations, but only for those explanatory factors that are explicitly included in the study. The strength of qualitative small-n methods is that they allow the evaluator to identify factors and issues that come to light during the research. This is one of the reasons why qualitative methods are so useful in EHA in what are often chaotic contexts.

The different strengths of large-n and small-n methods to indicate both what happened and why, explain why the best EHAs use mixed methods.

Mixed methods require both random and non-random sampling approaches, but whether they use large-n or small-n methods, humanitarian evaluators should explain their sampling strategy and choices. They should also explain any potential biases or other limitations inherent in their choices.

13 / Field methods

This section focuses on the methods used to collect data other than document review methods. Document review methods are discussed in [Section 10: Desk methods](#).

After an initial discussion of interviewing in general, a series of methods is discussed. For each method, the section discusses:

1. The extent to which the method is used in EHA
2. How it is done
3. What sort of data is collected
4. What sampling is generally used
5. Whether they are expensive or low cost
6. How many might be done in a small EHA with two weeks' fieldwork

The section also provides tips for each of the methods discussed.

The analysis of the data collected is dealt with in [Section 16: Analysis](#).

The development of the evaluation matrix is discussed in [Section 8: Inception phase](#).

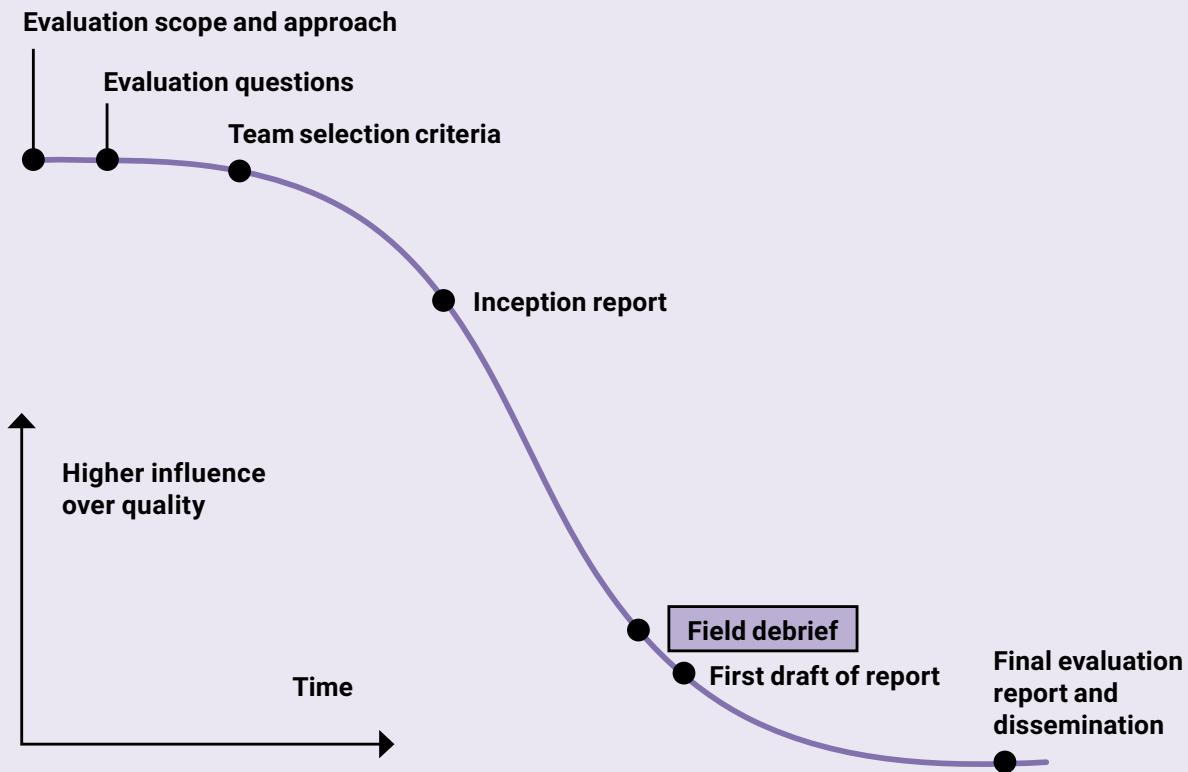


Keep in mind

The evaluation questions determine the methods used, but practical considerations of budget, time, ethics, and logistics may prevent the most appropriate methods being used for a particular question.

Evaluations can use a variety of methods to measure what they want. For example, the IRC evaluation of reconstruction in Lofa County used behavioural games as well as surveys to test if communities reached by the project behaved differently from control communities: 'we sought direct measures of behavioral effects of the CDR programme as revealed by behaviour in "public goods games"' (Fearon et al., 2008: 25).

Quality checkpoint: Field debrief



Debriefing in the field can help identify major gaps and influence the analysis of the data collected.

13.1 Interviewing as a technique

Interviewing lies at the heart of many social research methods and can be divided into three categories:

1. **Structured interviews.** These are used for quantitative surveys and interviewers follow a strict protocol for asking questions.
2. **Semi-structured interviews.** Very widely used in EHA. The interviewer probes the interviewee to draw out data and meanings.
3. **Unstructured interviews.** In these, the interview is an open conversation although it may have a central theme, such as a life history. Little used in EHA.

Interviews are dialogues, not just a transfer of information. This is as true of formal survey interview as of qualitative key informant interviews (Suchman and Jordan, 1990). Research on formal survey interviews shows

that responses vary not only according to how the question is asked but also with the interviewer's gender, ethnicity, age, apparent social class and body size (Dohrenwend et al., 1968; Eisinga et al., 2011; Flores-Macias and Lawson, 2008; Kane and Macaulay, 1993; Riessman, 1977; Suchman and Jordan, 1990; Webster, 1996).



Keep in mind

In some contexts interviewing people may put them at risk. Do not conduct interviews in such circumstances unless the interviewees are fully aware of the risks and accept them. While the interview may put them at risk, they also have the right to have their voices heard (see [Section 14: Engaging with the affected population in your evaluation](#)).

Structured interviews

Structured interviews are usually conducted in quantitative surveys. In order to minimise bias and the resulting errors in estimates, the interviews are carried out in a standardised way. Interviewers do not probe beyond the set questions.

Interviewers are usually called enumerators and they may be provided by a market-research company doing the survey or may be university students. Survey enumerators do not need the same levels of skills and context understanding that qualitative interviewers need, and therefore tend to be less expensive to employ.

Enumerators are required to ask the questions in a set sequence and should not explain the question in order to avoid biasing the responses since other respondents might have given a different response if they had received the same explanation. To emphasise its formality, the questionnaire used is usually referred to as the survey instrument.



Definition: Survey instrument

Survey instrument is the questionnaire used by the interviewer during formal survey interviews.

A survey instrument may have a large number of questions, organised into sets to examine different constructs such as leadership or skills. The survey instrument for teachers in the NRC review of Palestinian Education (Shah, 2014) had 91 questions measuring 10 underlying constructs. A standard part of the survey instrument is testing how well each question relates to the others for that construct (the internal reliability).

The questions are drawn from the survey instrument. It is essential to pre-test the survey instrument in order to identify and rephrase questions that might lead to queries or uncertain responses from participants (see Good practice example on pg 251 for WVI's practice on translating and back translating).

Although structured interviews may include some open-ended questions, these pose two problems:

- The response needs to be coded in some way so that it can be analysed, which can be expensive and time-consuming.
- The response to the open-ended question may raise other questions that the enumerator is not permitted or equipped to follow up on.

Semi-structured interviews

Semi-structured interviews are at the base of a range of methods widely used in EHA. In semi-structured interviews the interviewer uses an interview or topic guide and asks questions around this. In this approach the interviewer follows up on earlier responses and probes to uncover data and meanings.

Probes⁴ are questions that are intended to:

- **Clarify responses** – when an interviewer is uncertain what the interviewee meant
- **Get more details** – the most common use of probes in EHA interviewing
- **Get their analysis** – can be very useful for understanding why individuals acted in a particular way
- **Ask about variations and counterfactual cases** – useful for highlighting the rationale for decisions taken.

The probes used will depend on the evaluation questions and the scope of the evaluation.

Clarifying probes

These may include:

- Did I understand you correctly that...? This is often the best clarifying probe as it does not suggest that the interviewee was not clear in their earlier response.
- When you say ... what exactly do you mean?
- You used the term ... Am I correct in thinking that you are referring to...?

Detail probes

These may include:

- Can you give me an example?
- How did you deal with that?
- What response did you get to that?
- Can you tell me more about your experience?

Analytical probes

These may include:

- How would you characterise what happened?
- What was it about that stands out in your memory?
- What was important about that?
- What did you expect the results would be?
- How did you feel about that?

Variations and counterfactual probes

These may include:

- Has your approach changed over time? How has it changed?
What drove this change?
- Would you deal with it the same way the next time?
- I notice that you are doing ... but agency Y is doing ... with the same issue. What advantages does your approach have over theirs, and vice versa?
- Some of the people I talked to said ... What is your take on this?

Unstructured interviews

These interviews can range from the completely unstructured interviews such as may be used in ethnographic research to thematic interviews, where the interviewer introduces a theme, such as the story of the interviewee's life, and lets the interviewee run with it.

Unstructured interviews have not been used much in EHA, but some interview methods, such as focus group discussions, include an element of unstructured interviewing. One example of the use of thematic unstructured interviewing was the Disaster Assistance Committee (DAC) evaluation of the 2000 Mozambique Floods. Here, an evaluator conducted long interviews with members of the affected population to build a view of their experience of the aid response (Cosgrave et al., 2001).

Bias in interviewing

Clearly, all interviewers and interviewees have an agenda. Some agency interviewees may want their work to be widely known, others may want to conceal flaws. The affected population may be keen to see assistance continued, or may prefer to see it stop. The evaluator needs to remain aware of the potential for such biases, and that interviews are constructed processes, between the differing interests and values of the interviewer and interviewee. The analysis chapter deals with how to address bias in interview evidence.

As with all qualitative or mixed methods, triangulation is the most effective technique for controlling bias. Triangulation is at its most powerful when the purposive selection of interviewees is as inclusive as possible. Triangulation, dealing with interview bias, and data conflict, are discussed further in [Section 13: Field Methods](#).



Tip

When conducting interviews with the affected population, the timing of interviews has a large influence on who is available for interview. Evaluators should be conscious of whom they are excluding by interviewing at particular times of the day.

See [Section 11: Evaluation designs for answering evaluation questions](#) on bias in general.

13.2 Interpreters

Ideally, evaluators should be able to speak the language of the people they are interviewing. If not, they need to work with interpreters for some types of interview. Interpreting always raises problems in evaluation. It is not seamless and Temple et al. (2006) argue that interviewing through an interpreter is more akin to analysing secondary data than primary research.

While interpreters can be used for semi-structured interviews and group interviews, it is almost impossible to use interpreters to conduct effective structured interviews or focus group discussions. Survey interviews already take considerable time and adding interpretation makes them even longer. In focus groups, the mechanics of interpreting interfere with the concept of free-flowing exchanges.

For focus groups, Steward et al. (2007: 112) note that:

Moderators should speak the language in which the group is conducted as a first language. Even the most fluent person in a second language will have some difficulty framing questions and following up responses. Interpreters create even greater problems and should be avoided whenever possible. It is often the case that the perceptions and opinions of the interpreter colour the translation.

Needing to rely on an interpreter increases the time required to work through a set of questions. This can sometimes be useful in qualitative interviews as it gives the evaluator more time to reflect on the answers to previous questions while framing follow-up questions. Evaluators should make every effort to engage qualified interpreters. Knowledgeable interpreters can also add value by highlighting contextual issues in discussions after the interview that the evaluator might otherwise have overlooked.



Tip

Get the best possible interpreters you can afford. Professional interpreters are well worth their fee as they are less likely to distort the question or responses so allow for this in the budget.

Remember that interpreters need time to be recruited. Travel and accommodation arrangements should allow for presence of interpreters.

In one evaluation in Afghanistan, community elders were asked about the effect of the Taliban on their access to government services. The agency staff person acting as interpreter told the team that the elders complained that the actions of the Taliban had reduced their access to services. Later, the evaluation team's assistant told the team that what the elders had actually said is that they were the Taliban and that the government was refusing to provide them with services unless they surrendered their weapons.



Tip

Avoid using agency staff for interpreting if possible.

Although very far from ideal, it is not always possible to avoid using agency staff to interpret because professional interpreters are not available for the relevant languages, or because of security or access restrictions.

In the evaluation of CARE's response to the 2011-2012 humanitarian crisis in the Sahel (Gubbels and Bousquet, 2013: 18), the team noted that when conducting focus group discussions it was:

Obliged to rely upon one of CARE Chad's animators as an interpreter. This was a potential source of bias. However, by using probing questions when necessary, and assessing the overall pattern in responses across six villages, the evaluation team believes the issue of positive bias was avoided.



Tip

Cross-check interpretation by those with a vested interest. If you have no knowledge of the language and have to rely on agency staff to interpret, it is good to have an assistant who understands the language and can brief the team if there are any major inaccuracies. You can also openly record exchanges for later analysis to encourage accurate interpretation.

13.3 Qualitative interview methods

Key informant interviews

Key informant interviews are the backbone of EHA.

Key facts:

- They use the semi-structured technique and an interview guide.
- Typically take from 45 to 60 minutes.
- Provide qualitative data, best for issues of meaning and understanding.
- Good for answering questions about how and why things happened.
- Poor for gathering quantitative data.
- Need experienced interviewers who can probe to draw out meaning.
- Usually recorded with written notes, but may be recorded on a digital device.
- Sampling is usually purposive. The interviewees are selected to be the people best expected to answer the evaluation questions.
- The method is quick and relatively inexpensive.
- Two weeks' fieldwork might include 25 to 50 key informant interviews.
- Typically no more than four or five a day.
- Theoretical saturation (further interviews yield no new information) for a particular type of interviewee occurs at 6-12 interviews of a particular type (Guest et al., 2011) but more may be needed for credibility.



Keep in mind

Main ethical issues are that interviews should not endanger the interviewee and that confidentiality should be respected.



Tip

You can speed up interviews slightly by providing interviewees with details of the objective and the ground rules in advance.

Key informant interviews are usually conducted in a private space so that others do not overhear or join in. You can also use public spaces that have privacy norms, such as a coffee bar.

The evaluator is often perceived as a representative of the aid community. In many cultures it is rude to criticise something to someone assumed to represent the organisation that provided it. In such cultures it can be more useful to ask about how their neighbours view the assistance provided.



In depth: Interview guide

The interview guide sets out the general trend of the questions, starting with a reminder to tell the interviewee about the purpose of the evaluation and what the basic rules are (typically interview comments are non-attributable, directly or indirectly).

It is good to start with a question asking the interviewee about their role – almost everyone is happy to talk about themselves, even if they are shy. Not all the interview guide questions will be relevant for all interviewees and the interviewer should not ask irrelevant questions.

The questions should be sequenced in order to build rapport – avoiding the more contentious questions until later on in the interview.

The interview guide should cover at most 20 to 25 questions. If there are more, it will be necessary to prioritise or prepare separate guide for different types of interviewees.

The closing questions should include one about what they have learned or had reinforced by the response. Asking whether the interviewee was surprised by any area that has been omitted can help to resolve understanding about the intent of the evaluation and sometimes provides new leads.

The closing questions should also ask for suggestions for further interviewees and groups of interviewees. Ask if the interviewee would be available for further questions later on in the process if needed.



Tip

Avoid questions that elicit the 'right' answer. For example, if you want to know whether the implementation team paid attention to gender in the intervention, don't ask: 'Did you pay attention to gender?' as the answer will always be 'yes'. Instead ask: 'Did women or men benefit more from the intervention and if so why? What differences were there between the effects on men and on women?'



Household interviews

These are commonly used as part of quantitative surveys, but can also be used to collect qualitative data.

Key facts:

- Can use a formal survey instrument or an interview guide or a combination of both.
- Can take from 20 to 60 minutes depending on the complexity.
- Can gather qualitative and quantitative data, depending on the type of interview.
- Good for answering questions about the effect of assistance at the household level.
- Need interviewers who understand the context well and can pick up clues from observation to triangulate what they are told.
- Usually recorded on a survey form, but may be recorded on a digital device.
- The site for the interview is chosen purposively, but the households may be selected randomly (or large-n studies) or purposefully for small-n studies.
- The method is moderately expensive, unless it is part of a regular monitoring system.
- Usually done as a separate exercise if intended to be representative.
- The number of interviews per day depends on the complexity and travel time.

When used to gather qualitative data they generally only cover a few topics and are much shorter than a full key-informant interview. Their main advantage over group interviews is that the household is a private space and it is possible to address more sensitive issues than in group meetings.



Tip

Household interviews may provide good data for illustrative cases that highlight particular issues. Such illustrative cases can bring the human dimension home to readers and brighten up the evaluation report.

Group interviews

Key facts:

- They use the semi-structured technique and a truncated interview guide as at most five to ten questions can be discussed.
- Typically take from 30 to 45 minutes, but can take longer if the interviewees are interested.
- Gather qualitative data, best for issues of meaning and understanding.
- Good for answering questions about community perceptions of aid.
- Poor for gathering quantitative data.
- Need experienced interviewers who can relate to the community and probe to draw out meaning.
- Usually recorded with written notes, but may be recorded on a digital device.
- The site for the interview is chosen purposively, and the initial interviewees may also be chosen purposively, but as they are generally conducted in a public space, others may join in.
- The method is quick and relatively inexpensive.
- Two weeks' fieldwork might include 5-20 group interviews.
- Typically no more than 2-3 a day.
- There are no standard estimates for theoretical saturation for such interviews, but they can be expected to be 2-3 for focus groups and between 6 and 12 for key informant interviews.
- Prone to being dominated by certain individuals especially in hierarchical societies.



Keep in mind

May raise major ethical concerns as they are conducted in public, and speaking in public may place interviewees in danger.



Tips

- Manage the group. If you are meeting outside, don't remain standing but sit on the ground if possible. This will oblige the group to sit down and there will be less pushing and shoving as a result.
- Learn about local social norms and conform to these as far as possible.
- Steer questions away from the most vocal contributors. Ask how many agree with any particular viewpoint.

Formal meetings with the community structures

One variant of the group meeting that is commonly seen in parts of Asia is a meeting organised by community structures. These can be quite formal with a top table and someone to chair the meeting. Such meetings tend to be dominated by community leaders, which can make it difficult to elicit information from the wider community.

The best approach is to use questions that you expect will have different answers for different people in the group. Questions about, for instance, who was worst affected, whether men or women benefited most, which type of assistance was the most useful in the first week should elicit some differences in views. It can take some effort to start getting spontaneous responses, but when they get going, such meetings can provide useful data.



Good practice example: Establishing a dispute-resolution policy

In a UNHCR evaluation of age, gender, and diversity mainstreaming (Thomas and Beck, 2010), the Colombia country study used a highly participatory approach. Four communities in different parts of the country were selected to participate in the evaluation through the following steps (Mendoza and Thomas, 2009):

- A workshop with community members to construct a timeline of events leading up to UNCHR's participatory assessment (a key tool in the mainstreaming process) and the resulting action plan – to gauge awareness of the action plan and to evaluate whether the participatory assessment had generated changes in the community.
- Meetings with sub-groups of women, men, adolescent girls and boys, children, older people and people with disabilities – to ascertain whether different groups had different perceptions of the results of the participatory assessment and action plan.

Focus group discussions

Though the term focus group is very widely applied to group interviews, it is best reserved for structured interviews.

Key facts:

- Typically last two hours.
- Typically 6-8 participants.
- Based around a topic guide with 3-5 discussion topics.
- Most effective with partner agency staff, but can also be used with affected population.
- Generates qualitative data, good for exploring the views of particular groups (single parents, youth etc.)
- Need a comfortable, controlled space.
- Need a facilitator and a note-taker.
- Typically videoed or recorded ([see Section 14: Engaging with the affected population in your evaluation](#)).
- Samples may be randomly drawn from within groups to form the focus group. The group must be similar to each other without status or other differences.
- Morgan (1997) advises that few new themes emerge after the third focus group.
- Medium cost.
- Typically fewer than ten focus group discussions in a two-week evaluation.

- In some countries focus group participants may be paid for their time. This is not the practice in EHA, but at the least they should be offered refreshments for what is a relatively long time commitment.

The standard references for focus groups are Krueger and Casey (2009) and Steward et al. (2007).



Tip

Interact with the group before the discussion to spot the most talkative. Have a snack together before the start of the focus group discussion to identify those who may dominate. Place them next to the facilitator so that there will be less eye contact. Put the quietest opposite so that the facilitator looks at them more often.

In analysing the record of the focus group you will be looking at words, context, changes of view, frequency of particular comments, amount of discussion, intensity, and specificity, as well as the content.



Good practice example: World Vision's use of focus group discussions

"For FGDs, we typically have a team leader on every team, though on occasion we are unable to do that so we use the regional structure (could be two or three FGD teams in a region with one FGD team leader who supports them all, for example). But our FGD teams are comprised of four people at a bare minimum, and up to six: team leader, translator, two note takers, observer and facilitator. The observer and facilitator are opposite genders and switch roles as facilitator depending on the gender of the FGD; the observer, team leader and translator can also jump in to become note takers if the discussion is very lively."

Source: Kathy Duryee, WVI, personal communication, 2014

13.4 Survey methods

Using surveys

Surveys, especially formal face-to-face surveys, are regarded as very authoritative. In part, this is because many readers lack a sound understanding of statistics. Consider the following example:



Political poll reports a 2% fall in support for a party ...

what does that mean?

Many political opinion polls use a 95% confidence level and a margin of error of +/- 3%. The poll finds that support for the governing party has fallen from 35% in the last poll to 33% now.

Common assumption:

33% of the population now support the governing party.

- **Actual situation:**

33% of those sampled supported the governing party.

Common assumption:

There is a 95% chance that 33% of the population support the governing party.

- **Actual situation:**

There is a 95% chance that between 30% and 36% (33% +/- 3%) of the population supports the governing party. There is still a 5% chance that support is lower than 30% or higher than 36%.

Common assumption:

Support for the governing party has fallen.

- **Actual situation:**

Previously the level of support in the population for the governing party was 95% likely to be between 32% and 38% (35% +/- 3%), now it is between 30% and 36%. It is even conceivable that popular support has increased from 32% to 36%. However, not all outcomes are equally likely, as sample means are normally distributed around the population mean. The result is that it is likelier that the population mean is close to the sample mean than at the limits. This is why opinion polls seldom show large swings.

It is essential to know a bit about statistics in order to understand survey reports. While teams may report overlapping margins of error in before and after surveys, they focus on changes in the mean value of the variable of interest rather than point out the possible conclusion that there has been no change.

Nutrition surveys

Nutrition surveys are widely used to monitor nutrition programmes. Evaluations tend to use existing survey data rather than carrying out their own survey, as in the case of the Rwanda evaluation of food assistance for long-term refugees (Sutter et al., 2012). Such surveys are quite technical and are not discussed further here.

Formal face-to-face surveys

These surveys may use individuals as the sampling point, but in EHA it is more common to use households as the sampling point. The unit of analysis may either be the household or a household member, selected by some pseudo-random form such as the most recent birthday, a Kish grid⁵ or quota sampling (males and females alternately to ensure gender balance). Sometimes, however, there may be cultural problems about speaking to someone other than the initial informant (McBurney, 1988).

Key facts:

- Use a survey instrument that the interviewers follow rigidly.
- The duration of each interview depends on the length and complexity of the survey instrument.
- Gather quantitative information – good for answering ‘How many?’ questions. Can also ask closed qualitative questions as open questions are very demanding in terms of coding and analysis.
- Enumerators typically have only a few days’ training.
- Data recording on a survey form or digital device. Electronic recording is far superior.
- Interviewees are randomly selected so that the survey results can be generalised to the whole population. Every departure from true random sampling (e.g. pseudo-random or cluster sampling) increases the sample size needed to give statistically valid results.
- Typically several hundred interviews are needed, but can be as high as several thousand.
- Expensive because of the need to test and validate the survey and large number of interviews needed. Its use in EHA tends to be restricted to large-scale evaluations with a correspondingly large budget.

Face-to-face surveys can be a powerful means to establish the extent of different issues, including the scale of the effect of an intervention.

Surveys of the affected population are very effective for revealing perceptions of aid, limits of coverage, timeliness, etc. If you can afford it and have not already engaged them with other methods, it is well worth doing a survey of the affected population.

The sample size needed varies according to a range of factors including whether you want to talk about prevalence only in one group or you want to compare different groups or the same group over time. See Section 11: Evaluation designs for answering evaluation questions for advice on sampling.

Fowler (2009) provides a good, succinct introduction to surveys.

Formal surveys are becoming more common in EHA, especially in large and well-funded evaluations.

Examples of surveys in EHA

The following are examples of the use of formal samples in EHA.

- The surveys of the affected population carried out after the 2004 Indian Ocean Tsunami, the 2006 Pakistan earthquake and the 2007 Yogyakarta Earthquake (Bliss and Campbell, 2007a, 2007b; Bliss et al., 2006; Fritz Institute, 2005a, 2005b, 2005c).
- The LRRD study for the tsunami Evaluation Coalition (Christoplos, 2006) with 1,227 respondents in Indonesia and 915 in Sri Lanka.
- The follow-up LRRD study (Brusset et al., 2009) with 1,178 respondents in Indonesia and 965 respondents in Sri Lanka.
- The interagency evaluation of support for refugees in protracted displacement featured surveys in each of the case-study countries including Ethiopia with a sample size of 1,081 (Sutter et al., 2011), Rwanda with a sample of 1,200 (Sutter et al., 2012) and Bangladesh with a sample of 1,069 (Nielsen et al., 2012).

Once the survey data have been collected, they need to be entered and cleaned. This can be expensive and time-consuming, especially if the survey contains open-ended questions.



Tip

Use electronic aids for collecting survey responses. This will eliminate the cost of data entry and much of the cost of cleaning, as the survey form can be set up to reject invalid combinations in responses (such as a 16-year-old informant with 12 children), and the collected data can simply be uploaded to the main database without any further data entry. This can also speed up the survey process.

Digital tools can be used for assessment surveys as well as for evaluation surveys. The Multi-cluster Rapid Assessment Mechanism in Pakistan used personal digital assistants (PDAs) to gather assessment data (McRAM Team, 2009). Shirima et al. offer an example of the use of PDAs in a survey of 21,000 households in Tanzania and provide some insight into the practical issues involved (Shirima et al., 2007). Research suggests that using PDAs has the potential to reduce the logistics burden, cost and error rate of data collection (Seebregts et al., 2009).

Mobile phones are increasingly used by enumerators to directly enter the data. These have the advantage that data are uploaded immediately, preventing the risk of data loss if the phone is damaged or lost. This also allows for real-time quality control and supervision of the enumerator, to reduce the risk of data fabrication (Tomlinson et al., 2009). Mobile phones offer similar advantages to PDAs in reducing recording and entry errors (Zhang et al., 2012). Several software packages such as SurveyToGo or Snap Surveys can be used for setting up mobile-phone-based surveys. Different products have distinct pricing models, and vary in their support of specific devices. This field is developing very quickly, so it is important to check what is currently available and most suitable.

Surveys pose a number of logistic problems in EHA:

- **Lack of a convenient sampling frame.** A sampling frame is a more or less accurate list of the population from which the sample is to be drawn. The alternative when there is no other sampling frame is to use cluster samples, but this increases the number of interviews needed by a factor of two (nutrition) or four (mortality). See [Section 12: Sampling](#) for a discussion of the design effect.
- **Time.** Developing and testing the survey instrument can take time. Obtaining translations, training enumerators, and conducting the survey all consume time, and surveys often take longer than other methods in EHA.
- **Lack of baseline surveys.** Surveys are most useful when there is a baseline against which to compare the results. This is rarely the case in EHA, but the situation is improving in line with improvements in national statistics in many developing countries.
- **Lack of evaluators with the requisite skills.** Most EHA evaluators are from a qualitative rather than a quantitative background and have relatively little experience of conducting quantitative surveys.

Designing a survey instrument is a complex task. Survey questions, though simpler in some respects than semi-structured interview questions, are more complex in other respects, especially as the interviewer cannot adjust the question and probe for follow-up responses. Structured face-to-face surveys are based on the assumption that the questions will be asked in the same way of each interviewee. Survey questions need to be carefully designed and tested before being used, as they cannot be adjusted once the survey is under way.

The usual way to test survey questions is to conduct a pilot survey. How questions are asked can have a significant effect on responses, and a pilot survey may highlight some of these issues. Fowler's (1995) monograph on survey questions is a useful resource.



Tip

If the survey has to be translated, have a different person translate it back into the original language.



Good practice example: World Vision International practice on data-collection tools

Once the English version of data-collection tools is ready, the subsequent process is followed:

- 1.** A translator from the region where the survey will be administered (often local WVI staff, or sometimes a consultant) is enlisted to translate them into the local language or dialect.
- 2.** The tools are back-translated (into English). This is not always possible, but is an ideal.
- 3.** A training session is held with enumerators, during which they verify the accuracy of the translation, and ensure they all understand the questions in the same way. This is a very iterative process.
- 4.** Also during enumerators' training, a field test is done in communities using the same language as the survey tools (this can entail some logistical gymnastics if multiple languages are required).
- 5.** Following the field test, all survey responses are reviewed and any final changes to translations are made. Enumerators are also given feedback about the responses they recorded, and have a final chance to review and clarify any misunderstandings

Source: Kathy Duryee, WVI, personal communication, 2014

Online surveys

Online surveys are relatively easy to set up and conduct. They have largely replaced mail surveys and some forms of telephone surveys.

Key facts:

- Widely used in EHA.
- A very useful tool for getting views from widespread groups. Best for collecting views of staff of the implementing agency or partners, but can collect views from the wider humanitarian community.
- The survey can be completed remotely regardless of the time zone.
- Typically open for two to four weeks.
- Can gather qualitative and limited quantitative data.
- Respondents are opting in so response rates can be quite low. Respondents are usually targeted as a group rather than being randomly selected, which means the results should not be seen as representative of the wider population.

- Good for identifying potential key informants – always ask survey respondents if they would be willing to be contacted to provide further information.
- Interviewees may be invited, but are essentially self-selecting – so the findings from online surveys should not be generalised.
- Inexpensive.
- Typically from a few dozen to several hundred or more responses.
- One ethical issue is that common commercial practice is to offer survey participants some incentive, such as the chance to win a voucher or electronic device, but this is not done in EHA.



Tip

- Limit the number of questions to increase the response rate.
- At the start of the survey, state the number of questions or how long it should take to complete.
- Send several reminders to increase the response rate. If you send out individualised invitations the survey site can track who has responded and stop them from getting reminders.
- Carrying out an online survey during the inception phase allows you to identify potential key informants with strong views or well-developed analyses.
- Translating the survey into the appropriate languages may increase the participation by national staff.
- Dillman et al. (2009) is the standard reference for online surveys.

Examples of the use of online surveys

- The Stay and Deliver review of good practice in providing humanitarian assistance in different security and risk environments used an online survey in English, French, Spanish and Arabic to gather the views of 1,148 local staff (Egeland et al., 2011).
- The evaluation of the UNHCR response to the refugee influx into Lebanon and Jordan used an online survey to validate the findings of the evaluation (Hidalgo et al., 2015).
- The evaluation of the ECB project (Ky Luu et al., 2014).



Tip

Check if an agency conducts staff exit surveys – these can be a useful for identifying key issues for an evaluation. However, confidentiality should be respected.

IFRC has made use of such surveys and has set them up for large-scale emergency operations:

We did that for Haiti, for the Philippines as well. Where in the early weeks following the disaster we ask all our staff leaving the operation a limited number of questions: their perception, how well our fundamental principles are adhered to, or not adhered to, how well we are incorporating the views and opinions of the affected population. We let that run for several months. So by the time that we do a RTE or a humanitarian evaluation, we are able to provide to evaluators quite a lot of insights from the staff that has operated in the operation, often times as well those people have left. It's useful it's anonymous and it's definitely a practice I found very useful (Josse Gillijns, personal communication, 2015).

13.5 Observation

Observation is a useful technique in evaluation and is a core EHA method because it can help the evaluator understand the context better. Observation can be structured or unstructured.

Structured observation

Structured observation is where the observer uses a form, like the example on the following page, to record the observation. It has been little used in EHA but there is scope for greater use.

For example, if an agency had been training families in safe water collection, there might be an observer at the pump charged with recording behaviour to judge the effectiveness of the training. The observer would have a form with the following in the first column and columns for each person using the well. If the promoted behaviour is observed, the column for that user is ticked.

User Number:	1	2	3	4	5	6	7	8	9	...
Element of good practice										
Brought sealed container										
Open sealed container at well										
Kept cap from getting dirty										
Rinsed the container										
Rinse-water into the drain										
No excessive splashing										
Sealed container before leaving										
Note number: (if special obs.)										
Notes:										

While some structured observation systems can be complex,⁶ others can be relatively simple. Bentley et al. (1994) provide an excellent guide to the technique including several sample forms and guidance on developing them for each situation. Although structured observation is used most widely in sanitation or education there are many ways in which it could be used in EHA, including:

- Surveys of use of facilities, including water points, communal latrines, etc.
- The way in which men and women, boys and girls are treated differently at service points such as distribution centres or clinics
- The ways in which complaint desks deal with different complainants and types of complaint
- The level of attention and interaction during coordination meetings.

Not all structured observations require coding of data. In some cases the enumerator simply counts the number of people doing a particular thing in a particular time period. The enumerator does not need any specific skills other than the ability to persevere with a somewhat boring task. A structured observation form can be developed for any purpose – for looking at interactions in community meetings, for example – depending on the evaluation questions.

If the participants are aware of the observer's presence, this can lead to reactivity. In Bangladesh, Ram et al. (2010) found that the presence of observers increased hand-washing. They were about to measure this by fitting bars of soap with motion sensors. The soap moved significantly more on the days that the hand-washing was observed, and in 22% of households did not move at all on other days.

Unstructured observation



Tip

If possible, look around before engaging in individual or group interviews with the affected population. Doing so may give you information that allows you to probe responses and promote a franker exchange.

The Swiss Solidarity Tsunami evaluation team undertook systematic site walks at the start of research at each site. The team made individual notes and then compared them at the end of the day. Some team members had prior experience at some of the locations, and this gave them an important basis for comparing with the conditions they found (Ferf and Fabbri, 2014: 19).

Unstructured observation is where evaluators make observations when making transect walks or just roaming around the site. Lee (2000) identifies five types of simple observation.

- **Exterior physical signs.** How are people dressed? Are there signs of wealth or of distress or both? An example of this is where the community leaders during one evaluation in Angola insisted that the community were going hungry because of a cut in rations, but where bags of grain could be seen in many households.
- **Expressive movement.** Are children running around and playing? An example of this was in an evaluation in Malawi where a particular village did not meet the agency implementation team with singing and dancing as was the norm. Clearly the village was not happy with the implementation team.
- **Physical location is about how people organise space.** Is there evidence of shared space or are families keeping very much to themselves? Are people travelling on the roads or not? Of course some of this may be culturally determined but it can give an indication of underlying issues.
- **In situ conversations.** These can be the short exchanges that the evaluators may have as they are walking through a site. Are people complaining? What concerns are they expressing?
- **Time-related behaviour.** This is easier to capture with structured observation, but the evaluator may still notice things like people travelling from a food distribution after dark, which would suggest that distribution was not well organised.



Tip

Carry a camera where appropriate. Photographs are much more powerful than verbal descriptions of observations. Photograph interesting observations and use those in your report.

The use of filming for observations and for recording interviews is also growing, see discussion on informed consent and confidentiality in [Section 14: Engaging with the affected population in your evaluation.](#)



Good practice example: Using photographs to support key messages

In the response to the Kosovo crisis, an agency with substantial experience in the water and sanitation sector carried out a number of projects that fell below their usual standards. During the introduction to the debriefing meeting for the agency, the suggestion that the agency had failed to meet its own standards was strongly rejected by agency staff. However, the atmosphere changed when the evaluator used a number of photographs, for example a photo of a precariously balanced 5-tonne water bladder, which was located in an environment full of children, to illustrate neglect of safety rules and agency policies (Wiles et al., 2000).

13.6 Unobtrusive measures

Unobtrusive measures aim to avoid interfering in the lives of the affected population (Webb et al., 1966; Lee, 2000; Gray, 2014). As Patton (2014) points out, the aim of unobtrusive measures is to avoid the affected population reacting to the attempt to measure something.

The internet offers a wide range of unobtrusive measures, particularly in the monitoring of page visits, searches, and social media. The topics can be as mundane as researching how wardrobe malfunction affects interest in celebrities (Pfister, 2011) to the frequency of Google searches for the term 'earthquake'. Of course methods like these are unobtrusive, but this sub-section refers to non-reactive observational measures rather than to content analysis, whether of conventional documents, the internet, or other resources.



Good practice example: IFRC's use of social media as an unobtrusive measure in the response to Tropical Cyclone Haiyan

These days, every disaster generates a hashtag, sometimes more than one, and by analysing the traffic on Twitter, on Instagram ... you can find things that can be validated by your staff or by other interviews. It's an interesting tool to triangulate some of the findings that would come out of a humanitarian evaluation.

Source: Josse Gillijns, personal communication, 2015



Webb et al. (1966) suggested that unobtrusive measures can show up as wear, or building-up.

The classic example of wear measures given by Webb et al. was the observation of wear on carpet tiles as an indicator of visitors' interest in different museum exhibits. Another example of a wear measure was Patton's monitoring of coffee consumption during a week-long training to infer which sessions were the least interesting (Patton, 2014). In an EHA setting, wear on different paths can suggest the intensity of use of different water points, and so on.

An example of a building-up measure was Blake's (1981) study of graffiti in Hawaii as an unobtrusive measure of ethnic relations. In an EHA setting, an example of a building-up measure could be the rate at which pit-latrines fill up, or the rate at which rubbish is generated.

Unobtrusive measures are very powerful as they avoid the risk that the act of measurement will affect the measure. While there are some examples of unobtrusive measures in the management of humanitarian action, there are very few examples in EHA. One example of a management unobtrusive measure was measuring mortality recording stores' issue of winding sheets (for wrapping corpses for burial) among refugees in Western Tanzania. These were freely available on request but the refugees were too superstitious to request one for any other purpose than a burial. This was an erosive measure as it depleted the stocks of winding sheets.

One evaluation example of an unobtrusive accretion measure was in the inter-agency RTE of the Mozambique floods in 2007 (Cosgrave et al., 2007). Here the accretion measure was grass growing in the temporary shelters that displaced villagers had built. This was strong evidence that the inhabitants had returned to their villages and were returning to the temporary site only for distributions. An example of an erosion measure was in a refugee camp where the path to the marketplace toilet was well worn, but the path to the adjoining hand-washing station was not.

The evaluation of the rapid response to movement programme in DRC found that the collecting of grass by IDPs for temporary shelters was an unobtrusive measure of delays in issuing plastic sheeting (Baker et al., 2013: 22).

13.7 Learning-oriented methods

The methods listed here are particularly well suited to learning-oriented evaluations. These are usually relatively participatory and should involve those who need or want to be involved in the learning. Methods should be conducive to reflection and to admitting mistakes as well as successes. This list is far from exhaustive.

In the activist culture of humanitarian operations, it is often said that there is not enough time for participatory and reflective exercises. But only a small portion of humanitarian action is so time-critical that it would be harmed by taking time out for evaluation. Planned and facilitated well, a number of learning-oriented activities can be completed in a couple of hours, or a day, and the value of pausing to reflect and learn collectively merits the time investment.

After identifying who is the focus of the learning – for example, specific staff members, partners, or members of the affected population – processes need to be designed that enable them to reflect. These include the following:

- Create a safe space for reflection and discussion where participants feel they can talk freely and admit mistakes
- Ask simple, open questions, usually the most effective way of encouraging reflection (see ‘Questioning and Listening’ in Chapman et al., 2005)
- Use participatory and creative processes
- Ensure that the lessons learned are well documented
- Provide good facilitation
- Encourage those who do the reflection and learning to take ownership of the evaluation findings
- Build in a process for follow-up.

Story-telling

Telling stories is a natural way to communicate. Focusing and facilitating this process can be a creative way to facilitate learning, especially when it is accompanied by a process of questioning and reflection to deepen analysis. This sub-section describes two ways of using story-telling for learning purposes: using metaphors and the MSC technique.



Good practice example: Storytelling with Oxfam GB in Aceh

Just after the Asian Tsunami, Oxfam did an internal learning review in Aceh. To facilitate self-learning among staff and to make it more participatory, a storytelling approach was used. This was facilitated by the Global MEAL adviser for Oxfam GB, who explained the exercise as follows: 'I didn't want the usual "hello I'm from Oxfam, and we've come to evaluate the programme" because then you've already introduced bias'. Staff were asked to not wear any Oxfam-branded clothing and were dropped off outside the town. They were asked to walk around the community, sit down beside somebody and just say 'tell me your story'. They had to sit back and listen, then come back with what they found. It was brilliant. The staff really enjoyed it, they learned things they didn't know about their programme: mistakes they'd made, gaps, people who had been left out, wrong targeting etc. ...They felt they had got a lot out of it, and we had a much better picture of what was going on in the community, than if we had done a more traditional evaluation where we tell people where we come from and why we are there.'

Source: Vivien Walden, Oxfam GB, personal communication, March 2015

Using metaphors

This is best facilitated in a workshop, for example through the following steps:

1. Participants, individually or in groups, draw images (for example, of a tree or a river) to represent their experience of the programme being evaluated. The different parts of the tree or river represent different parts of the programme.
2. Participants tell a story about the drawing, without interruptions, to their colleagues.
3. Those listening to the story ask questions about it to deepen analysis and learning and to offer their reflections.
4. At the end of this process, the group lists the main achievements, challenges, and lessons learned from the storytelling about their combined experience of the programme.
5. This can be focused on the future by asking the following questions at the end:
 - a. How can we build on what has gone well?
 - b. How can we deal with the challenges?
 - c. How will we apply what we learned here?





Good practice example: Using metaphors and storytelling in a review

The Tearfund review of capacity building in disaster management used the metaphor of the tree as the basis for storytelling. In a monitoring and learning workshop, each Tearfund partner was asked to draw a tree.

- The roots represented values and principles underpinning the programme.
- The trunk represented partners (organisations and individuals).
- The branches represented activities.
- The fruit represented programme achievements.
- The leaves represented lessons learned.
- Broken branches on the ground represented internal challenges.
- Clouds in the sky represented external challenges.
- Buds represented activities planned but not yet implemented.
- The storytelling was followed by a learning review.

Despite some initial scepticism, participants said that they found it was a visual and creative participatory process that had: encouraged reflection and learning; led to the joint realisation of achievements, challenges, and lessons learned; and provided a snapshot of the whole programme. They also recognised the challenge of translating learning into action.

The Most Significant Change (MSC) technique

The MSC technique is a participatory approach that can be used with staff or local community members. When it is used with the affected population it can have an outcome or impact focus. It builds from the field level upwards, using stories that capture what happened, when, why it happened, and why it was important. This technique was used in the IFRC evaluation of the response to the 2007 Peruvian earthquake; participants were asked to identify the most significant changes that had occurred as a result of the intervention by the Peruvian Red Cross (Martinez, 2009).

The steps are as follows:

1. Stories are collected at the field level in response to two questions. During the last month [or other time period], in your opinion, what was the most significant change that took place in the lives of people participating in the project? Why was this the most significant change?
2. The most significant of these stories are selected by a panel of stakeholders or staff.

- Once the changes have been recorded, the stories are read aloud, often in a workshop setting, and the participants discuss and reflect upon the value of the reported changes.

Advice for using this technique is offered by Davies and Dart (2005).

Workshops

Workshops with agency staff are a common feature of EHA. Workshops are often held:

- At the start of fieldwork. This allows the evaluation team to introduce themselves, present the inception report, and get feedback from field staff. It is also an efficient way to arrange later key informant interviews.
- At the end of fieldwork. This allows the evaluation team to debrief on what they have found and test their findings and conclusions with the workshop participants.

The evaluation team can use both of these workshops to collect data, identify key issues, and fine-tune their findings, either from the inception phase or the fieldwork. As noted in Section 5: Framing your evaluation, workshops can be used to develop or validate the ToC, and to establish what typifies adequate, good, and excellent to build an evaluative rubric (see Section 10: Desk methods).

The Oxfam RTE report on its drought response in Kenya reflected ‘feedback from debriefing workshops with partners and staff’ (Murphy et al., 2011: 2). This is quite a common approach. Start-up workshops are less common than debriefing workshops, but have the potential to increase stakeholders’ ownership of the evaluation. They are also useful for organising interviews and field visit logistics.



Tip

In the start-up workshop, ask what questions the participants would like the evaluation to answer. It will not be possible to address all of these, but even this discussion can clarify what the evaluation is actually going to do, and increase buy-in.

Workshops can also be used for broader objectives. The impact assessment of cash transfers in West Sumatra held both a start-up workshop and a final learning workshop to present findings from the assessment. This learning workshop also gave stakeholders an opportunity to share best practice and develop some practical steps for advocating that donors support cash transfer approaches (Aspin, 2010). During the three-month RTE of the response to the Haiti Earthquake, local NGOs used a workshop with the evaluators to begin a discussion of how to improve collaboration with international actors (Grünewald et al., 2010: 18).

Key facts:

- Widely used in EHA, most commonly for debriefing at the end of the fieldwork.
- Usually held with a group of persons who are knowledgeable about the evaluated topic.
- Often used to develop or test the ToC or to validate the evaluation plan or conclusions.
- Generate data on the validity of what is proposed.
- Can be used to develop recommendations based on the findings and conclusions of the evaluation.
- Need a workshop venue, seating, flipcharts, projector, etc.
- Typically a meal will be provided.
- Workshop participants' organisations are usually purposively selected but the organisation will decide who attends. The main problem with workshops is that the most useful potential participants are often too busy to attend.
- This is a medium-cost method.
- Typically there will be one or two workshops but there may be more in some specific cases for a small two-week evaluation.

Workshops can be less useful if groups with very different interests or with major differences in status or authority attend the same event. This makes it difficult to reach a consensus on particular issues.



Tip

Consider having cascaded workshops at different levels. A cascaded series of workshops feeds the conclusions of one workshop into the subject matter for the next. For example, an initial one-day workshop for field staff or partner organisations could identify issues that need resolution at their level. A workshop for middle managers on the following day could consider these issues and identify some of the underlying problem that have led to them. They could then suggest a number of ways in which these issues could be addressed. Finally, a half-day workshop for the senior management team could review the outcomes of the first two workshops and decide what action to take in order to resolve the issues.

See [Section 13: Field methods](#) for an example of the use of workshops for developing recommendations.

Learning tools in workshops

Workshops can serve to capture knowledge from participants and can also serve to help bring existing stakeholder knowledge to the surface and give them a chance to articulate and analyse it. One established format for learning workshops is the after-action review, discussed below, but many other tools can be used in workshops to promote learning, including:

- **Dotmocracy:** where participants indicate their agreement with different statements by putting dots on them (www.dotmocracy.org).
- **Idea rating sheets:** where any participant can generate an idea and then other participants can agree or not (www.idearatingsheets.org). The Idea rating sheets overcome some of the problems with dotmocracy.
- **Mini-rubrics:** where participants can use different colour dots to indicate their views on the value or other aspects of parts of an intervention. They can also be used to quickly present evaluation results (Davison, 2014).
- **Hierarchical Card Sorting:** a participatory card-sorting option designed to provide insights into how people categorise and rank different phenomena. Similar to Participatory Rapid Appraisal (PRA) ranking but for use with literate groups.⁷
- **Rich Pictures:** exploring, acknowledging and defining a situation through diagrams in order to create a preliminary mental model.⁸
- **Diagramming:** asking stakeholders to represent relationships as a diagram or to develop problem trees or mapping.

There are many more tools that have been developed for PRA or for training that can be used to collect data with specific groups of participants.

After-action review workshops

After-action reviews (AARs) can be used as an alternative learning approach to evaluation and also as a method to answer evaluation questions.



Definition: After-action review (AAR)

An after-action review is a structured discussion of a humanitarian intervention that enables a team to consider and reflect on what happened, why it happened, and how to sustain strengths and improve on weaknesses.

An AAR is usually a facilitated process involving all those who have been identified as the focus for learning, conducted in a workshop setting. An open atmosphere that fosters trust among participants is essential for a successful AAR. The general principle is 'no attribution, no retribution'. A neutral and objective facilitator is essential to ensure that the discussion stays focused on issues, remains positive, and does not deteriorate into self-justification or blame.

The review should address the following questions:

- What was expected to happen?
- What actually happened?
- What went well, and why?
- What could have gone better, and why?

An AAR includes the following steps. For each step, individual participants could write their responses on sticky notes, which the facilitator clusters to reflect particular themes, or the group could respond to each question collectively, and the facilitator writes the responses on a flip chart.

- Participants, individually or in pairs, write down their understanding of the objective or intent of the action. (Because humanitarian action often takes place in chaotic environments and plans can quickly become outdated, this is a useful way to find out if the objective is clear and shared, and redefine it if necessary.)

- Participants then write down what actually happened – possibly working as a group to construct a timeline of key events and changes over time in the situation or the programme. Cards that can be placed on a board or wall to show the timeline are useful for this exercise.
- The group then addresses two questions: What went well? What could have gone better?
- This is followed by a discussion on what could be done differently the next time, at the end of which the facilitator summarises all the lessons that have emerged and asks participants to vote for what they regard as the three most important lessons.

The key lessons learned, and any actionable recommendations, are documented and circulated to all participants. A timeframe, for example six months to a year, may be agreed on for assessing progress towards implementing the recommendations.

Key facts:

- In EHA, AARs are typically used when an agency wants to learn lessons from its operations.
- Typically last no more than half a day or a day.
- Needs a facilitator and a separate note-taker.
- Must observe the principles of no attribution and no retribution – the organisation must have a culture that permits the identification of problems without later looking for scapegoats.
- Good for helping the evaluation team to identify key lessons from the operation.
- Needs a suitable venue and key staff must make the time to participate.
- Participants are purposively selected.
- Relatively inexpensive.
- Usually only one AAR is conducted per site, so a multi-site evaluation may have more than one AAR.

The USAID (2006) guide provides details of the process.



Good practice example: Interagency after-action review of the response to the 2004 Indian Ocean earthquake and tsunamis

This review by staff (Baker, 2005) from five NGOs drawn from four tsunami-affected countries lasted for two days and identified three main lessons:

- The need for early socio-economic analysis to assist in programming and programme monitoring, for joint rapid assessments.
- A central role for community consultation and participation.
- The importance of preparedness planning, notably the need to build local capacity for an emergency response.

14 / Engaging with the affected population in your evaluation

Humanitarian agencies are paying renewed attention to strengthening their accountability to affected populations. It is a key commitment in the UN's Transformative Agenda to improve the effectiveness of humanitarian action. There is growing interest in strengthening 'communicating with communities' (CwC)⁹. The 2015 State of the Humanitarian System reported:

Greater awareness at the field level of the importance of engaging with affected people...so that conflict- and disaster-affected populations are not seen "purely as recipients", and that interventions are designed to centre more on their needs and preferences...[but] progress in accountability to aid recipients has mainly been at the level of rhetoric rather than reality.

(ALNAP, 2015: 72; see also Ashdown, 2011; and Brown and Donini, 2014).

What does this mean for EHA? How can evaluations engage more effectively with the affected population as well as in the earlier implementation stages?

Engaging with the affected population has been a weak part of EHA despite various efforts to increase accountability to this critical group of stakeholders – these are after all the people in whose name humanitarian action is undertaken.¹⁰ If we are serious about listening to the affected population, it is essential to engage with them during an evaluation.

14.1 Engaging with the affected population

It is only by engaging with the affected population and hearing their perspectives and views that we can know whether humanitarian programmes and projects have in fact been relevant to their needs, and thus be able to address the critical OECD-DAC evaluation criteria of relevance. The crisis-affected people are usually the best judge of the effectiveness of humanitarian work. We cannot improve the quality of humanitarian action without listening to their views. This means building quality consultation into evaluations.

The Humanitarian Accountability Partnership (HAP), now CHS Alliance, has worked on the issue of engaging affected populations for many years, developing standards that underline that listening and responding to feedback from crisis-affected people must happen at all stages of the project cycle, from planning and implementing through to monitoring and evaluating humanitarian programmes (HAP, 2013). Much of their learning has now been incorporated into the Core Humanitarian Standard (2015).

In practice, agencies have different reasons for engaging with the affected population in their humanitarian programmes, although this may not always be explicit. Brown and Donini (2014: 20-21) identify three rationales for agencies to engage with the affected population:

1. **A value-based or normative rationale:** agencies believe it is the right thing to do, for example in order to fulfil a moral duty, or to respect the fundamental rights and dignity of affected groups.
2. **An instrumental rationale:** because it makes humanitarian programmes more effective, for example, by gathering information to inform programme decisions, and better meeting the needs of those affected by crisis.
3. **An emancipatory rationale:** because it addresses structural inequalities and the root causes of crises, for example giving voice and agency to marginalised groups, or more ambitiously, transforming power structures and dynamics.

Evaluation plays a vital role in assessing how effective agencies have been in achieving these aims of engagement. Where the rationale is implicit, an evaluation may have to make it explicit, for example from interviews during the inception phase. As agencies attempt to be more accountable to the affected population, for instance as part of the Transformative Agenda, evaluation is vital in assessing whether they have achieved this from the perspective of the affected people. Good practice example on pg 270 summarises the findings of research conducted in the Philippines to hear the perspective of affected communities after Typhoon Haiyan, as many humanitarian agencies invested time and resources in being more accountable to the affected people.



Good practice example: Listening to affected people about their experience of agency efforts to be accountable to them

Through research carried out between November 2014 and February 2015 as part of the Pamati Kita Common Services project, communities affected by Typhoon Haiyan were consulted about their perspectives and experiences of humanitarian agencies' efforts to be accountable to them. The research adopted a qualitative approach in the spirit of a listening project. The consultations revealed that affected people had a strong preference for face-to-face communication over more technological means of communication, such as the SMS hotlines, which had been favoured by many agencies. Overall, affected people described their relationship with international humanitarian agencies as quite distant, in contrast to the perspectives of the agencies, which believed they had been much more accessible.

Sources: Ong et al. (2015); Buchanan-Smith et al. (2015)

Evaluations that have engaged actively with the affected population are usually much richer than those that have not. See, for example, the joint evaluation carried out by CARE and Save the Children after the Haiti earthquake in 2010, in the [Good practice example](#) below. Only certain types of evaluation might have good reasons for not engaging with the affected population, for example in an evaluation that is focused entirely on institutional processes rather than on the effectiveness of a programme in alleviating suffering.



Good practice example: People First Impact Method

Nine months after the Haiti Earthquake, CARE International and Save the Children Fund commissioned a joint independent evaluation of their humanitarian response. While the evaluation used OECD-DAC evaluation criteria and cross-cutting themes to assess the aid efforts to date, it also provided a snapshot of how different groups that were representative of Haitian society perceived and experienced the areas of the global humanitarian response in which the two organisations had participated.





This was done using the People First Impact Method that put the experience of Haitian people at its centre, and worked from that experience to determine the effectiveness of wider agency efforts, including those of CARE and Save the Children. National staff from both organisations were trained over two days to build their communication, listening and facilitation skills, and then to conduct focus group discussions using thematic open-ended questions. This meant that the discussion could be non-prescriptive, extending beyond the work and projects of either agency, in order to gain knowledge about people's real-life experience and thus to help answer the two agencies' questions: 'Are We Doing the Right Things?' and 'Are We Doing Things Right?'

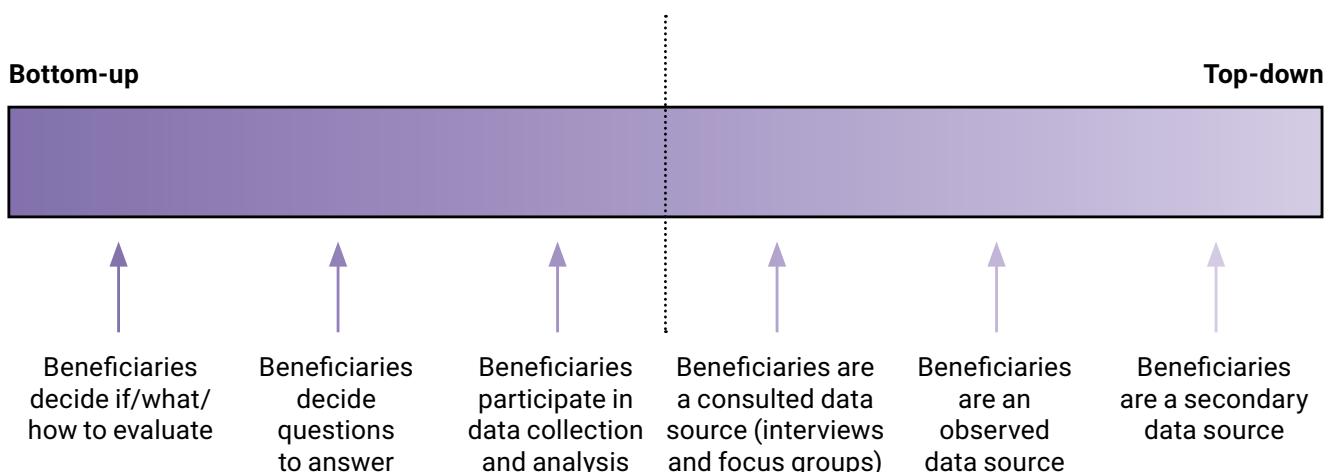
Source: O'Hagan et al. (2010)

Evaluations of humanitarian action that are truly bottom-up and participatory (see [Section 4: Types of evaluation](#)) – at the left-hand end of the spectrum in [Figure 14.1](#) – are very rare in the humanitarian sector. This would mean involving the affected population in all aspects of the evaluation, from the planning and design phases, to gathering and analysing data, identifying findings and formulating recommendations, and in [dissemination](#).¹¹ The following sections provide guidance on the more common ways of engaging with the affected population in EHA.

14.2 Planning to engage with the affected population

Only rarely is the affected population involved at the start of an evaluation, in determining the need for it and its potential scope, although there are precedents. One such example is the feasibility study carried out by OCHA for a joint impact evaluation (Beck, 2011). This requires allocating time for consultation before the evaluation begins.

Figure 14.1: Participatory Continuum in Evaluation – an example by the IFRC



Source: IFRC-PED, 2014

More often, engaging with the affected population means planning for consultation with them during the evaluation process. As demonstrated in IFRC's participatory continuum in evaluation, this tends to be in the 'top-down' half of the participatory spectrum, but is important nevertheless for the reasons stated above. Evaluation commissioners must plan for this kind of consultation from the outset, making sure both that the resources available and the timescale are adequate for substantial periods of fieldwork.

Consultation with the affected population has been weak in EHA usually because too little time and too few resources have been allocated to the fieldwork; this is often the part that gets cut when schedules have to be shortened and when budgets are tight. Yet, as an early ALNAP and Groupe URD document on participation of crisis-affected people in humanitarian action underlined (ALNAP, 2003), this is as much about a mindset as about a set of tools, in this case one that places engagement and consultation with the affected population centre-stage in the evaluation process, and resources it accordingly.

Evaluation managers should also make sure to recruit evaluators with the right set of skills and knowledge. They should be experienced in carrying out consultations with affected populations and some team members should have prior knowledge of the context and understand community structures and power dynamics. This may be particularly important in a conflict-related crisis. If there are sensitive protection issues to take into account, it is important that evaluation team members have the appropriate skills and experience to know how to deal with them.

**Tip**

In recruiting evaluators to carry out consultations with affected populations, in addition to interviewing them and reviewing their CVs, find out whether they have done similar work before and, if appropriate, take up their references.

**Tip**

In planning for consultation with the affected population in challenging contexts, for example because of inaccessibility and/or insecurity, build contingency time into the evaluation schedule in case travel is delayed or alternative arrangements have to be made.

14.3 Designing the evaluation to engage with the affected population

The first step in designing the evaluation is to identify which evaluation questions in the ToR are relevant for consultation with the affected population. Questions about the wider impact of the humanitarian action and about its appropriateness and relevance to the needs of the affected population must clearly be directed to the affected population. But questions about the cost-efficiency of interventions, or about an agency's internal management arrangements, may not be relevant for consultation with the affected population.

Second, it is important to review what already exists in terms of feedback from the affected population, especially in cases where there has been a high level of engagement throughout implementation, for example through feedback mechanisms such as hotlines, post-distribution monitoring and regular community consultation, and where that feedback has been well-documented. For example, after Typhoon Haiyan in the Philippines, some agencies such as WVI had rich computerised databases that recorded and analysed feedback received, having invested in a range of different accountability mechanisms.¹²

Third, it is important to consider which groups among the affected population are to be consulted, and thus how the population is to be disaggregated. Make sure to consider UNEG's guidance on incorporating human rights and gender equality in evaluations (2014). A good example is the evaluation of UNHCR's Age, Gender and Diversity Mainstreaming in Colombia (Mendoza and Thomas, 2009). Disaggregation usually aims to elicit different experiences of the crisis

and the response. Thus, for example, men and women, older and younger people, richer and poorer sectors may have experienced the crisis and response differently. Consider also the social ‘fault lines’, especially in a conflict-related crisis: for example, ethnicity may be a major fault line in the conflict with different ethnic groups experiencing the crisis in very different ways.

Fourth, there is a need to select methods and a sampling approach for engaging with and consulting the population. In practice, qualitative methods are usually used to engage with the affected population. Sometimes the affected population have been consulted using ‘mixed methods’, combining quantitative formal surveys with qualitative methods, which can be particularly revealing if done well, but require greater resources (Gayfer et al., 2014). This has been more common in well-funded joint evaluations, and increasingly in impact evaluations.

As described in [Section 12: Sampling](#), purposive sampling is often used because of the limitations of constrained access and difficulties of reaching certain geographic areas or certain groups, which make random sampling very difficult. In this case, local knowledge is critical in order to agree on the purposive sample.

14.4 Methods for engaging with the affected population

[Table 14.1](#) summarises five of the most common methods for consulting the affected population, and their pros and cons. Participatory rapid appraisal (PRA) is also very common and is addressed in the following sub-section.

[Focus groups](#) and [group interviews](#) are particularly effective for listening to the affected population in more open-ended discussions. These are usually guided by a checklist of issues to be covered, while allowing space to follow up on new or unexpected information that arises during the discussion. Focus groups are also an opportunity to consult different population groups separately, for example to talk with women and men separately, or with different age groups. A focus group should comprise six to eight people from a homogenous group within the population. In practice it is rarely possible to control the size and composition of the group, so it may become a much larger ‘group interview’.

It may be appropriate for courtesy and protocol reasons to start the consultation in a community meeting, to explain the evaluation. This is also an opportunity to ask for the community’s overall perspective on the programme or project, but is unlikely to elicit everyone’s voices. For instance, women, young people and marginalised groups may be reluctant to speak out.

Table 14.1: Common methods for consulting the affected population

Method	Pros	Cons
Focus groups & group interviews	Good for consulting a particular group. Good for open-ended discussion.	Controlling the size and composition of the group. Require skilled facilitation.
Community meetings	Use existing structures. Can involve large numbers.	May be dominated by certain individuals or groups. Difficult to disaggregate responses.
Individual & household interviews	Easier to ask sensitive issues. Can be more in-depth. Can be illustrative	More time-consuming. Cannot make generalisations from a small sample.
Semi-structured key informant interviews	May provide a good overview and insights that can be followed up in group or individual interviews.	Need to be aware of 'gate-keepers' and potential bias of key informants.
Formal surveys	Provide comparable data-sets and may be easier to generalise. Quantitative data convincing to decision-makers.	Time-consuming and relatively expensive. May not be feasible in conflict environments.

Focus group discussions can usefully be followed up with a number of individual or household interviews to pursue particular themes and/or follow up on sensitive information that it may not be possible to discuss in a group. These can be particularly useful as illustrative case studies, demonstrating the effect at household level of a particular aspect of the crisis and/or response, that has emerged as a common pattern, for example through the focus group discussions.

Key informant interviews are widely used in EHA as a way to consult the affected population. Key informants are usually individuals in the community who are in a good position to comment on the overall humanitarian response and its consequences, for instance who received humanitarian assistance and who did not and why. They may also be able to talk about some of the unintended effects of the response, positive or negative. Examples of key informants include health workers and teachers who have a lot of contact with community members and are in a good position to observe changes and trends and which groups are more vulnerable and marginalised, for instance. After the tsunami in Thailand, evaluators found that monks and policemen were informative key informants, although they had not received humanitarian aid.

Formal surveys may also be used as a way of consulting the affected population although, as described in Section 13: Field methods, this is according to pre-determined questions; they do not allow for open-ended discussion.

Participatory Rapid Appraisal



Definition: Participatory Rapid Appraisal

Participatory Rapid Appraisal (PRA) techniques (or Participatory Rural Appraisal) is a group of methods enabling local people to enhance, analyse, and share their knowledge and learning in ways that outside evaluators can readily understand.

Although PRA techniques were originally designed for a development context, they are well-suited to engaging with the affected population.

Key points:

- PRA methods generate mostly qualitative data.
- Used in EHA for consulting the affected population.
- Materials needed depend on the specific method.
- Sites are purposively selected.
- Participants may be self-selected or may be from a particular group.
- Can be done as part of a group interview.
- Number of PRA sessions can range from a few to dozens.
More PRA sessions are needed if it is the main technique in participatory evaluations.



Keep in mind

Inexpensive for the evaluators, but makes large demands on participants' time. Is this appropriate or reasonable based on the phase of the crisis and/or current context?

A strong point of most PRA techniques is that they are community based, enabling the evaluators to capture the knowledge of the community and not just of one or two members. Most are strongly visual and most are also accessible to non-literate groups. Participation in PRA exercises has a real cost for the community, however. As the Swiss Solidarity evaluation ten years after the 2004 Asian Tsunami noted during the initial qualitative research: Formal and organised gatherings, such as focus group discussions and participatory rural appraisal (PRA) tools, did not always yield reliable data, as the communities had been over-researched in the years since the tsunami. The use of these methods was therefore limited to situations where the team felt that they could add value to the research and a more open 'informal' semi structured interview method was applied (Ferf and Fabbri, 2014: 18-19).

Table 14.2: Common PRA techniques

Technique	Description	Potential use
Calendar	The group constructs a diagram that shows changes over time – for example, in agricultural work, gender-specific workloads, or disease.	A 24-hour calendar for women could show the amount of time spent on obtaining relief items (such as food aid) in the early stage of a crisis. A seasonal calendar could show periods of greatest food scarcity, which could then be compared in the evaluation with the timing of food distribution.
Timeline	The group constructs a timeline of events.	One timeline could record key events or moments of insecurity during a conflict-related crisis, and a second could record when humanitarian assistance was provided.
Proportional piling	The group is given a pile of 100 stones, beans, or seeds to represent a total in a given category (such as household income) and is asked to divide the pile up to illustrate the relative significance of various elements within that category (such as sources of income).	Knowing the relative significance of different sources of livelihood can indicate the relative importance of a specific relief intervention, such as a cash-for-work scheme or food aid. This technique can also be used to identify the poorest in the community.
Ranking	The group is asked to rank different items, either against each other or two dimensionally according to certain criteria. In pairwise ranking a matrix is drawn up with the same items on both axes and the group is asked to rank each item against the items on the other axis.	Ranking could be used to understand how well different types of humanitarian assistance (such as food aid, non-food items, seeds and tools) meet recipients' needs. If this is done two dimensionally, it could capture different needs in the household – for example, of elderly people, women, men, or children.
Transect walk	The evaluators walk through the village with a small group or with key informants and ask about whatever they observe – for example, who has access to grazing in a pasture or who lives in a home.	This can help evaluators understand the different effects of a natural disaster on different areas or groups in a community and to explore whether the humanitarian response was sensitive to these differences. It provides good opportunities for observation.
Social mapping	A visual method to show the relative location of households and the distribution of different types of people (such as male, female, adult, child, landed, landless, literate and non-literate) together with the social structure and institutions in an area.	Can help evaluators to understand how location and vulnerability are interlinked in a community. It can help evaluators focus on particular groups to answer coverage questions.
Venn diagram	Circles representing different categories overlap where the categories hold a given value in common.	These diagrams are usually used to show different institutions – for example, international, national, and local humanitarian organisations – how they relate to each other, and their relative significance in responding to a crisis.

More information on the use of PRA methods can be found in *Methods for Community Participation* (Kumar, 2002) and *Participatory Rapid Appraisal for Community Development* (Theis and Grady, 1991). The PRA approach stems from the work of Robert Chambers (1994). Online resources are available from the World Bank (Bank, 2007) and the Canadian International Development Agency (CIDA, 2005).



Tip

PRA methods look simple but require good facilitation skills. They should be used with a relatively homogenous group, such as recently displaced women or residents who remained in a settlement or town after it was attacked. It can be useful to carry out PRA exercises with men and women separately, or with leaders and non-leaders separately.

PRA facilitation needs to walk a fine line between community control and elite capture of the process (Adbullah et al., 2012).



Good practice example: PRA in a participatory evaluation in Uganda

The evaluation of the Farmer Field School (FFS) programme in Uganda (Foley, 2009) made extensive use of PRA tools. This was a food security and livelihood intervention similar to many recovery interventions in humanitarian settings. All 12 groups in the programme used:

- Wealth ranking to understand perceptions of relative wealth and how they might have been reflected in programme targeting.
- Seasonal calendar to see production cycles and input needs and judge how well programming aligned.
- Proportional piling to investigate household food, income, and expenditure and suggest what types of change might be expected from effective FFS programming.
- Pairwise ranking to explore the relative severity of coping strategies and to link these ideas with those about vulnerability and seasonality.

14.5 Particular challenges in engaging with the affected population in EHA

Challenges in engaging with the affected population, and possible solutions, are presented in Table 14.3 below.

Table 14.3: Common challenges of engaging with the affected population

Challenge	Potential solutions
Obtaining access to the affected population, especially in insecure conflict-related crises	Telephone surveys, meeting affected people in more secure environments, e.g. in market towns. See <u>Section 15: Constrained access</u> .
Cultural and language barriers	Ensure that the fieldwork team has appropriate language skills or has access to interpreters. Recruit local researchers/evaluators to be part of the team.
Developing robust methods to get beyond the anecdotal	Pay attention to the design of qualitative methods, for example using peer reviewers to advise on the robustness of the methods. Use <u>triangulation</u> .
Affected people have limited time for consultation	Be alert to time pressures and design your consultation methods accordingly, e.g. while people are waiting for relief distribution, in the evening when they return from work.
'Gatekeepers' attempt to influence which groups the evaluation team is able to meet	Be alert to political and power dynamics within communities and adapt the interviewing approach accordingly. See Tip on 'gatekeepers' on pg 280.
How to consult with a traumatised population	Ensure evaluators have appropriate skills and experience to engage with the affected population in a sensitive manner. See also <u>Table 14.4</u> .



Tip

When consulting the affected population, be alert to the influence of 'gatekeepers' in the community, for example community leaders who may wish to influence which groups or individuals the evaluation team consults. You may have to make a special effort to reach marginalised groups, for instance those living on the edge of the village or the back of the camp, but be sure that you are not putting them at risk by talking to them (see [Table 14.4](#)). Cross-check with other key informants about the different groups in the community.



Tip

When consulting the affected population consider using different entry points to those used by the agencies you are evaluating for needs assessments and ongoing monitoring, to avoid bias and to reach the 'unheard'. For example, carry out a transect walk through the community/camp, and randomly select households to visit and interview (see [Section 11: Evaluation designs for answering evaluation questions](#)).

The 'Listening Program' run by CDA describes how:

We of course followed protocol and met with the village chiefs, and community leaders and government officials, etc – but we always asked for permission to meet with others and sometimes had to insist on leaving time to meet with others (women, youth, marginalised groups). We asked local NGO staff and CBOs staff (community based organisations closer to the ground) who to speak with, and asked principals, teachers, nurses, local businesses, tuk tuk drivers for suggestions (See Anderson, Brown, Jean, 2012).

14.6 Ethical issues in engaging with affected populations

There are several ethical considerations to keep in mind in planning an evaluation (see [Section 2: Deciding to do an evaluation](#)). This sub-section expands on issues and considerations that are particularly important when engaging with the affected population.¹³ [Table 14.4](#) summarises some key challenges and possible ethical responses.

Table 14.4: Key challenges and possible responses when engaging with affected populations

Challenge	Ethical response
Are population groups at risk of abuse or targeted violence?	Do not interview people unless it is safe to do so.
Do politically powerful groups dominate the consultation process?	Be alert to power dynamics and seek ways to reach marginalised and less powerful people.
Are members of the affected population traumatised by their experience of the crisis?	Interview sensitively to avoid distressing interviewees, for example causing them to re-live a traumatic experience. For questions with a high risk of leading to re-traumatisation, ask yourself if this information is really necessary. Ethically, it is inappropriate to ask sensitive questions if the evaluators are not in a position to link traumatised people to a relevant service.
Has there been a breakdown of trust and relationships within the affected population as a result of the crisis?	<p>Ask trusted members of the community to introduce the evaluation and consultation process.</p> <p>Ensure complete transparency in explaining the purpose, constraints, intended use of the evaluation, and how data will be used and stored.</p>

In addition to considerations of ‘do no harm’, which are explained in Section 2: Deciding to do an evaluation, data protection, confidentiality and prior informed consent are important safeguards that should be put in place before interviewing and carrying out primary data collection in any evaluation, but most particularly in EHA where there are protection issues to be explored.

Confidentiality and its link to **data protection** should be part of all evaluators’ ethical code. This includes procedures to ensure the privacy of the respondents during the data-collection phase. The general principles underpinning data protection and confidentiality in evaluation is that everyone ‘owns’ their own life experiences and that attributable data are available to the evaluator only on a negotiated basis (Kushner, 2005: 74). Thus, the name of an interviewee from an affected population should not be included in an evaluation report unless they have explicitly granted permission.

It is also important to obtain **informed consent** before interviewing members of the affected population, especially if there are protection issues to be considered. This 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 about the evaluation before proceeding with interviews or other data-gathering exercises.
- Allow sufficient time for potential participants to reflect on and decide about taking part.
- Consider the skills of the interviewer (some evaluation teams include members with basic counselling skills if there are protection issues to be explored).
- Equip interviewers with information on services available to the interviewees (e.g. health and social services). Be aware that special considerations apply for all data-gathering exercises expected to cover issues relating to sexual violence.
- Reassure participants that they can withdraw from the interview at any time.
- Consider how best to ensure that interviewers comply with ethical procedures developed for the evaluation, and are consistent in approach (UNICEF 2015: 131-132, based on CP-MERG, 2012).

15 / Constrained access

15.1 What is constrained access?

Most EHA fieldwork methods require evaluators to observe and conduct interviews and exercises with the relevant stakeholders. But, as mentioned in [Section 1](#), evaluators often have little or no access to the people who have received humanitarian assistance, or even to the regions where they live. This is especially true of conflict settings. The risks range from humanitarian agencies (and those evaluating their work) being a direct target for attack (for example if they are associated with parties to the conflict), to inadvertently being caught up in an ongoing conflict, to endangering the affected population by seeking to obtain information from them.

Constrained access¹⁴ is clearly an issue for the people who are implementing humanitarian programmes as well as for evaluators. While a relatively high level of risk may be acceptable as the cost of providing life-saving assistance, it may be unacceptably high for conducting an evaluation. The decision to take risks should always be in proportion to the likely benefits.

What do we mean by constrained access? Examples of different scenarios and degrees of constrained access include:

- A context that is too insecure for any kind of formal evaluation, such as in parts of Somalia and Syria.
- While local evaluators may have greater access because they are more accepted and less visible than foreigners, for example in most of Afghanistan and during the conflict in Darfur, in other cases, such as the Balkans or North Moluccas, national staff may be at greater risk than expatriates.
- It may be possible for the full evaluation team to reach some groups in the affected population, for example IDPs in or near to towns, but not in rural areas, as was the case during the conflict in Darfur.
- International evaluators' access to a country or to certain regions may be subject to administrative constraints, such as the need for visas, travel permits, or no-objection certificates.

Constrained access is also a challenge for monitoring humanitarian interventions and may limit monitoring data in insecure environments.

This section focuses on how to decide whether to proceed with an evaluation when access is constrained. It looks at different approaches to evaluation in such situations and when an evaluation may have to be carried out remotely. Just as remote management is at the cutting edge of humanitarian assistance, remote evaluation is, in many ways, at the cutting edge of EHA. A number of agencies and evaluators have been experimenting with different approaches, some of the most promising of which are presented here.¹⁵ Section 15.5 looks at ways of ensuring credibility of the evaluation when access is constrained and some, if not all, of it has to be done remotely.

15.2 The importance of an evaliability assessment

When access is constrained, it is vital to conduct a feasibility or evaliability assessment before starting an EHA. Section 2: Deciding to do an evaluation explains the basic concept of an evaliability assessment as a means to have a clear basis on which to proceed, and to make the evaluation as useful as possible. If access is constrained, the assessment should focus on whether it is feasible to evaluate a given activity or programme, with credibility, and whether it is worth doing so.



Good practice example: MSF Evaluation Unit

Upon receiving an evaluation request, the MSF Evaluation Unit starts thinking through the ToR and possible design options. The Unit considers what adjustments may need to be made to the 'gold standard' of designs to suit the particular context of an operation. In situations of constrained access, such as cross-border operations in Syria, the main challenge is obtaining primary data on outcomes as well as direct feedback from beneficiaries. The evaluation design will therefore be more focused on process and activity than on results and it will have to rely more on secondary data and perceptions. These options and limitations are clearly shared with the commissioners of the evaluation, so that they may determine if an evaluation would meet their learning and accountability needs or if other options should be considered.

Source: Sabine Kampmueller, MSF Evaluation Unit, personal communication (September 2015)



Key questions that should be asked in order to answer if it is worth doing an evaluation include:

1. What are the main risks that the evaluation faces in this context? These may include:
 - a. operational risks such as the personal security of the evaluators and potential disruption to the programme if resources are diverted to the evaluation
 - b. financial risks if there are additional costs associated with working in an insecure environment, and
 - c. protection risks to which the affected population may be exposed if they participate in the evaluation
2. What are the implications of these risks for the evaluators' access to affected populations?
3. What secondary and other data are available if the evaluators cannot gain access to the affected population?
4. What other options are available to the evaluators to gain access to the affected population ([see Section 15.3 on remote methods of evaluation](#))?
5. How will this affect the credibility of the evaluation?
6. What are the possible alternatives to an evaluation, for example a reflective learning workshop with staff, peer learning among agencies, or more limited evaluative activity?

An evaluability assessment may be particularly important where agencies feel they are under pressure from donors to fulfil a contractual obligation to evaluate to assess systematically whether an evaluation is feasible.



Good practice examples: Evaluability assessments

In 2009 the IASC commissioned an evaluability assessment (Cosgrave, 2010) of a proposed evaluation of humanitarian interventions in Central and South Somalia. As well as addressing issues that usually form part of an evaluability assessment, such as clarifying the objectives and defining the scope of the evaluation, the key question was that of how to conduct an evaluation when even the simplest monitoring was very difficult. The main element of the evaluability assessment was a commentary, based on key informant interviews and a document review, on a draft ToR for the evaluation dating back to 2008. The evaluation was successfully conducted in 2011 (Polastro et al., 2011).





In 2014 Global Affairs Canada, part of the Canadian government, carried out a 'Scenario and Risk Analysis' before its planned South Sudan Country Program Evaluation (CPE). The core question was how the current conflict in South Sudan would affect its inception and data-collection portions. A subsidiary question concerned the necessary contingency plans. A systematic scenario analysis explored the current situation and anticipated a range of possible outcomes. A risk register was subsequently developed, differentiating between operational, financial, report quality and reputation risks. On the basis of this analysis it was decided that the evaluation should go ahead, and a number of measures put in place to mitigate some of the risks.

15.3 Ways to overcome constrained access

Agencies and evaluators have experimented with a number of creative ways to overcome constrained access to affected populations. Some of the most commonly used are presented in Table 15.1.

Where access is constrained, evaluators may have to rely to a great extent on secondary data, although this is often poor in such contexts.



Definition: Crowd-sourcing

Crowd-sourcing uses a large number of volunteers either to collect data or to analyse imagery data, usually through indirect means. This type of data is called crowd-sourced data.

Example of crowd-sourcing for monitoring: Satellite imagery analysed by crowd-sourced volunteers

In 2011 UNHCR enlisted the help of volunteers to tag three different types of informal shelter to provide it with an estimate of the IDP population in the Afgooye Corridor in Somalia. The volunteers processed 3,909 satellite images in just five days and added over 250,000 tags.

Source: <http://irevolution.net/2011/11/09/crowdsourcing-unhcr-somalia-latest-results>.

Table 15.1: Ways to overcome constrained access

Technique	How to use this method	Potential pitfalls
Use local researchers/evaluators to carry out interviews with the affected population (Norman, 2012: 30-35) (see <u>Good practice example on pg 289</u>)	<p>Plan for a training workshop at the outset (Sida, 2013) and an analysis workshop after the fieldwork has been completed.</p> <p>If appropriate, ask the local researchers/evaluators to record their interviews for verification purposes.</p> <p> Tip Make sure to allow enough time for these workshops, according to the skills and experience of the local researchers/evaluators.</p>	<p>Feasible only if local researchers/evaluators will not be put at risk.</p> <p>In long-running conflicts it may be difficult to find researchers/evaluators who are not perceived as being associated with one of the parties.</p>
Carry out surveys online, by phone, and/or SMS (see <u>Good practice example on pg 290</u> and <u>Section 13: Field methods</u>)	<p>These can be used for relatively short and straightforward surveys, for example to find out when people received assistance, what and how much. Phone surveys might also be used for field-based staff.</p> <p>Hotlines can be an opportunity for the affected population to raise questions of concern, especially if they were set up during programme implementation (Walden, 2013: 3).</p>	<p>SMS and online surveys are subject to self-selection bias and need to be interpreted with care.</p> <p>Phone surveys may also be associated with bias, e.g. accessible only to those who have mobile phones.</p> <p>In a highly politicised environment, local people may not trust and/or be reluctant to use a hotline.</p>





Table 15.1: Ways to overcome constrained access

Technique	How to use this method	Potential pitfalls
Interview members of the affected population in accessible areas <u>(see Good practice example on pg 290 and Section 13: Field methods)</u>	<p>Find out if members of the affected population travel regularly to more accessible areas, e.g. to more secure market towns, and arrange to carry out interviews and focus group discussions with them.</p> <p>Request members of the affected population to travel to areas accessible to the evaluation team, if it is safe for them to do so.</p>	<p>May introduce bias, e.g. if it is deemed safe for men but not for women to travel.</p> <p>Hard to triangulate findings if the evaluation team can meet only with particular groups.</p>
Remote observation (Norman, 2012: 45-48) <u>(see Section 13: Field methods)</u>	<p>Satellite imagery can be used to check infrastructure (e.g. built through cash-for-work programmes), or to review settlement patterns.</p> <p>Key informants and members of the affected population can be asked to take videos and photographs, using cameras with built-in Global Positioning Systems.</p> <p>If programmes are using remote monitoring, such as providing staff with tablets, smartphones, or cameras, these can be used to capture data.</p>	<p>While this may be suitable for observing the physical landscape and infrastructure, it may reveal little about how it is being used, e.g. who has access to particular infrastructure such as water points.</p>
Crowd-sourced data, e.g. Twitter, Facebook (Sida, 2014)	<p>Crowd-sourced data could be used to look at how widespread the use of particular facilities is through mobile phone tracking.</p> <p>Check if remote monitoring has been done through social media such as Facebook and Twitter, and if/ how the data could be used in the evaluation (see <u>Good practice example on pg 257</u>).</p>	<p>Respondents are self-selected, thus introducing bias, e.g. young people are more likely to use social media than older people.</p> <p>There may have been social media campaigns undertaken by parties to the conflict.</p>

If the programme has developed remote monitoring systems during implementation, it may be possible to build on them. For example if call centres or hotlines have been set up, these might be used in the evaluation for data collection; see Good practice example on pg 290.



Keep in mind

As with all evaluation, the data-gathering process should not place people at risk.



Good practice example: The evaluation of FAO's cooperation in Somalia, 2007 to 2012

As part of this evaluation, and recognising the constrained access that the team had to many geographical areas in Somalia, an in-depth study of FAO's cash-for-work (CFW) programme was commissioned as part of the evaluation. The fieldwork took place over a month.

An international researcher/evaluator (an anthropologist) was appointed to lead the study. Although for security reasons she had no access to the CFW sites, she worked with five carefully selected local researchers, and was based in Dollow in fairly close proximity to the CFW sites. Five days were allocated to training the local researchers, who used qualitative PRA methods for data collection.

A 'snowball technique' was applied (see [Section 12: Sampling](#)), whereby information was collected iteratively over the one-month period. The local researchers visited each village over three days, returning to Dollow for debriefing and to discuss their findings with the team leader. This was an opportunity to examine unexpected issues, often generating further questions to be asked during the next field visit. PRA techniques and tools produced material that could easily be analysed and was essential to the iterative research and training process. Triangulation was used in the form of photographic and audio recordings made by the local researchers. Using data from FAO's call centre set up for monitoring purposes during programme implementation, another Somali consultant was employed to conduct phone surveys which were also used to triangulate information collected in the field with CFW beneficiaries in other regions.

In addition to the CFW study, the evaluation team also requested some 'intermediaries' (who were able to travel), for example members of farmer and livestock professional associations, to come to Mogadishu to meet with the team in a safe space.

Sources: Buchanan-Smith et al. (2013); Tessitore (2013)



Tip

Consider whether certain evaluators might be less at risk, due for example to their nationality or ethnic identity. Keep in mind, however, that you have a duty of care towards all evaluators.



Good practice example of interviews by phone: Oxfam GB Evaluation of Somalia Drought Response

For its cash-distribution programme, Oxfam GB collected mobile phone numbers of participating beneficiaries (the RTE suggests that 10-15% of beneficiaries registered their phone number) and during the RTE in September 2011 a small number of people were called to provide feedback on the programme. This yielded some noteworthy results. For example, of 12 numbers called, five answered (one of whom actually returned their missed call). The purpose of these conversations was to assess beneficiaries' knowledge of their selection criteria, their impression of the process, understanding of the project, and whether they knew how to lodge any complaints. The feedback from these calls yielded several insights. For example, all the respondents said they understood the selection criteria, most indicated they knew how to contact an official from the organisation if necessary, although none was aware when, how and how much money would be received.

Source: Featherstone (2012: 13-14)



Good practice example: Bringing members of the affected population to more accessible areas

In evaluating an IDP programme in DRC, it was important for Groupe URD, the evaluation team, to speak to the affected population directly. Obtaining access to the settlements would, however, have involved passing through rebel-held territory. Although foreigners were at risk of being kidnapped, male IDPs could transit relatively safely through the area. Local partners had previously taken advantage of this to do distribution of assistance items. Therefore, Groupe URD developed precise criteria to help the IDP population select a range of representatives (e.g. farmers, religious or traditional chiefs). Four or five representatives per settlement were thus selected and were asked to travel to a secure village to be interviewed. IDPs also had the option of identifying someone who was already in the village whom they felt could accurately speak on their behalf.

To make the trip worthwhile, local partners offered the representatives supplies that they could take back to their communities.

The downside of this kind of approach is that the evaluators cannot control whom the community will choose. Only those who can pass through the insecure area can come (for example female representatives cannot travel in certain conditions).

Source: Bonaventure Sokpoh, Groupe URD, personal communication, 2015



Good practice example: Remote evaluation based on staff interviews, MSF in Pakistan

In 2009, MSF initiated an inter-sectional evaluation of its remotely managed projects worldwide. Due to security concerns in Pakistan, the evaluator had little access to the programme or the affected population. To help gain a more complete picture of the programme, the evaluator contacted current and former MSF staff. International staff were easily tracked down (and interviewed in meetings or by phone), but this was not so for national staff. The evaluator used a snowballing technique to identify staff that would be helpful to speak to and to get their contact details. This was a time-consuming process but it paid off as it increased the sample of respondents considerably. Prior to the field mission, the evaluators followed up on leads, but much of the work was done in-country from the capital city and from a project coordination office (in Peshawar city). It was important to travel to Peshawar, to ensure the highest possible proximity to the actual project site, to find relevant people for interviews and to understand the context better.

Source: Mzia Turashvili, MSF Evaluation Unit, personal communication, November 2015

15.4 Credibility of remote evaluation

If the evaluators have had limited access to the affected population, this raises the question of how to ensure that the findings will be credible.

Consider the following:

1. Be explicit in the evaluation report about the limitations and constraints faced by the evaluation team, and how these have affected the evaluation findings. Be clear what population groups or geographical areas the findings relate to. Be careful in your analysis not to generalise to groups and areas the team has not been able to visit.
2. Triangulation is always important in EHA. Where access is constrained, triangulation becomes even more critical. For example, if you are dependent upon teams of local evaluators that have been working on their own, compare data gathered by different evaluators in the same location to identify any evaluator bias. Triangulate information collected from the affected population through remote surveys, such as online or phone surveys, with information from key informants in the area, and agency staff.

3. Where you feel you cannot be conclusive in your findings because of the constraints you have faced, consider turning uncertain findings into hypotheses to be tested by the agency after the evaluation. You may want to recommend further research and data collection on particular issues.



Tip

If you obtain conflicting information from different sources that you cannot reconcile, and you have not been able to visit the people or area concerned, be prepared not to use that information in your analysis, or make it explicit in your report that you have not been able to reconcile two (or more) different accounts.

15.5 Other options to remote evaluation

- **Remote monitoring:** consider using remote monitoring if remote evaluation is not thought to be feasible. Remote monitoring should be easier than evaluation if it focuses principally on inputs and outputs, rather than on outcomes and the wider impact. Third-party monitoring is also worth considering, for example if one NGO is asked to monitor the work of another one.
- **Peer learning:** organisations working in a relatively inaccessible area could be invited to come together to share learning and experience. This could also be used to encourage reflection and learning. In 2013 and 2014, ALNAP co-hosted with the DEC peer-learning workshops for agency staff working on evaluation, M&E, accountability and learning related to the response in Syria. This was a useful opportunity to seek advice and share learning (Sida, 2013 and 2014).

In some cases, you may consider advising that remote monitoring activities be improved prior to going forward with an evaluation. There are a significant number of resources and lessons on remote monitoring, including:

- From Somalia, see Polastro et al. (2011: 29).
- From the Syria response, see ACF's 2014 Learning Review (2014: 54-55)
- From a Humanitarian Innovation Fund (HIF)-funded project, Tearfund's report on monitoring and accountability practices for remotely managed projects implemented in volatile operating environments (Norman, 2012).

16 / Analysis

This section focuses on the methods used to transform the collected data into the findings that form the basis of the evaluation's conclusions and recommendations.

The section covers:

- The analysis needed to answer evaluative questions.
- The analysis needed to answer causal questions.
- Qualitative data analysis of primary data such as interview notes.
- Qualitative data analysis of secondary data, such as progress reports.
- Statistical data analysis of primary data, such as survey data.
- Numerical analysis of secondary data, such as from distribution data records.
- Moving from the findings to conclusions and recommendations.



Definition: Primary data

Primary data is data collected for the purpose of the evaluation.

Interviews and surveys conducted specifically for the evaluation are examples of primary data.



Definition: Secondary data

Secondary data is data collected for other purposes but is used by the evaluation.

Secondary data includes agency policy documents, progress reports, other evaluations and relevant scientific literature. See [Section 11: Evaluation designs for answering evaluation questions](#) for details on secondary documentation.

Both secondary and primary data provide evidence for evaluations.



Definition: Evidence

Evidence is the available body of facts or information that can support a particular proposition or belief.

See the ALNAP evidence study for more on the quality of evidence (Knox Clarke and Darcy, 2014).



Tip

Evidence tables facilitate writing the report by gathering all the evidence about a particular question or theme in one place. Record all evidence and emerging findings gathered through interviews and documents in an evidence table. This is a table where the evaluation team records pieces of information against the evaluation question or theme. This serves as a pre-coding of the data. This helps make evident which particular issues have strong evidence, and which do not. When working to draw findings, conclusions and recommendations, the evaluation team can see which questions or themes require a stronger evidence base. Evidence tables facilitate writing the report by gathering all the evidence about a particular question or theme in one place.

16.1 Big-n or small-n

The basic approaches for analysing big-n and small-n data are different. Big-n or quantitative data are usually analysed statistically, and small-n or qualitative data are usually analysed using coding (see below for a definition of coding). As noted in [Section 11](#), these two categories are not watertight and there are many overlaps. Some designs are linked with particular means of analysis.

One key difference between big-n and small-n methods is the means used to ensure accuracy and reliability. For big-n methods, accuracy and reliability are achieved through applying a particular method in a particular way. For small-n methods, accuracy and reliability are achieved partly through the method itself and partly through triangulation (introduced in [Section 13: Field methods](#)).

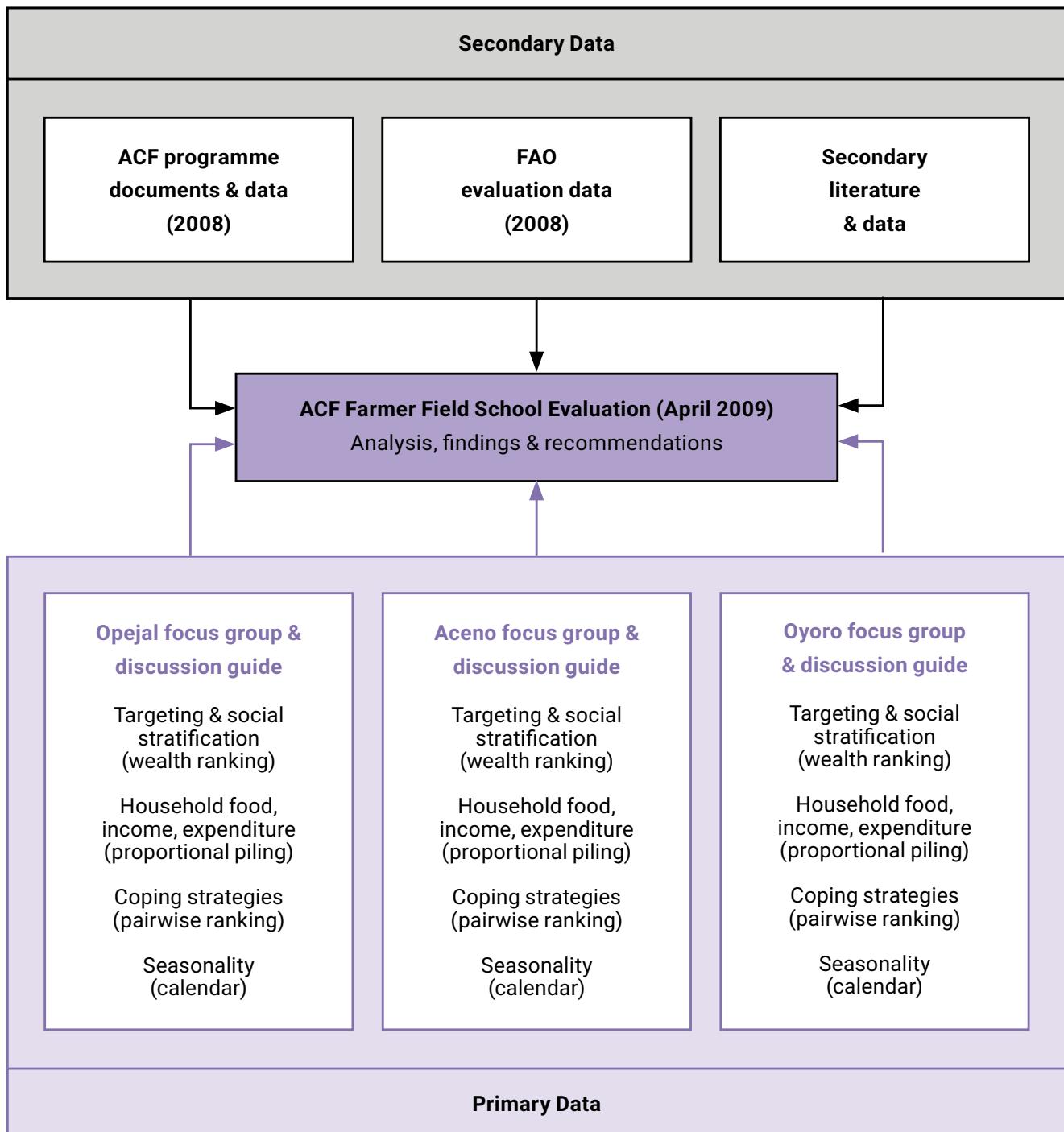
Implicit in triangulation is the idea that mixed and qualitative methods draw on each other to support a particular finding and conclusion.

The findings in any evaluation should be based on multiple sources (see [Section 10: Desk methods](#) and [Section 13: Field methods](#)).

The conclusions in turn may be based on multiple findings. The key principle is that evidence is triangulated as this helps to guarantee its quality.

The following example from a participatory evaluation in Uganda (Foley, 2009: 49) shows how different sources of evidence were intended to contribute to the evaluation.

Figure 16.1: Evaluation methodology diagram



16.2 Analysing evidence to answer normative questions

One way to answer normative questions, such as ‘to what extent did the shelter programme comply with the Sphere Standards’, is to search for triangulated evidence of breaches of or compliance with standards.

While this approach may be suitable for a very basic evaluation of a simple intervention, it can lead to the suggestion that any negative conclusions about compliance are based on anecdotes or bias rather than on rigorous evidence.

A more rigorous approach is to break down the relevant norm or standard applicable into separate elements. See Section 5: Framing your evaluation for examples of norms or standards that might be applicable. For example, to comply with the Sphere Standards for shelter you first have to comply with the six Sphere Core Standards (Sphere Project, 2011: 49):¹⁶

- People-centred humanitarian response
- Coordination and collaboration
- Assessment
- Design and response
- Performance, transparency and learning
- Aid worker performance

And then with the five detailed Sphere minimum standards for shelter (Sphere Project, 2011: 239):

- Strategic planning
- Settlement planning
- Covered living space
- Construction
- Environmental impact

Examining the extent to which an intervention met each of these standards makes it possible to answer the overall question. The Sphere Standards include indicators that can be used directly with some standards (if the indicators are judged to be appropriate in the particular context). Other standards can be answered only by making an evaluative judgement, and the techniques described in Section 16 can be used in that case.

16.3 Analysing evidence for evaluative questions

Evaluative questions (see [Section 6: Choosing evaluation questions](#)) ask evaluators to make value judgements. They can pose a problem for evaluators as such judgements can be criticised as being too subjective. This is an issue for all types of evaluation, regardless of design or methods. For example, a survey may establish that some critical factor has increased by 20% due to an intervention, but the evaluator has to make a judgement about whether this represents good or bad performance in the circumstances.

One way to approach evaluative questions is to break them down to separate descriptive, normative, and other elements from the purely evaluative element. Another approach is to take a rigorous and transparent approach to answering evaluative questions.



Definition: Evaluative reasoning

Evaluative reasoning is the analytical process by which evaluators answer evaluative questions.

Evaluative reasoning synthesises information on quality and value by combining:

- Evidence about performance on a particular dimension and interpreting it relative to definitions of ‘goodness’ to generate a rating of performance on that dimension.
- Performance ratings on several dimensions to provide an overall conclusion about how good performance is on the whole (Davidson, 2014: 1).

Examples of the use of evaluative reasoning through the medium of evaluative rubrics as advocated by Davidson (2014) are given below. Oakden states that rubrics ‘offer a process for making explicit the judgments in an evaluation and are used to judge the quality, the value, or the importance of the service provided’ (2013: 5).



Definition: Evaluative rubric

An evaluative rubric is a table that describes what the evidence should look like at different levels of performance, on some criterion of interest or for the intervention overall.

Evaluative rubrics are not scored numerically like the rubrics used for document analysis, but are scored with indicators of quality such as excellent or poor. The definition of ideal performance can be drawn from standards such as the Sphere Standards (Sphere Project, 2011) or from project and agency policy documents.

Evaluative rubrics have two elements:

- The descriptors of performance – poor, adequate, good, excellent and so on.
- Criteria that define what good looks like. In Davidson's model this is presented as a list of criteria under each of the descriptors of performance for a particular dimension.

Performance	Excellent	Very good	Good	Adequate	Poor
Descriptors of performance	Clear example of exemplary performance or best practice for food distribution; no weaknesses.	Very good or excellent performance in virtually all aspects. Strong overall but not exemplary. No weaknesses of any real consequence.	Reasonably good performance overall, might be a few slight weaknesses but nothing serious.	Fair performance, some serious but not fatal weaknesses.	Clear evidence of unsatisfactory functioning; serious weaknesses across the board on crucial aspects.
List of characteristics for this performance					

Source: Based on Davidson (2004: 137)



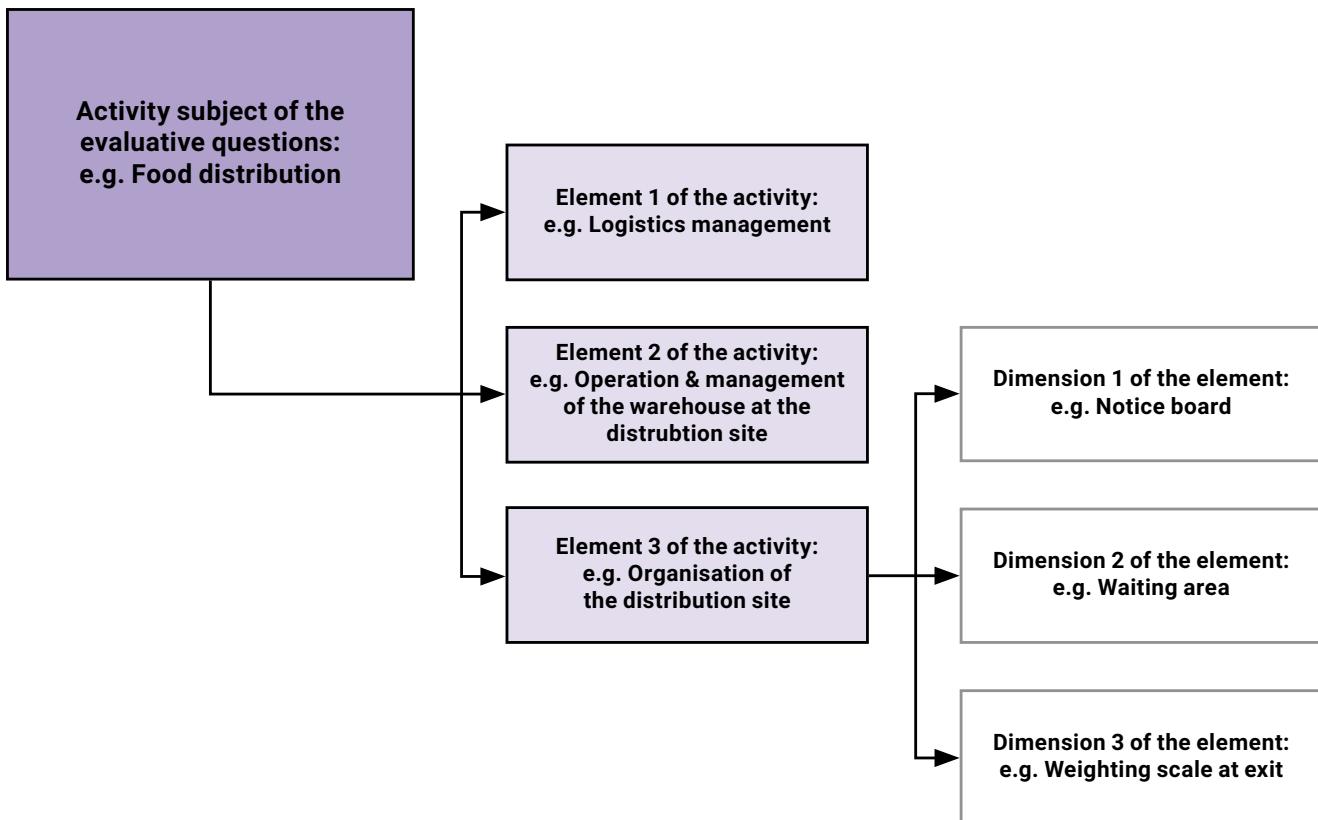
Tip

Develop the evaluation rubric before the main data collection Define what you think are the characteristics of different levels of performance before you begin your main data collection. Do not reinvent the wheel. Review conceptual models or standards in the sector to develop the criteria for your rubric (see Section 5: Framing your evaluation).

Davidson uses a list of characteristics for each descriptor of performance for each element of the activity as in Oakden (2013: 6-7). For this example, the list of characteristics of excellent performance in the organisation of a food distribution site (one element of a food distribution activity) might include the following dimensions:

- Illustrated notice board showing ration scales for each group in relevant languages
- Shaded waiting area with access control and access to water and toilets
- Sufficient large-capacity weighing scales so that recipients can check their rations – with test weight
- Clear signs directing people to the correct lanes
- Clean and dry queuing lanes
- Scoops in good condition and clearly marked
- Queue control measures to prevent crushing

Essentially, this is breaking down a more complex activity into elements and dimensions to allow a more transparent assessment of the quality of the activity (determined by the criteria expressed in the rubric). For instance:



The following rubric shows a few dimensions of the organisation of the food site element of food distribution:

Activity	Food distribution				
Element	Organisation of the distribution site				
Dimension	Excellent	Very good	Good	Adequate	Poor
Notice board	Illustrated notice board showing ration scales for each group in relevant languages	Notice board without graphics showing ration scales for each group with relevant languages	Notice board showing ration scales for each group but missing some key languages	Notice board showing overall ration scales, but missing group information or languages	No notice board
Waiting area	Shaded waiting area with access control and access to water and toilets	Shaded waiting area with access control and access to toilets	Shaded waiting area with access control but no toilets	Waiting area with access control	No formal waiting area
Weighing scales at exit	Sufficient large-capacity weighing scales so that recipients can check their rations – with test weight	Sufficient large-capacity weighing scales so that recipients can check their rations – but no test weight	Large-capacity weighing scales so that recipients can check their rations, but with queues	Weighing scales at exit but not large enough for full rations amounts	No weighing scales at exit

Criteria drawn from UNHCR (1997); Jaspers and Young (1995); Sphere Project (2011).

Evaluative rubrics can use quantitative as well as qualitative performance criteria, and in the following example of the rating for agreement with a statement such as ‘the distributors treated recipients respectfully’, the scores for survey answers are used to establish the extent to which recipients were treated in a dignified manner.

Category	Excellent	Very good	Good	Adequate	Poor
Dimension scores	Over 90% agree to a considerable or high degree	80-90% agree to a considerable or high degree	60-80% agree to a considerable or high degree and no more than 15% agree to a limited or very limited degree	40-60% agree to a considerable or high degree and no more than 15% agree to a limited or very limited degree	Less than 40% agree to a considerable or high degree

Source: Based on Oakden (2013: 10)

The different dimension scores can be combined into an overall score for the element as follows.

Overall score for the element	Excellent	Very good	Good	Adequate	Poor
Dimension scores	All dimensions score excellent	Almost all dimensions score very good or excellent; no poor or adequate scores	Most dimensions score good or better; no poor scores, more very good or excellent scores than adequate scores	Most dimensions score adequate or better, no more than 20% of scores are poor and fewer than better than adequate scores	More than 20% of dimensions score as poor

The same approach can be used for combining the scores for the elements into an overall evaluative judgment for the activity.



Tip

Develop, or encourage your team to develop, evaluative rubrics in a workshop with stakeholders.

Holding a workshop with stakeholders to define a rubric for measuring timeliness reduces the likelihood of argument over this compared to if the evaluators or the evaluation manager arbitrarily seek to establish such a rubric.

Davidson (2014) provides an excellent overview of evaluative reasoning. No humanitarian examples of the application of evaluative rubrics were found while this guide was being written, but Oakden (2013) provides an example of their use. Evaluative rubrics provide a powerful tool to contest the perennial problem of evaluative judgements in EHA being dismissed as merely subjective.



Tip

Use evaluative rubrics or encourage your team to use them, where you consider that some stakeholders may contest the answers to evaluative questions as being subjective or based on anecdote or bias.

16.4 Analysing evidence for causal questions

Causal questions (see [Section 6: Choosing evaluation questions](#)) ask about the causal links between interventions and outcomes. Causality is a problem in EHA because most interventions feature many actors and changes are influenced by multiple factors, rather than just a single one. Thus EHA usually refers to contribution rather than attribution.

Causal inference is the cornerstone of [impact evaluation](#).



Definition: Causal inference

Causal inference is the establishing of a relationship between a cause and an effect.

It should be understood that, as Cook et al. point out, 'causal inference, even in experiments, is fundamentally qualitative' (2002: 6). Deciding what the relationship between variables means is a qualitative judgement, even if it is based on quantitative evidence. For example, deciding which factors should be checked as possible confounding factors is a qualitative judgement.

Davidson (2009) offers eight strategies for establishing causal inference:

1. Ask those who have observed or experienced the causation first hand. In the Cook–Scriven dialogue on causation in evaluations (Cook et al., 2010), Scriven makes clear the people can observe causation at first hand. Cook notes that they have to rule out alternative explanations. In EHA, if you observed market traders reducing their food prices as soon as a food distribution is announced, you would have good reasons for inferring that the announcement caused the reduction once you had talked to the traders to rule out alternative explanations.
2. Check if the content of the intervention matches the nature of the outcome. For example, if there was a reduction in severe malnutrition following the distribution of mattresses, it would be foolish to conclude that the distribution led to the reduction.
3. Look for distinctive effect patterns (the modus operandi method) while searching for and eliminating alternative explanations.
4. Check whether the timing of outcomes makes sense. If B invariably happens after A, then it is likely that A may lead to B.
5. Look at the relationship between 'dose' and 'response'. For example,

- do children who have had the same exposure to a traumatic event show fewer symptoms if they have more therapy sessions than do those who have fewer therapy sessions in the same time interval?
6. Use a comparison or control (experimental or quasi-experimental design).
 7. Use statistical models to eliminate other potential causative factors. This is the approach used in statistically based quasi-experimental studies which use techniques such as stratification, multivariate models, logistic and linear regression, and the analysis of covariance to exclude the influence of variables which are not of interest.
See Pourhosingholi et al. (2012) for a brief description of the techniques.
 8. Identify and check the causal mechanisms. You can follow the theory of change and check the evidence for each part of the causal chain.
See Funnell and Rogers (2011) for a description of theories of change and their use.

See Alexander and Bonino (2015) for a discussion of causation in EHA. Rogers (2014) provides a good overview of causation. Befani (2012) provides a thorough academic review as part of the study by Stern et al. (2012).

16.5 Analysing evidence to answer descriptive questions

Descriptive questions (see Section 6: Choosing evaluation questions) sometimes require the use of statistical methods (see Section 16.6) to describe the intervention by identifying averages and so on. The analysis of coded qualitative data may also be used to identify overarching themes and issues in the intervention.

Qualitative data analysis of primary data

Qualitative data analysis should be a rigorous and logical process that turns the evaluation data into findings. Formal qualitative data analysis usually depends on attaching categories to particular pieces of evidence and then drawing those pieces of evidence together. Assigning such categories is called coding.



Definition: Coding

Coding assigns categories to particular pieces of evidence.

The categories are described by codes, which may simply be headwords that describe the category or, in more complex coding schemes, a set of alphanumeric characters. In EHA coding is almost always limited to the use of headwords that remind the coder of the category.

Coding in EHA

At its very simplest, coding can consist of reading the field notes and attaching labels to different categories of evidence, or simply recording field notes in line with categories in the first place. Simple approaches can include:

- Reading through the notes and attaching labels (codes) to evidence about the related category.
- Cutting and pasting notes into a new structure organised by themes.
- Using coloured markers or stickers on physical documents.
- Adding text labels, comments, or highlights to electronic documents.



In Depth: Coding in EHA

More complex evaluations need more rigorous approaches, such as coding field notes. In EHA the codes usually take the form of keywords and complex codes are seldom used. A piece of evidence can be a single observation, a sentence or a phrase from a report or an interview, a survey response, a survey summary, or any other statement of fact or opinion. Typically, a semi-structured interview of 45 minutes can produce anything from 10 to 20 pieces of evidence. Saldaña (2012: 1-41) provides an excellent introduction to the topic.

Budget and time limitations mean that post-fieldwork coding is very unusual in EHA. One exception is the CARE Cambodia Disaster Preparedness and Mitigation evaluation where the whole team was engaged in ‘translating and coding qualitative data and writing the draft report’ (Ramage et al., 2006). The usual pattern in EHA appears to be that evaluation team members write report sections based on their fieldwork notes, or that the evaluation team leader does it on the basis of written briefings from the team members and her own notes.





Current best practice in EHA is to code qualitative data as it is collected. The initial codes are established by the evaluation questions and by the issues that emerge during the inception phase. A typical humanitarian evaluation may have 20 to 30 codes. Additional codes can be drawn from (after Miles et al., 2013):

- The conceptual framework (see [Section 5: Framing your evaluation](#)). For example, if we were using the sustainable livelihoods approach as our conceptual framework, we would have codes for vulnerability, the different assets, structures, processes, livelihood strategies and outcome.
- The problem areas. Common problem areas in humanitarian action are targeting, gender, and coordination.
- The hypothesis. If we were examining the hypothesis that cash aid led to anti-social behaviour then we would have codes for different types of anti-social behaviour. Evidence of a lack of such behaviour would also be included.
- Key issues. In a shelter project, key issues could include rent levels, more than one family occupying a shelter and so on.

The following code table shows the codes for the Inter-Agency RTE of the Swat Valley crisis (Cosgrave, 2009). The codes were derived from an evaluation question, common problem areas, key issues, issues raised in the ToR, or issues that emerged during the fieldwork (such as neutrality, access, bureaucracy, delays and the contextual factors). Several of the codes were derived from multiple sources, but only the main source is listed.

Code	Source of code	No. of pieces of evidence	No. of different sources
Change from 2009-2010	ToR issue	57	23
Connectedness	Eval question	64	37
Early recovery	ToR issue	41	26
Gender	Problem area	26	16



Code	Source of code	No. of pieces of evidence	No. of different sources
Government	ToR issue	53	24
Learning for future	ToR issue	61	24
Military	Eval question	35	16
Monitoring	ToR issue	23	16
Neutrality	Fieldwork	92	36
Reviews	Eval question	7	5
Achievements	Eval question	81	39
Funding	Key issue	91	36
Coordination	Key issue	144	39
Access	Fieldwork	32	21
Learning Changes	ToR issue	22	12
Security	Problem area	56	30
Assessments	Eval question	43	32
Camps-Host-Other	ToR issue	58	34
Consultation	Eval question	12	11
Gaps	Problem area	109	47
Targeting	Problem area	82	45
Bureaucracy	Fieldwork	38	21
Delays	Fieldwork	18	14
Context Issues	Fieldwork	57	29
Total		1,306	160 in total





Coding takes place through reviewing collected data and abstracting evidence:

1. Interviews and other notes: The evaluation team enter their interview notes into the evidence tool, coding them at the same time.
2. Paper documents: Key segments are entered into the evidence tool.
3. Electronic documents: Key segments are copied, cleaned, and pasted into the evidence tool. (Cleaning is needed to remove hard-carriage returns and formatting codes from the text.)
4. Survey reports: For online surveys the process can be automated if the survey questions are aligned with the codes. The result is manually reviewed and additional codes added as needed.

Coding is much easier if all of the data-collection tools, interview guides, survey forms and so on are aligned with the initial codes.



Tip

Use the initial categories to organise all the data collection, because if the categories later change, it is easier to adapt from organised data than from disorganised data.

Use a simple spreadsheet evidence tool to facilitate the coding process, which can be automatically collated to have a single summary of the evidence.

Code	Evidence	Source	Initials	2 nd Code	3 rd Code
The topic code for this piece of evidence	Details of the piece of evidence – typically a snippet of text – from one sentence to a paragraph. The average word count for a piece of evidence in the example given above was 22.	A code referring to the specific source, be it an interview, meeting, document or observation	Who wrote this up	A second code for the same text	A third code for the same text



Although the tool can be used with a single code for each piece of evidence additional codes can be added for the same piece of evidence if necessary. If more than three codes apply to the same piece of evidence, the text can be re-entered with the additional codes.

The sources of evidence can be varied. In the case of an RTE in the Philippines, over 40% of the evidence came from specific comments made in answers to open-ended questions in an online survey (Cosgrave and Ticao, 2014). This, combined with telephone interviews, played a strong part in the evaluation as many of the critical key informants had already left the Philippines.

Data-collection method	Pieces of Evidence	As %
Online survey	944	42.1%
Telephone interviews	523	23.4%
Semi-structured interview (two or more interviewees)	390	17.4%
Semi-structured interview (individual interviewee)	230	10.3%
General meeting	117	5.2%
Emailed comments	21	0.9%
Detailed discussion (>10 minutes on one or more topics)	11	0.5%
Observation	2	0.1%
Total unique pieces of evidence	2,238	100%

Drawing conclusions without CAQDAS software

The evaluation team uses the codes to sort all the data and collate it by code. The team leader or a team member assigned to particular codes then reviews all the data collected under a particular topic before drawing findings. When considering the coded data, there will inevitably be some conflicts.

Evidence conflicts can occur in quantitative data as well as qualitative data. An example of quantitative evidence conflict is where amounts distributed by month do not match monthly amounts dispatched from warehouses –

there can be complex reasons for this, including dispatch and distribution in different months, different definitions, local storage at distribution sites, how discrepancies at distribution sites, returns, or losses are dealt with, and so on.

In considering evidence conflicts:

- Give greater weight to the views of those most affected. For example, the views of beneficiaries regarding whether or not they were consulted have far greater weight than the views of agency staff on the same topic. The view of beneficiaries who have missed out because of gaps in coordination should be given more weight than the views of people who attended coordination meetings.
- The views of someone with significant experience in a sector might have greater weight than those of someone with little experience. This is not always the case, however – a person with a lot of experience may be locked into a particular approach or mindset.
- Consider giving greater weight to views that contradict the apparent self-interest of the interviewee. For example, pay close attention when beneficiaries say they do not want more of a particular type of assistance. Similarly, pay close attention to criticism of their own programme by agency staff particularly if they do not give other indications of disaffection.
- Give greater weight to those views that triangulate with other evidence.

This approach ensures that the findings of the evaluation report are strongly grounded in the evidence gathered. Where there are significant evidence conflicts, these should be reported, together with the reasoning used in weighing them.

CAQDAS software for post-fieldwork coding

Post-fieldwork coding can be a slow process and is rarely carried out in EHA. However, there are powerful software tools that can help, such as Computer Assisted Qualitative Data Analysis or CAQDAS. But:

- These are expensive
- Coding consumes a great deal of time
- The software packages are difficult to grasp quickly.

For this reason, CAQDAS packages are seldom used in EHA. LaPelle (2004) shows that general-purpose word-processing software can be used for basic CAQDAS tasks. Koenig (2004) provides a very good practical overview of CAQDAS software, including potential pitfalls.

Miles et al. (2013) note that 'Some CAQDAS programs include wonderful features to support analyses, but there is a steep learning curve for most programs'. Silver and Lewins (2014) provide a step-by-step guide to some of the more popular packages. Nevertheless, Baugh et al. (2010) describe the time need to learn a package as one of the biggest limitations for the use of CAQDAS.

Qualitative data analysis of secondary data

This topic has already been partially addressed in [Section 10: Desk methods](#). The main point here is to emphasise that the secondary data can also be entered into the evidence tool. The main evaluation fieldwork may give access to additional secondary data that will need to be analysed using the methods described for desk studies.

Portfolio analysis

Portfolio analysis is a powerful technique for examining strategy by looking at project data. At its simplest, portfolio analysis can just consist of descriptive analysis on such topics as what percentage of aid went to which partners or sectors and so on. It can also examine average grant types by amounts. This is the type of portfolio analysis undertaken by the evaluation of the Danish humanitarian strategy (Mowjee et al., 2015) and by the joint evaluation of peace-building support for Southern Sudan (Bennett et al., 2010).

On a slightly deeper level it can look at which types of project, sectors, or partners were most likely not to request a budget extension or to encounter problems.

Portfolio analysis can also involve using a rating tool or summary scorecard, as in the case of the Australian evaluation of NGO assistance after the October 2005 Pakistan earthquake (Crawford et al., 2006). It can also involve the rating of individual projects.

A deeper portfolio analysis can consist of establishing a number of rating tools and then rating each project in the portfolio against them. This was the approach taken in the CERF five-year evaluation (Channel Research, 2011) for the areas of gender, vulnerability, and other cross-cutting issues.

16.6

Statistical analysis of primary data

This is not a comprehensive presentation of statistical analysis. The subject is very technical and there are many texts on the topic, all of them many times the length of this Guide.

It should be noted that data collected in surveys is not just numerical but can also include categorical data. The four data categories are:

- Categorical
 - Nominal – such as gender (fe/male)
 - Ordinal – where the categories are ranked (social class, for example, or ‘very unhappy, unhappy, happy, and very happy’ on a survey question)
- Numerical
 - Interval data – numerical data for which there is no true zero (such as a food security score)
 - Ratio data – numerical data where there is a meaningful zero and where numbers can be multiplied in a meaningful way (such as the weight gain in a feeding programme)

The statistical procedures that can be used vary with the type of data.

Cleaning and data entry

Using electronic data collection greatly simplifies data cleaning and data entry. Data still need to be reviewed as software will not correct every error. Electronic data collection has largely removed what used to be the onerous task of data entry.

Missing data

The problem with missing data (from non-responders or partial responders) is that there is always a concern that those who did not respond are somehow different from those who did, and that this may bias the results. This is a genuine concern, particularly with internet surveys, but non-response is also an issue with face-to-face surveys. The reasons for non-response may include:

- Interviewee was not available (the survey may have a protocol for repeat attempts or alternative interviewees)
- Refusal to answer particular questions
- Inability to answer particular questions (don’t know)
- Enumerator skipped it.

Good questionnaire design and testing can help to reduce the last three of these categories. The first is a particular concern because a potential interviewee may be absent for reasons that bias the survey results. For example, if a survey found that 10% of the selected sample was not available when the enumerators called the reason might be that they were engaged in some essential activity outside the village. Excluding these could bias the survey as it may be excluding a specific and distinct livelihood group. Again, pre-testing can identify problems like this.

Statistics

The use of statistics use in EHA is of two types – descriptive and inferential.



Definition: Descriptive statistics

Descriptive statistics are used to summarise key aspects of a population.

The most common descriptive statistics are measures of central tendency (commonly called averages) and measures of dispersion or spread (see pg 313). Such statistics are often used to give a partial answer to descriptive evaluation questions. They are also useful for answering normative questions, especially when the norm or standard in question has a related numerical indicator. For example, the answer to a question such as 'To what extent did we meet the sphere standards for the quantity of water supplied?' should probably include data on the mean amount supplied daily per person, as well as the range before a discussion on whether this level of supply met the standard of a 'sufficient quantity of water for drinking, cooking, and domestic and personal hygiene'.



Definition: Inferential statistics

Inferential statistics are used either to make inferences about a population from a sample, or to make inferences about hypotheses.

Inferential statistics are used both to answer descriptive questions (such as: what percentage of the affected population used most of their cash grant for shelter costs?) and causal questions (such as: to what extent did the community-managed acute malnutrition programme reduce levels of severe and acute malnutrition?).



Tip

Always give the confidence level and interval for inferential statistics.

For example, if 50% of the sample used most of their cash grant for shelter, but that due to a small sample size the interval for the 95% confidence level was 20% to 80%, this is a very pertinent fact since it is 95% certain that between 20% and 80% of the population used most of their cash grant for shelter.

Central tendency

The most common measure of central tendency for numerical data is probably the mean (usually called the average) such as average family size or the average time in a transit camp. Central tendency is often used to answer descriptive questions, and it can describe the average number or volume of cash grants distributed and so on. Using averages like this can quickly give an idea of the scale of interventions.

The mean is one measure of central tendency. Two other possible measures of central tendency are the median, which is the value for which half the values are large and half are smaller, and the mode, which is the most common value. The median value can often be a more useful measure of central tendency.

New refugees in an existing refugee camp

Consider a refugee camp that has a population of 2,000 who have been there for 20 years. A new influx of 8,000 took shelter in the camp one month ago. The average length of time refugees are in the camp is $(2,000*20*12+8,000)/10,000 = 48.8$ months – over four years. However the median is only one month (if you lined up all the refugees by the date of arrival, numbers 4,999 and 5,000 would have been there for one month). The mode (the most common time in the camp) is also one month.

Dispersion

Measures of dispersion (or spread) show how the values are distributed around the mean or median and can be very important in answering descriptive questions. For example, it would be a poor description to say that the average cash grant for livelihood activities was \$300 without also mentioning that the

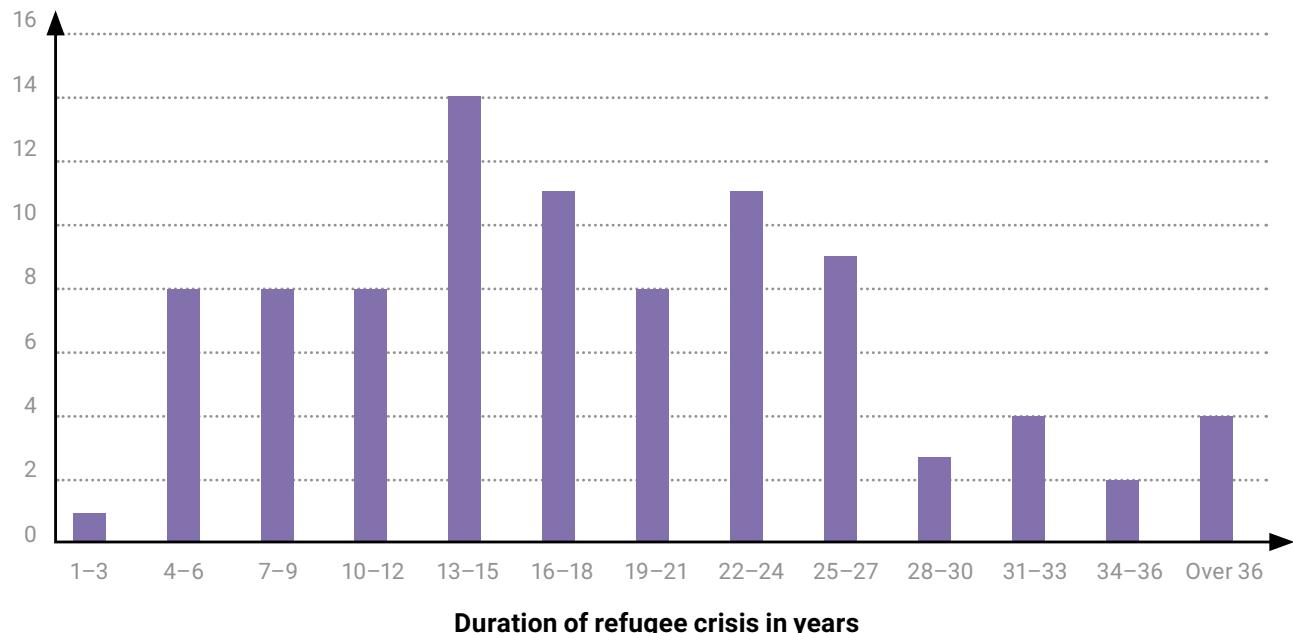
range was between \$30 and \$600, but that 90% of recipients received between \$200 and \$400. Options include:

- The range, the difference between the highest and lowest values (e.g. family size ranged from one to 19 members).
- Sub-ranges, the difference between the highest and lowest values applicable to a particular group (single-headed households received a grant of between \$150 and \$450).
- The inter-quartile range – the difference between the values that are 25% and 75% of the ranked values.
- The variance – the sum of the squares of the difference between each value and the mean.
- The standard deviation – the square root of the variance. This has the advantage of being in the same unit as the mean, so if the measurement is in years, the standard deviation would also be in years.

However, the variance and standard deviation are quite technical measures. Distribution charts are often more informative for primary intended users. The following chart presents the information more simply than would a statement of the mean duration of displacement and the standard deviation.

Figure 16.2: Frequency distribution of refugee crises in 91 countries of origin from 1978–2014 with over 5,000 refugees at peak (excluding the new crises of Syria and South Sudan)

Number of countries



Source: Author's own

Statistical hypothesis testing

The main advantage of big-n data collection is that it makes it possible to test hypotheses through the use of statistics: Hypotheses include:

- Hypotheses about the characteristics of a single group based on the characteristics of a randomly drawn sample (for example, estimating the population mean).
- Hypotheses about the difference or relationship between two groups (for example, between the assisted group and a comparison group that did not receive assistance, or between the assisted group before and after the intervention).

The usual approach is to test for the difference between the hypothesis and the null hypothesis. If, for example, this hypothesis is that those who received a shelter package one year ago had a higher household food security score than an equivalent group who did not, the null hypothesis would be that they do not have a higher household food security score.

The null hypothesis is similar to the counterfactual, which is the condition that would have prevailed if there had been no assistance. In this instance, the equivalent group that did not receive the shelter package represents the counterfactual. It is never possible to actually examine the counterfactual (because it is impossible simultaneously to test the results of providing and denying assistance at the same time to the same person, so the null-hypothesis is an approximation).

In hypothesis testing we can conclude either that the hypothesis is true or that it is false. If we conclude that it is true we could be right or wrong. If we are wrong in concluding the hypothesis is false, that is a false positive error and we conventionally use a value of 5% for this: that is, our conclusion that the hypothesis is true has a 5% chance of being wrong.

If we wrongly conclude that the hypothesis is false this is a false negative error. Conventionally, this is taken as 20%, but see [Section 12: Sampling](#) where it is suggested that a 10% chance of a false negative is more appropriate.

Statistical tests

The tests used depend on the type of data (categorical or numerical, as described earlier) and on the number of dependent and independent variables.

- Dependent variable: the outcome of interest – in this example household food security scores. This is a numerical variable.
- Independent variable: the input factor of interest – in this case the supply of shelter kits. This is a nominal variable (received kit/did not receive kit).

The choice of an appropriate test can be a complex technical matter.

The precise statistical tests will also depend on the evaluation design.

Different analyses are appropriate for multiple regressions, for interrupted times series, discontinuous regression and so on. Evaluation teams can undertake such designs only if they have the technical capacity to analyse them.

As a rule, we should test all apparent outcomes for statistical significance.

In this case, for instance, if we found that the assisted group had a household food security score that was 5% higher than the non-assisted group, we should test how probable it was that this result could have occurred by chance.

Consult a reliable text for details on how to conduct statistical tests. Many current university statistics texts seem to be based on the use of a software package such as SPSS.¹⁷ Such packages are quite expensive. Microsoft Excel has basic statistical capabilities, which can be improved with a free add-on such as Real Statistics (Zaointz, 2015), which adds a range of functions including some that are useful for processing big-n data.

16.7 Numerical analysis of secondary data

The numerical analysis of secondary data is similar to the analysis of primary data, but there is a greater focus on descriptive statistics, trends, and correlation. The review of protracted displacement (Crawford et al., 2015) was built in part around the analysis of secondary data on refugees collected by UNHCR, but also used secondary data on IDPs from the Internal Displacement Monitoring Centre, development data from UNDP, economic data from the World Bank, stability data from the Fragile States Index, as well as financial data from UNHCR, WFP, and UNICEF.



Tip

Always generate a scatterplot of any numerical data you are examining.

Scatterplots highlight any anomalies, show to what extent the data are correlated, and may lead to useful lines of enquiry. Any spreadsheet programme can produce a scatterplot. It is important to remember that correlation does not imply causation. It may be that the correlation is entirely by chance or that the apparently dependent and independent variables are in fact being influenced by a third or more other variables.

Figure 16.3: Correlation between overall humanitarian funding and ReliefWeb references in 2010 (each point is one country)



Source: Author's own

The review of protracted displacement (Crawford et al., 2015) used scatterplots to test relationships in the data before statistical testing. In this case the scatterplot was testing the relationship between humanitarian funding and ReliefWeb references to a country. The interest was in whether the distribution of ReliefWeb references could be used as a proxy for the distribution of humanitarian funding for recent years for which financial data was not yet available.

Common statistical tests used in numerical analysis

Any evaluation team should be able to conduct statistical tests for:

- Correlation. What is the probability that a correlation has happened by chance?
- Tests for independence, such as the chi-squared tests. Are two factors related in some way or independent?

Common spreadsheet software includes the functions necessary for such tests. Such basic statistical testing is essential to establish whether or not a correlation is likely to have arisen by chance. Tests of independence are needed to establish whether two factors are independent before wasting time investigating their relationship.

For example, in the evaluation of the Consortium of British Humanitarian Agencies (Cosgrave and Polastro, 2012), the team used a chi-squared test to check the relationship between grant committee membership and the percentage of grant applications that obtained funding. The impact of committee membership on grant success was not statistically significant at the 5% level.

16.8 From evidence to recommendations

Evaluations reports usually include evidence, findings, conclusions and recommendations. Evidence is usually presented to illustrate how the team reached their conclusion on a particular finding. The sum of evidence collected used in an evaluation includes the team's notes, all data collected, and the reference set they have collected. Only a fraction of this can be presented in the report.



Tip

Make the evidence presented in the report count. Include quotes from beneficiaries and use photographs, tables and charts to illustrate and summarise key points (making sure to gain consent and respect confidentiality as appropriate, see [Section 14](#) for more on this).



Definition: Finding

A finding is factual statement based on evidence.



Definition: Conclusion

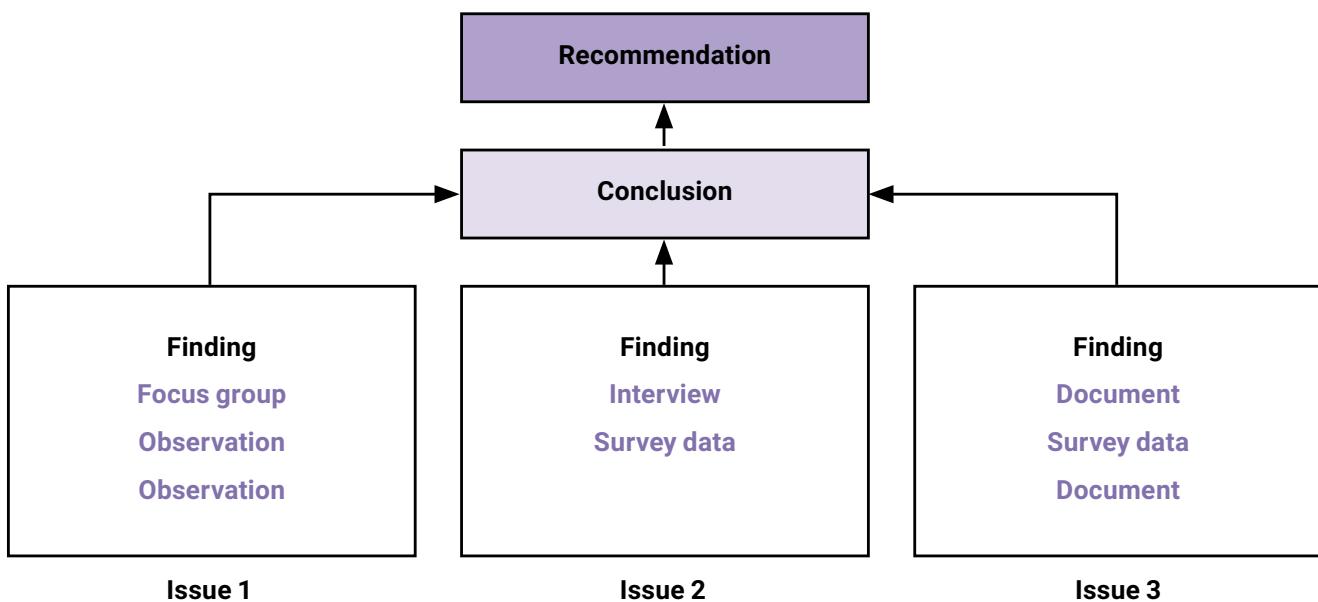
A conclusion is an inductive statement based on one or more findings.



Definition: Recommendation

A recommendation is a course of action the evaluators suggest as a way to address one or more conclusions.

There needs to be a clear logical flow, or process of analysis, from the findings through the conclusions to the recommendations. This should be presented as transparently as possible in the evaluation outputs. For example, findings should be clearly backed up by evidence, conclusions on findings, and recommendations on conclusions. This broadly follows the structure of a scientific report.



Jane Davidson (2007) argues for a different way to arrange an evaluation report. She suggests that the answers to the evaluation questions should be presented first, followed by the evidence and conclusions. She also advocates structuring the report around the evaluation questions and including everything related to each question in the relevant section, including any recommendations.

Conclusions can be presented either at the end of each section or in a final section. Presenting conclusions at the end of each section keeps them closer to the related findings and helps to make the logical flow more transparent.

It is also possible to have a final section that summarises the conclusions.

Recommendations

Recommendations can be:

- Omitted from the report. The logic behind only drawing conclusions is that the commissioning agency is best placed to take the conclusions forward into recommendations. This approach is rare in EHA as few if any organisations have the necessary structure to translate conclusions into recommendations. The RTE of Plan International's response to Tropical Cyclone Haiyan in the Philippines did not include recommendations in the final report (Cosgrave and Ticao, 2014).
- Generated by the evaluation team and presented in the evaluation report. This is the most common approach in EHA, and most EHA evaluation reports present recommendations.
- Developed through a consultative process with stakeholders. Again this approach is rare in EHA. The principles behind this approach are that it encourages ownership of the recommendations, and that the stakeholders are best placed to suggest remedies for issues identified in the conclusions. The joint evaluation of the Yogyakarta earthquake (Wilson et al., 2007) used this approach (see [Good practice example on pg 346](#)).
- Refined through a consultative process with stakeholders from draft recommendations initially developed by the evaluation team. This was the approach taken in the RTE of the 2010 Pakistan Floods (Polastro et al., 2011) (see [Good practice example on pg 346](#)). ACF policy is that recommendations should be general so that the stakeholders can develop them (ACF, 2011).

Useful recommendations

Recommendations are useful only if they are:

- **Specific** – it must be clear exactly what is being recommended.
- **Related to verifiable actions** – it should be possible to tell whether the recommendation has been implemented or not.
- **Directed** – the person or entity responsible for implementing the recommendation should be identified; responsibility may be further clarified in a management response to the report.
- **Practicable** – recommendations can involve new or unusual ways of doing things, but they should bear resources and other constraints in mind.
- **Time-bound** – a timetable for implementing the recommendations should be given wherever possible.
- **Consistent** – recommendations should not contradict or seem to contradict each other.
- **Prioritised** – it should be clear which recommendations are of primary concern and which are secondary.
- **Economical** – the recommended actions should clearly deliver benefits in proportion to their costs.

Recommendations should also be limited in number. The fewer recommendations a report contains, the easier it is for the commissioning agency to use it. An experienced evaluator may, however, notice dozens of performance issues and has an ethical duty to raise them. Various strategies can help resolve this conflict:

- Offer a general recommendation, and then provide details of how this could be implemented by different actors.
- Make minor recommendations verbally and put them in an annex.
- Rank recommendations by importance.
- Group recommendations by the focus of the recommendation.

Endnotes

11 / Evaluation designs for answering evaluation questions

1. The relevant chapter is available online at: us.sagepub.com/sites/default/files/upm-binaries/61527_Chapter_15.pdf.
2. Blinding is not always effective, but studies rarely test for this (Hrobjartsson et al., 2007; Fergusson et al., 2004). Sackett (2007) notes that in three studies of the effectiveness of blinding, it was successful only half of the time.

12 / Sampling

3. The standard deviation is the square root of the sum of the squares of the differences between the individual values of a variable and the mean value of that variable. The standard deviation is in the same units as the variable.

13 / Field methods

4. Drawn in part from the Michigan State University web page on interview probes (Kennedy, 2006).
5. This is a grid with the interview number on one axis and the number of adults in the household on the other. The intersection of the two gives a random number for selecting the interviewee (usually in ascending order of age).
6. The NRC review of Palestinian Education (Shah, 2014: 72) used a structured observation instrument for classroom observation. This instrument had 52 observation items to measure five underlying constructs.
7. See www.mande.co.uk/docs/hierarch.htm.
8. See betterevaluation.org/evaluation-options/richpictures.



Endnotes

14 / Engaging with the affected population in your evaluation

9. See the CDAC network.
10. Almost 75% of the EHAs assessed by ALNAP between 2001 and 2004 were judged unsatisfactory or poor in consulting with and encouraging participation by primary stakeholders, especially members of the affected population (Beck and Buchanan-Smith, 2008).
11. For guidance on conducting this kind of participatory evaluation see Better Evaluation's work betterevaluation.org/plan/approach/participatory_evaluation, and Alexander and Bonino (2014).
12. See Buchanan-Smith et al. (2015).
13. This sub-section is heavily based on the Evaluation of Protection companion Guide (Bonino, 2016). The content here is adapted to relate more broadly to EHA. For any specific ethical considerations relating to the evaluation of protection programming, please refer to the Companion Guide.

15 / Constrained access

14. For a full discussion of what is often referred to as 'contracting humanitarian space', see Collinson and Elhawary (2012).
15. See also Bush and Colleen (2015).

16 / Analysis

16. The Sphere Project will be replacing the six common standards with the nine Core Humanitarian Standards (CHS Alliance, 2015) in future versions of the Sphere Handbook (Sphere Project, 2015).
17. Andy Field (2006) is an example.

Notes

Notes

Communicating and reporting findings and results

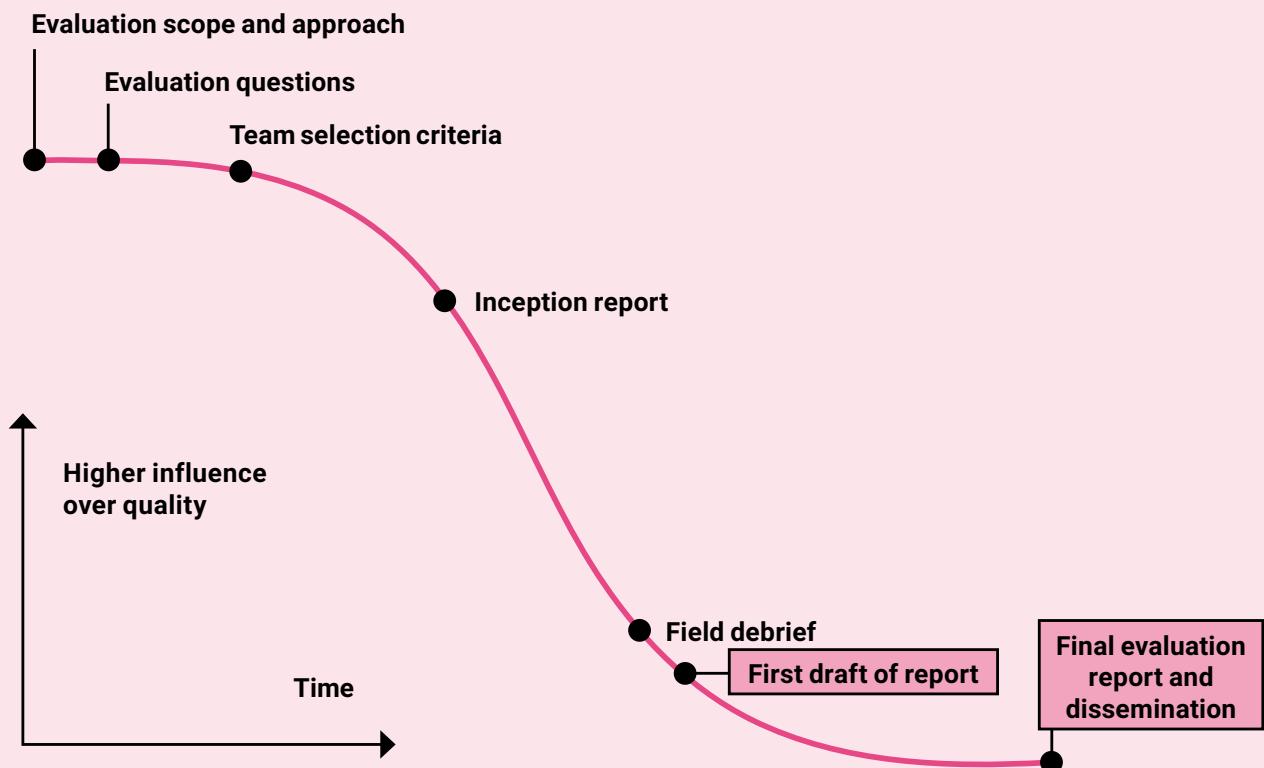


17 / Reporting and communicating evaluation findings with a utilisation focus

As highlighted in Section 3: Think early and often about evaluation utilisation, it is important in planning an evaluation to identify both what the primary intended users need to know and how this information can best be communicated to facilitate its use. How should the evaluation findings be presented to users? In addition to the evaluation report, what other products should be made available?

This section provides guidance for the evaluation team in writing up their findings in a report that is of high quality, engaging and accessible. It goes on to offer some ideas for how to present evaluation findings to maximise their use. Finally, it gives guidance for the evaluation manager and commissioning agency in the approval process to ensure a quality evaluation report, as well as in dissemination and communication activities after the report has been written.

Quality checkpoint: First draft of report & Final evaluation report and dissemination



This section covers two key quality control points. First, the draft report provides an opportunity for the evaluation manager to review the quality of analysis and influence how results are presented in the final report. Second, with the evaluation report in hand, the evaluation manager can now focus on dissemination of the findings and results, adjusting how these are presented to suit the needs of different stakeholders.

17.1 Key outputs

Outputs required from the evaluation team normally include: the inception report, debriefing workshops, advice provided by the evaluation team directly in the field, and the evaluation report. [Table 17.1](#) summarises the main evaluation outputs, their timing, and approval processes. The evaluation team may be requested to engage in dissemination activities as well.

Table 17.1: The main evaluation outputs

Output	Timing	Review and approval process
Inception report	At least two weeks before the fieldwork to allow time for comments and adjustments to the fieldwork methods and plans.	Circulation to selected stakeholders for comments, usually stakeholders with a direct interest, and/or primary intended users. Meeting between reference or steering group and evaluation team.
Advice provided directly in the field	As issues arise in the field and in response to observations, especially in RTEs.	None.
Debriefing workshops	Midway through an RTE for evaluation team members to test emerging findings. At the end of fieldwork in other types of EHA, usually to present and test preliminary findings.	Stakeholders' feedback on evaluators' preliminary findings and conclusions, validating or challenging.
Debriefing note	At the end of the fieldwork (can be as simple as the presentation at the debriefing workshop).	Final version should reflect the feedback received in the debriefing meeting as well as the preliminary findings.
Draft evaluation report	Three to four weeks after the end of the fieldwork, longer for more complex evaluations; may undergo several revisions.	Initial quality check by evaluation manager. Circulation to reference group or steering group, including stakeholders with a direct interest and intended users of the evaluation, for their comments. Meeting or workshop may be held between reference group or steering group and evaluation team to discuss key issues.
Final evaluation report	About two weeks after final comments received on penultimate draft.	To be approved by evaluation manager and/or wider management group.

17.2 Evaluation report

This is the main means of presenting the evaluation's findings, conclusions and recommendations. As stated in USAID's 'How-To Note on Preparing Evaluation Reports' (USAID, 2012: 1):

A key factor in using evaluation findings is having a well-written, succinct report that clearly and quickly communicates credible findings and conclusions, including easy-to understand graphics and consistent formatting.

Most agencies have their own requirements for how reports should be structured. If there is flexibility in how the report is structured, and the commissioning agency does not have a set format, the evaluation team should pay attention to what will be most useful to the principal intended users of the evaluation. Structuring the report around the evaluation criteria may be useful for the evaluation manager and for the evaluation department so they can make comparisons across evaluations, but it may be less useful for programme managers at whom many of the recommendations may be targeted.

For example, if you are evaluating a multi-sectoral humanitarian programme, you might want to structure at least part of the report by sector, for instance by WASH or by shelter.

An alternative is to structure the report according to the evaluation questions. Evaluators should consider sharing the report with the commissioning agency early in the process, for example with the inception report, to ensure it will meet the users' needs.



Examples of how to structure an evaluation report

For a report structured according to the OECD-DAC criteria, see IFRC's evaluation of its livelihoods programme after the Haiti earthquake in 2010 (Thizy, 2013).

For a report structured by broad themes explored during the evaluation, see the Joint Evaluation of the Global Logistics Cluster (Majewski et al., 2012).

The ALNAP Quality Proforma (ALNAP, 2005), developed as a way to assess the quality of humanitarian evaluation reports, provides a useful guide for writing an EHA report and looks for the following:

- Evidence that the evaluation assessed the intervention against appropriate international standards (e.g. international humanitarian and human rights law; the Red Cross/NGO Code of Conduct, Sphere).
- That it is informed by a contextual analysis of the affected area and population (including relevant historical, social, economic, political and cultural factors), and that this is drawn upon to support the analysis of the intervention; and that there is a clear analysis of the crisis, including key events (and a chronology where appropriate).
- That certain cross-cutting issues are addressed, including gender equality, advocacy, and consideration of vulnerable and marginalised groups.



Tip

Prepare your report outline before fieldwork begins to help you structure data collection and analysis to fit the requirements of the final report, thus reducing the work involved in producing it. It is particularly useful for collating and analysing qualitative data collected in interviews.

A short report is more likely to be read than a long one, but must provide sufficient evidence for the findings, conclusions and recommendations. The fewer the evaluation questions in the ToR, the easier it is to balance these requirements.¹



Tip

If the evaluation report is expected to adhere to a particular format, this should be stated in the ToR. This may include length as well as presentational details such as fonts, spacing, margin sizes and so on. Some agencies may be able to provide a report template.

The report structure

Table 17.2: Evaluation report structure

Section	Should contain
Title page and opening pages	<p>Key information, including:</p> <ul style="list-style-type: none"> • Evaluation intervention being evaluated • Timeframe of the evaluation and date of the report • Locations (country, region, etc.) of the evaluation – may include maps • Evaluators' names and/or organisational affiliations • Name of the commissioning organisation commissioning the evaluation • Table of contents which also lists tables, graphs, figures and annexes • List of acronyms
Executive summary	<p>A stand-alone section (usually 1-3 pages) that includes:</p> <ul style="list-style-type: none"> • Overview of the humanitarian action being evaluated • Evaluation objectives and intended audience • Evaluation methods • Most important findings and conclusions, following the sequence in which these are presented in the main report • Main recommendations
1. Introduction	<p>The scope and purpose of the evaluation, intended audience, team composition, and structure of the report.</p> <p>The overarching evaluation questions.</p> <p>Were there any changes to the evaluation questions proposed in the ToR?</p> <p>How was scoping done?</p>
2. Methods	<p>A description of the main methods used, their appropriateness and why they were chosen.</p> <p>If you have evaluated against a theory of change, you could include it here, or make reference to it and include in an annex.</p> <p>The nature and scope of involvement of the affected population.</p> <p>Key constraints to carrying out the evaluation (e.g., lack of time, constrained access to affected population, lack of baseline data), and their effect.</p> <p>Any biases in the evaluation process or evaluation team and how these were mitigated.</p> <p>International standards used as reference points in the evaluation, e.g. Sphere, LEGS, and any conceptual frameworks used, e.g. the malnutrition framework or the livelihoods frameworks referred to in Section 5: Framing your evaluation.</p>
3. Context	Contextual analysis of the crisis to which the intervention is responding, e.g. affected area and population, key events.



Section	Should contain
4. Main sections	<p>Organised by evaluation criteria, by evaluation questions or other framework appropriate to the evaluation and its intended users, these chapters present the evidence and findings.</p> <div data-bbox="457 377 1383 552" style="background-color: #e0e0e0; padding: 10px;">  Tip Presenting the conclusions and recommendations at the end of each section emphasises (and helps to ensure) they are grounded in the relevant evidence, findings, and conclusions. </div>
5. Conclusions	<p>Flow logically from and reflects the report's central findings.</p> <p>Provide a clear and defensible basis for value judgements.</p> <p>Provide insights pertinent to the intervention that has been evaluated and to the purpose of the evaluation.</p>
6. Recommendations <small>(see more on writing recommendations in Section 16)</small>	<p>Should be:</p> <ul style="list-style-type: none"> (a) Clear, relevant and implementable, reflecting any constraints to follow-up (b) Follow on from the main conclusions and reflect consultation with key stakeholders (c) Presented in priority order, with a timeframe for implementation, suggesting where responsibility for follow-up should lie (d) Limited in number.
Annexes	<p>Annexes usually include:</p> <ul style="list-style-type: none"> • TOR • List of persons interviewed and sites visited • List of documents consulted and secondary data used • More details on the methods, such as data-collection instruments • Evaluators' biographical data and/or justification of team composition • Evaluation matrix • Chronology of the issue or action being evaluated – this is essential in any evaluations in which timeliness is a criterion <p>Other annexes could address topics on which a detailed discussion would be out of place in the main report, or present results from specific methods (such as a summary of the responses to an online survey).</p> <div data-bbox="457 1664 1383 1799" style="background-color: #e0e0e0; padding: 10px;">  Tip Use annexes for supporting elements and for detail that would clutter the main report. </div> <div data-bbox="457 1837 1383 2016" style="background-color: #e0e0e0; padding: 10px;">  Tip If the report is circulated mainly in electronic form, consider presenting annexes as a separate document to make the report shorter and less intimidating when first opened. </div>

Source: Based on UNEG (2010) and ALNAP (2005)

The importance of the executive summary

Organisations may have standards on the length of the executive summary, but it is generally suggested that it be between two and five pages long.

Keep in mind that many stakeholders may only have time (or make time) to read the executive summary, so not only does it have to be a good representation of your evaluation but it must also stand on its own. It may be worth including a summary of the main recommendations, for example in a table, categorised in 'crucial/important/desirable' at the end of the executive summary.

The executive summary should be omitted in early drafts of the report to avoid getting comments on the summary rather than the full report.



Tip

Allow time to write and review the executive summary. It will be the most widely-read part of the report, so don't leave it until the last minute to write, when the team is running out of energy!

Making your report more engaging

There are a number of ways to make the report more accessible, readable and engaging. Some examples are listed below. Some are very easy to implement while others require more planning. They may need to be considered and agreed upon prior to the fieldwork.

- **Sub-titles:** Use clear sub-titles that convey the main message of the section or sub-section.
- **Bold:** Bold (or italics) can help to highlight key words in important (but long) sections of the report, for example, a key word or phrase in each paragraph. It is important to stay consistent in what type of words or phrases are highlighted (e.g. the theme of the finding) (DFID, 2005: 43).
- **Pull-out quotes:** Use pull-out quotes to highlight any particularly strong or important statements.
- **Direct quotes:** Direct quotes that relate to key findings can help bring the report to life. Although it is often vital to conceal a person's name, you can assign quotes using titles such as 'key informant from partner organisation', 'local school teacher', etc. (Sutter et al., 2011). See more on obtaining consent and respecting confidentiality in [Section 14](#).

- **Boxes:** Elements such as definitions, mini case studies, or good practice examples could be put into text boxes. These help to break up the report and make key messages or stories stand out.
- **Photographs:** These need to be carefully chosen and appropriate permission obtained (copyright or permission of identifiable individuals who feature in the photograph). See discussion on informed consent and confidentiality in Section 14: Engaging with the affected population in your evaluation. Does the commissioning agency have guidance on what images may or may not be used?
- **Stories:** Small stories can be used to personify a finding and help make an evaluation report less dry. They should be used sparingly, to support or exemplify a point, since they also tend to oversimplify issues. It is important to obtain the permission of any individual profiled in the story, particularly if it may be used outside the report or if the report is published.
- **Tables, charts and figures:** Tables, charts, graphs and other figures can communicate key findings in a concise way. Tufte (1990; 1997; 2001) provides good advice on the presentation of visual information, see also examples in Section 10.
- **Data visualisation:** The range of possibilities for these is constantly evolving, from stagnant infographics to web-based visuals that the reader can explore. For clear guidance and tips on data visualisation, see Evergreen Data (stephanieevergreen.com/blog).



In surveys, 83% of UNICEF country staff and 84% of country partners state cluster outcomes justify their investments

UNICEF (2013) Evaluation of UNICEF's Cluster Lead Agency Role in Humanitarian Action (CLARE), Evaluation Brief

Different reports for different audiences

Evaluation reports often have a range of intended audiences. The evaluation manager and the programme staff who implemented the intervention that was evaluated usually want to see detailed analysis, and especially the evidence on which findings and conclusions are based. They are likely to want a detailed report that spells this out. But other users, for example senior managers, are unlikely to read a lengthy report and need to have the key findings and their implications summarised in a short document. Thus it may be worth producing different versions of the final report – a detailed one that includes everything, and a short stand-alone summary. Danida publishes short briefing papers with every evaluation. See discussion on communication channels below.



Good practice example: Circulate a stand-alone briefing

A 16-page stand-alone summary was prepared and disseminated for the synthesis joint evaluation of support to IDPs, commissioned by Sida, the Danish and Dutch Ministries of Foreign Affairs, and ECHO in 2005.² This made the key findings of the full 140-page report much more accessible.

See Borton et al. (2005)

Circulating and commenting on the draft report

Draft evaluation reports are usually circulated by the evaluation manager to key stakeholders, including those stakeholders with a direct interest and intended users of the evaluation, many or all of whom may have been interviewed in the course of the evaluation. Giving them an opportunity to respond to any errors of fact, understanding or analysis is an effective way to ensure quality. It is also an important part of promoting stakeholder ownership of the evaluation. Occasionally all interviewees will be given the opportunity to read and comment on the draft report, although this rarely includes the affected population.

Process tips

When sending out your request for feedback on the draft report:

- Allow sufficient time for this – at least one to three weeks. Recipients of the draft may need to circulate it within their agency and collate one set of comments.
- Explain your timeline restrictions to commenters. You may consider giving a strict date for comments upon which assumption that silence implies consent will be applied.
- It is helpful to let stakeholders know the ‘maturity’ of the report. The level or depth of comments that are appropriate on a first draft is very different from those a final draft.
- Clarify the type of feedback requested and how it should be shared (see Good practice example on pg 338).
- If there only a few people commenting on the report, you may want to circulate it in Microsoft Word, so they can use the comments and track changes functions, and each person giving feedback is identified. If there are many stakeholders commenting on the report, consider circulating it as a PDF that cannot be changed, accompanied by a template in Word for comments.
- Where possible the evaluation manager should collate all comments into one document to make it easier for the evaluation team to address them. This also helps assure that all comments on a particular section are considered at the same time and so makes the revision process more efficient.

It is good practice for the evaluation team to indicate how they have responded to reviewers’ comments by providing brief notes against the collated comments, such as:

- Corrected
- Nuanced
- Detail added
- Deleted
- Not accepted – with a detailed reason.



Tip

Number paragraphs in the draft report to make it easier to make comments on specific passages. Do not number lines, however, as this makes the review process more intimidating and you may also end up with more comments than the evaluation team can handle!



Good practice example: Feedback process in the Canadian Red Cross

The Canadian Red Cross shares evaluation reports internally for feedback and to verify facts before sharing them more widely with all stakeholders. In the instance of the Evaluation of Health Emergency Response Unit deployment in the Philippines for Typhoon Yolanda, the draft evaluation report was shared for comments with all those interviewed or surveyed at the initial stage.

The feedback was quite useful and stakeholders welcomed the opportunity to vet the findings. This gave momentum to the feedback process and allowed the report to be released earlier, while appetite for it was still high.

To gather feedback the evaluation manager asked that comments be organised into three categories:

- Correction and errors: any factual edits
- Differences of opinion or additional information
- Recommendations: thoughts on what had been proposed

She offered stakeholders the opportunity to send in their comments via email or to phone to discuss their thoughts.

Approving and finalising the evaluation report

After the evaluation team has responded to all comments received, they should submit a final draft report for approval, usually by the evaluation manager or by a wider management group. There are a number of useful reference documents that can be used by the commission agency for the final quality check, such as the ALNAP Quality Proforma, specifically developed for EHA (ALNAP, 2005), and the UNEG Guidance Document that provides a 'Quality Checklist for Evaluation Reports' (UNEG, 2010).

In essence, the evaluation report should be assessed against the following:³

- Does it adequately answer the evaluation questions? If not, are the reasons for not doing so clearly articulated and are they acceptable?
- Does it provide evidence to support its findings?
- Do the conclusions flow logically from, and reflect, the report's central findings?
- Are the recommendations relevant to the purpose of the evaluation, and are they supported by evidence and conclusions? See [Section 16: Analysis](#) for more detail on getting from evidence to recommendations.
- Is the report coherent and free from internal contradictions?
- Is it accessible to its intended readership and users, e.g. in terms of its language, whether it is succinct, and clearly laid out?
- Does it follow any layout and formatting requirements?

17.3 Dissemination⁴

Plan for dissemination early

All too often, planning for dissemination of the evaluation findings and recommendations is left until after the evaluation report has been finalised. But if you are serious about utilisation, the entire evaluation – including dissemination – must be planned and budgeted with this in mind (O'Neil, 2012). This goes far beyond publishing the report.

As Bamberger et al. (2012: 166) explain many potentially useful evaluations have little impact because the findings are not communicated to potential users in a way that they find useful or comprehensible – or, even worse, because the findings never ever reached important sections of the user community.



Good practice example: Clarifying the dissemination list in the ToR

The IFRC 'Framework of evaluation' states that the ToR should include an initial dissemination list 'to ensure the evaluation report or summary reaches its intended audience. ...The dissemination of the evaluation report may take a variety of forms that are appropriate to the specific audience' (IFRC, 2011a: 15).

Here are some questions to help you think through dissemination:

- To which key groups do the evaluation findings and recommendations need to be communicated?
- Why does this information need to be communicated?
- What do these different audiences need to know? What would they like to know?
- What is the best means to communicate with each of these groups?
- Are there any special considerations or limitations particular to the users of the evaluation to be kept in mind (e.g. patchy internet connection, language, high staff turnover)?
- What is the best timing for dissemination (e.g. upcoming strategy revision, new planning cycle)?
- Who is responsible for the communication?

Evaluation reports in the public domain?

A few organisations, usually larger ones such as UN (for example, UNHCR) and donor agencies, have clear policies about putting all their evaluation reports in the public domain in the interests of public accountability. But many organisations do not, preferring to decide on a case-by-case basis whether they should be made publicly available. This may increase the risk of more critical evaluation reports being 'buried', but in some instances reports cannot be published for security reasons, for example if staff or partners could be put at risk from publication of the findings.

As good practice, it should be clear from the beginning of the evaluation process whether the report will be in the public domain. If necessary, the evaluators should seek clarification on this when they are first appointed.



Tip

Personal interactions in the field, in briefings, and in workshops are more likely than written reports to lead to learning. Be sure to plan for these, especially in evaluations with a learning focus.

Designing a dissemination strategy

As an evaluation manager, you should consider designing a dissemination strategy for each of your evaluations, however expansive or limited the communication and take-up of the evaluation findings should be.

A dissemination strategy means identifying the communication channels and products that best suit the needs of the various audiences and users of the evaluation. Some recommendations may have different target audiences. Sometimes a dissemination strategy has to be designed for each conclusion and recommendation to make sure that they reach their intended audience and are appropriately used. For instance, strategic and operational recommendations speak to quite different levels in an organisation. You might want to consider different communication methods for these two audiences, focused on what each needs to know.

Communicating and feeding back to stakeholders and interviewees in-country

Apart from debriefing workshops at the end of the fieldwork phase of an evaluation, there is often little investment in communicating the evaluation findings back to stakeholders and interviewees in-country. If the primary intended users of the evaluation are in-country, this should be given due attention. Consider budgeting for members of the evaluation team to return to the evaluation site(s) to report to key stakeholders, such as programme staff. Also consider ways to feed back the findings to members of the affected population who have been involved in the evaluation in the evaluation, see WVI's Good practice example below.



Good practice example: World Vision International feeds back to affected people

World Vision International uses community validation processes to feed back to communities it has consulted, whether during assessments, after community consultations, or in evaluations. These processes are used to inform communities about how the consultations influenced the agency's decisions. WVI staff may create posters, use flip charts and other materials to help present the findings to children and adults. More creative methods have included puppet shows for children.

17.4 Different communication channels

This section presents a number of different channels and products for sharing evaluation results or lessons, whether at the head office, in the relevant country, or more widely across an organisation. This list is far from exhaustive – be creative.

Meetings, personal communication and briefings

- Targeted meetings can be held between the evaluation team and intended users. If these are designed workshop-style, and customised to suit the needs of each group, they may encourage greater engagement by users.



Tip

Could summary presentations used for meetings and workshops be circulated more widely? Could these be translated? This can make it easier for the commissioning agency to circulate the evaluation findings locally.

- One-to-one briefings can be offered to key users by the evaluation team or the evaluation manager.
- Writing different versions of the report and short, accessible summaries or briefing documents. As Chris Lysy puts it in his blog/comic, FreshSpectrum:⁵ ‘Don’t just write one 200-page report. Write one 60-page report, two 30-page reports, five 10-page reports, ten 5-page reports, and twenty one pagers. Each with a different purpose, tailored for different audiences’ (August 2014).
 - ODI RAPID uses the 1:4:25 rule – 1-pager for policy-makers, 4-pager for information and 25-page full report.
 - UNICEF’s evaluation of its Cluster Lead Agency Role in Humanitarian Action is a good example of a 4-pager (UNICEF, 2014).
 - See the WFP Office of Evaluation’s briefs.⁶



Tip

If the full evaluation report is not going to be translated into relevant languages, consider translating at least the executive summary or brief into local languages for dissemination to local stakeholders.

- Consider adding a 'cover sheet' to the evaluation report, which categorises lessons learned, and provides a snapshot of why this report may be of interest to the reader (Oliver, 2007: 18).
- Short emails can communicate findings, conclusions, or recommendations to key users.



Tip

What tailor-made inputs based on the evaluation could be provided to managers? For instance, could speaking points be drafted for managers who are likely to have to report on the evaluation in relevant meetings?

Audio-visuals

- **Videos:** Findings and recommendations can be communicated using video footage from the evaluation or recorded interviews with the evaluation team. These can provide powerful feedback on key issues in the words of the affected population and are more likely to be remembered than long evaluation reports. See Good practice example on pg 344 and discussion on informed consent in Section 14.



Tip

PowerPoint makes it possible to add voice recordings to slides and convert the presentation into a video. If you cannot go to a meeting in person, it may be possible to send a video summary of the evaluation findings.

- **Podcasts** (audio recordings that can be listened to on MP3 players and computers) discussing key points from the evaluation.
- **Webinar** (online panel discussions or presentations), such as the webinar ALNAP co-hosted with DEC on using Contribution to Change in the Philippines (www.alnap.org/webinar/19) or the Interaction webinar series on impact evaluations (www.interaction.org/impact-evaluation-notes).



Good practice example: Using video to disseminate key messages or lessons

USAID's Office of Learning, Evaluation and Research commissioned a videographer to accompany an evaluation team in the field. The filmmaker followed a project evaluation team for over three weeks, to document most phases of the evaluation. One objective was to generate learning about the evaluation process, convey some of the challenges in the field (particularly in conflict-affected areas), and allow those managing commissioning, designing and undertaking evaluations to better understand the complexity and value of evaluation. The video is available at: www.usaidlearninglab.org/library/evaluating-growth-equity-mindanao-3-program

Two further examples are Groupe URD's video on the Haiti RTE (vimeo.com/15198053) and OCHA's video on the Inter-Agency RTE for the 2011 Pakistan Floods (ocha.smugmug.com/Film/OCHA-Films/i-rdN3hsX/A).

Source: Paul Barese in AEA 365 wwwaea365.org/blog/?p=8768 Linda Morra Imas, IDEAS Evaluation listserv exchange, April 2013

Internet-based options: social media and blogs

Social media is evolving rapidly, and new platforms and uses are continually emerging. Social media is increasingly used to publicise the release of evaluation reports or evaluation findings and recommendations. At present, most social media platforms are externally facing. But there is much room for creativity here.

One of the key rules for social media is learn its rules or particular features and adjust the content accordingly. For example, it may be appropriate to publicise the publication of a new evaluation report via Twitter, but this makes for quite a dull blog or Community of Practice post, for which lessons learned might be more appealing.

Here are some examples:

- **Blogs:** Details of the progress of the UNHCR Age and Gender Diversity Mainstreaming evaluation for Colombia (Mendoza and Thomas, 2009) were updated on a blog called 'It Begins with Me. It Begins with You. It Begins with Us'. (See itbeginswithme.wordpress.com).
- **Photo stories:** These could be presented in PowerPoint or PDF as well as via social media such as Flickr.com, a photo-sharing platform

that lends itself quite nicely to this. See the Asia Development Bank's example of an evaluation photo story: www.flickr.com/photos/71616117@N06/sets/72157630309259904/with/7452881380.

- **ALNAP Humanitarian Evaluation Community of Practice:** partnerplatform.org/alnap/humanitarian-evaluation.
- **LinkedIn Groups:** The European Evaluation Society conducts discussions on LinkedIn.
- **Twitter:** Groupe URD (@GroupeURD), for instance, is very active at tweeting about its evaluation work.
- **Reddit:** Comparable to an online bulletin board, Reddit lets users submit links that other users rate, thus creating a virtual ranking. Content is organised by topic, called subreddits. There is an evaluation-specific one here: www.reddit.com/r/Evaluation.

Evaluation databases

If people cannot find your report, they cannot use it. Many organisations are investing in improving their evaluation databases. For example, IFRC has revamped its online database (see www.ifrc.org/fr/publications/evaluations), and Tearfund has developed a Google Docs-based tool to share across the organisation details of planned and completed evaluations (Hallam and Bonino, 2013; Warner, 2014).

Established at the end of the 1990s, ALNAP's Evaluation Library⁷ offers the most complete collection of humanitarian evaluative materials to date – evaluation reports as well as evaluation methods and guidance material, and selected items on evaluation research. Make sure to upload your evaluation outputs with ALNAP.⁸



Tip

Include the annexes with the online version of the evaluation report. It is also possible to upload a version without annexes. This avoids the problem of someone being able to locate the report in five years' time, but being unable to find the annexes.

17.5

Facilitating take-up of an evaluation

A number of factors influence the utilisation of evaluation findings and results. See [Section 3](#) and Hallam Bonino (2013), for more on this. For instance, is there an internal demand for evaluation? Fostering and sustaining this is a higher-level and longer-term process to which a single evaluation makes only one contribution.

As discussed earlier, involving the primary stakeholders throughout the evaluation process is key to utilisation and take-up of the findings (see [Section 4](#)), as is a well-planned and well-funded dissemination phase (see [Section 17.4](#)).

Take-up can also be facilitated by:

- Involving some of the key users in formulating the recommendations. This can enhance stakeholders' ownership of the recommendations.
- Identifying champions within the organisation who are committed to action and to change.
- Building the findings of individual evaluations (or from a number of evaluations on the same topic) into training materials, for example as case studies.
- Clearly allocating responsibility for follow-up. In larger organisations, this is usually done through a formal management response matrix.



Good practice example: Consultative development or refinement of recommendations

The Yogyakarta evaluation made no draft recommendations. Instead, the conclusions were discussed at a multi-stakeholder event in Yogyakarta, which included beneficiaries, government representatives, local and international NGOs, and staff from the four agencies evaluated. The stakeholders reviewed and amended the conclusions and made some recommendations to INGOs regarding future responses. (Wilson et al., 2007: 2-3)

In the case of Pakistan, the evaluation team went with draft recommendations to be refined by stakeholders. After the draft report was prepared the team returned to Pakistan to hold three provincial and one national workshop with key stakeholders involved in the humanitarian response to the floods.



The team leader presented the findings, conclusions, and recommendations during the workshops. Stakeholders then jointly validated and prioritised recommendations and defined who would be responsible for implementing each recommendation and the timeline for implementation. The main changes in the formulations resulted from group discussions. The evaluation team considered that this process boosted ownership of the evaluation recommendations and fostered real-time learning among stakeholders engaged (Polastro et al., 2011: 4).

World Vision International also uses workshops to develop recommendations from some evaluations. They bring together a large group of stakeholders, sometimes up to 100, grouped into teams, into an interpretation workshop, which may last up to two days. The workshop participants might include senior leadership, sectoral staff, project partners, etc. During the process, different sections of the evaluation findings are provided to each group rather than the full report. These might be organised around the OECD-DAC criteria, for instance. Then each group has to work through their section of the report to answer a series of questions. This is a springboard for participants to understand how to interpret findings and then to use them appropriately during the rest of the workshop with the final aim of generating recommendations. In this way, the participants gain confidence in using the findings and they're far more likely to refer to the report in future for information and decision-making.

Source: (Personal communication: K. Duryee, World Vision International, 2013; Chamberlain, Jensen and Cascioli Sharp (2014); Vallet (2014)



Definition: Management response matrix

A record of management's response to each evaluation recommendation and the steps managers plan to take to address it, with a target date and responsible party for each step.

Figure 17.3: Sample management response matrix

	Further funding required?		Management response			Comment	Action to be taken	Timing	Responsible unit
	Yes	No	Accept	Partially accept	Reject				
Recommendation 1									
Recommendation 2									

Source: FAO

The value of the management response matrix is that it encourages discussion among the intended users about the recommendations and whether or how they will implement them. It can also serve as an accountability tool (UNEG, 2010). Some agencies, such as WFP and FAO, have a process of following up on the management response matrix, perhaps six months or a year later, to monitor whether the proposed action actually was taken.

The following are some points of good practice for management responses, distilled by UNEG (2010: 7-8). These are relevant to both large and small evaluations.

- **Clearly defined roles and responsibilities** should be communicated to all key evaluation stakeholders.
- **Establish an agreed deadline** by which management or other key stakeholders should provide their formal response to the evaluation.
- Management should **nominate a focal point** to coordinate the management response.
- The management response should clearly indicate whether **management accepts, partially accepts or rejects the recommendations**. If the latter, the reason(s) should be given. In the first two cases, actions to be taken should be mentioned in detail, indicating the timeframe and specific unit(s) responsible for implementing them. When more than one unit is mentioned, it should be clear which is responsible for which action(s).
- Management responses should be **disclosed in conjunction with the evaluation report**.

The limitation of the management response matrix is when a number of agencies have been involved in a joint evaluation. But it may still be possible for a single agency to identify the recommendations that are pertinent to it and to construct a corresponding management response matrix. UNEG (2010), in its good practice points, suggests forming an ad hoc group of management representatives of the different agencies/partners involved in a joint evaluation to elicit a coordinated management response.



Good practice example: Involving key stakeholders throughout the evaluation

For the ECB evaluation of the response to the Yogyakarta earthquake (Wilson et al., 2007), the evaluation team presented summary evaluation findings to the steering committee and to field staff so that these primary stakeholders could work with the team to draw up conclusions and recommendations. A final meeting was held between the evaluation team, members of the steering committee, government officials, and local people to review and amend the preliminary conclusions and to make further recommendations.

17.6 Evaluation syntheses, thematic reviews, and meta-analyses to facilitate take-up

As Hallam and Bonino (2013: 78-80) explain:

Most potential users of evaluation results want to know more than just what one single evaluation or study found. ‘They want to know the weight of the evidence’ (Weiss, 1998: 317). Dozens of studies, evaluations and reviews may cover the same issue or theme. Looking at these in their totality, through meta-analysis and evaluation synthesis, can yield far richer evidence and findings (World Bank, 2009: 70).



Definition: Evaluation synthesis

An evaluation synthesis follows a systematic procedure to organise and summarise the findings from a set of (possibly quite disparate) evaluations. This can be an opportunity to ask particular questions about the set of evaluations, and/or identify common trends in the evaluation findings.



Definition: Thematic review

A thematic review draws out the findings and learning from a series of evaluations of humanitarian action in a particular sub-sector, e.g. protection, shelter. These can be commissioned to answer/ address a particular set of overarching questions.



Definition: Meta-evaluation

Meta-evaluation is an overarching evaluation designed to aggregate findings from a series of evaluations, usually against a set of benchmarks. It can also be used to denote the 'evaluation of an evaluation', in other words to judge the quality of a set of evaluations.

There are several useful guidance documents on synthesis processes. Though they may not be specific to EHA, the principles still hold true. For instance, the Oxfam GB and the Feinstein International Center (n.d.) guidance note for its Humanitarian Evidence Programme aims to synthesise research in the humanitarian sector.

There are several examples of agencies gathering and synthesising learning across a range of evaluations of humanitarian action:

- CARE has carried out a meta-review of evaluations and after-action reviews from 15 emergency responses, and drawn together key lessons from this (Oliver, 2007: Annex 1A).
- The WFP Office of Evaluation Annual Evaluation Report makes strategic recommendations for WFP based on the synthesis of findings, conclusions, lessons and recommendations from all evaluations completed by the Office of Evaluation during the previous year, and reports on related evaluation activity to strengthen the evaluation function in WFP and the international system:
www.wfp.org/content/annual-evaluation-report-2013-0.
- Periodically, NORAD produces a synthesis report of lessons from evaluations (Disch et al., 2008; Stokke, 2007).
- ACF and Tearfund produce annual learning reports that are heavily based on evaluation findings. Respectively, these are the ACF Learning Review (ACF, 2014) and Tearfund's Learning and Impact Report (Tearfund, 2015).

- Since 2013, UN agencies carry out annual meta-evaluations to review how adequately gender has been integrated into their evaluation reports, using a pro-forma developed by UNEG. These meta-evaluations are used for accountability purposes to assess whether agencies are meeting specified minimum requirements based on the UNEG Norms and Standards for Evaluation. In early 2016, for example, WFP commissioned independent consultants to carry out such a meta-evaluation. In WFP's case, most of its evaluations relate to humanitarian assistance.

Some meta-reviews have explored the extent to which findings and recommendations from past evaluations have been taken up (Oliver, 2009). This can focus on issues where there appears to be continued failure to learn, and can usefully explore the reasons for resistance to change and propose how such blockages can be addressed in future.

ALNAP Lessons Papers aim to make the lessons of previous responses available in a concise and readable format, in order to inform, and thus improve the performance of future humanitarian action. They are aimed at staff designing and implementing humanitarian responses. The papers are mainly based on evaluation findings held in ALNAP's Humanitarian Evaluation and Learning Portal (HELP), as well as additional material. See all ALNAP Lessons Papers here: www.alnap.org/what-we-do/lessons.

Endnotes

- 1.** The report contents described here are based on the ALNAP Quality Proforma (ALNAP, 2005).
- 2.** See www.sida.se/contentassets/ea2b7ee9cdff40a7a0a54022ebf1167d/support-to-internally-displaced-persons-learning-from-evaluation.-synthesis-report-of-a-joint-evaluation-programme---summary-v_3325.pdf.
- 3.** Based on UNEG (2010) and ALNAP (2005).
- 4.** This section draws heavily on Hallam and Bonino (2013), Capacity Area 3, and Warner (2014).
- 5.** See freshspectrum.com.
- 6.** See www.wfp.org/evaluation/lessons/evaluation-briefs.
- 7.** See www.alnap.org/resources/results.aspx?type=22.
- 8.** www.alnap.org/account/submitresource.aspx.

Notes

Notes

Humanitarian impact evaluations



18 / Humanitarian impact evaluations

18.1 Why is impact evaluation important?

Most evaluations of humanitarian action focus on outcomes, often descriptively. In other words, they describe whether an intended outcome has been achieved, for example whether malnutrition rates have fallen, but struggle to establish a causal relationship between the outcome and the programme or intervention being evaluated.¹ It is even less likely that they capture impact in terms of the wider effects of humanitarian action, even though the ToR for most evaluations of humanitarian action include impact-related questions (Proudlock and Ramalingam, 2009; Knox Clarke and Darcy, 2014: 40-44). There are still few examples of evaluations of humanitarian action that focus solely on impact. There are many reasons for this, some of which are described below in the sub-section on challenges.

The growing interest in impact evaluation of humanitarian action is because donors and operational agencies want to establish the impact of internationally funded humanitarian actions on the lives and livelihoods of the people they aim to assist. For example, donors are asking questions about the impact of their – often large – investment in humanitarian action, usually from an accountability perspective. Operational agencies may want to know which humanitarian responses are most effective, often for learning purposes.

18.2 What is impact in EHA?

The definition of impact, according to the OECD-DAC criterion, is as follows: Impact looks at the wider effects of the programme – social, economic, technical, 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 (based on Beck, 2006: 21).

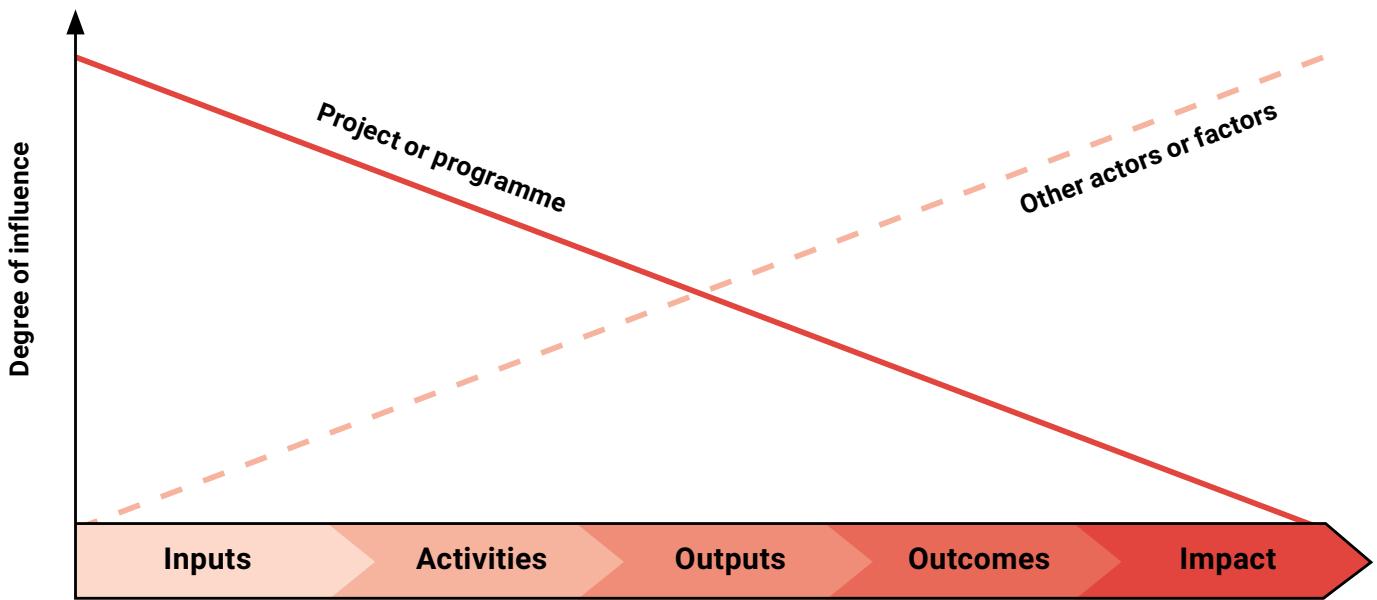
In practice, agencies interpret impact in different ways in relation to evaluating humanitarian action. Some focus on which outcomes can be attributed to the

intervention. For example, can a reduction in malnutrition rates be attributed to the provision of food aid? Other agencies want to know about the wider impact, including the ripple effects of humanitarian action, intended and unintended, positive and negative. For example, to what extent has the provision of food aid affected cereal markets or local food production?

The essence of impact evaluation is that it explores cause and effect or 'causal inference'. It shifts the focus away from the effectiveness and efficiency of the intervention, to examine whether people are better off or safer as a result. Establishing this causal relationship is the challenge.

In some contexts an impact evaluation may be asked to look at the long-term, lasting and significant changes that can be attributed to humanitarian action. After a natural disaster, such as floods or droughts for example, this might include evidence of improved preparedness or resilience to future disasters. In a protracted conflict, where there is widespread and long-term displacement, such as in South Sudan or as is associated with the Syria crisis, the evaluation may be more focused on the immediate impact, for example the impact of providing humanitarian assistance to displaced persons on the host population, rather than lasting impact.

Figure 18.1: Relative Influence along the results chain



Source: Smutylo (2001)

The fundamental challenge entailed in impact evaluation is attribution, in other words isolating the impact that is due to a given intervention from the many other factors at play. As demonstrated in [Figure 18.1](#) (Smutylo, 2001), the influence of the action of a single agency (or even multiple agencies) decreases along the results chain as the influence of other actors and factors. For example:

- Although humanitarian agencies may have delivered food aid effectively and provided primary health care to the affected population, the positive impact of such interventions may have been minimal if the conflict intensifies, those communities become displaced, and the humanitarian crisis deepens.
- Conversely, the food security of an affected population may have improved, but the extent to which that can be attributed to humanitarian action is open to question. For example, other extraneous factors could have played a role, such as when greater security means that markets can function more effectively, or seasonal factors such as harvest time mean that there is more food available.

The challenge of attribution is not unique to EHA, but it is amplified in most humanitarian contexts because of the difficulties of assembling high-quality evidence. This, in turn, can be due to the following contextual challenges in humanitarian crises, as described in [Section 1: What is Evaluation of Humanitarian Action?](#):

1. **Challenge:** The dynamic and fluid environment in most humanitarian contexts, with many and unpredictable factors affecting outcomes and impact, including a range of diverse actors.

Potential solution: consider joint impact evaluations (see below).

2. **Challenge:** Lack of data, including:
 - a. basic data required to design certain evaluation methods, such as on the population or number of people affected by the crisis
 - b. baseline data on key indicators related to wellbeing, for example on livelihoods, or access to education, against which it is possible to assess whether there has been a change
 - c. available and high quality monitoring data that shows change over time (monitoring data are usually focused on process and outputs rather than outcomes)

Potential solution: use recall, and ask members of the affected population about their previous situation and how it has changed; use alternative methods to deal with the absence of basic data, e.g. satellite imagery to count dwellings.

In addition to these contextual challenges, some evaluation-specific challenges affect evaluation in the humanitarian sector:

1. The need for rapid action in an unpredictable environment, which means that there is little time for advance preparation for an impact evaluation, from the early stages of the crisis and response (Puri et al., 2014). Impact evaluations tend to be planned late in the programme cycle.
2. Selecting the most appropriate design and blend of approaches that are best suited to answer specific cause-and-effect questions.
3. Impact evaluation requires different skill levels than conventional evaluations. The data collection and analysis requirements may require a more research-oriented set of skills. These skills have generally been scarce in the humanitarian sector (Alexander and Bonino, 2015; Proudlock et al., 2009).

As a result of these challenges, and especially the challenge of isolating the impact of a particular intervention, impact evaluation of humanitarian action tends to focus more on partial attribution or on the contribution of a particular intervention to change (see Section 1: What is Evaluation of Humanitarian Action?).

18.3 Planning an impact evaluation

The particular challenges of undertaking an impact evaluation mean that plenty of time needs to be allowed for the planning stage. Consider the following questions:

Are the conditions right for an impact evaluation?²

As high-quality impact evaluations require considerable resources, ensure that at least one of the following conditions are met to justify the investment:

- Is the intervention significant enough (in terms of size, policy prominence, or potential consequences) to call for such a specific type of evaluation?

- Is it strategically relevant in terms of the potential learning and influence of such an impact evaluation?
- Is there untested or contested evidence of ‘what works, for whom, and where’ that the proposed impact evaluation could illuminate?
- Are the conditions conducive for the evaluation to be influential? In other words, how will the findings be used? Ensure you are clear about this in order to be utilisation-focused.
- Is the programme or intervention? For example, has it been running for long enough to show results?
- Is there a sufficient budget for the sample sizes that will be needed to demonstrate impact at the likely effect size ([see Section 12: Sampling](#)).



Tip

After addressing the six questions above, consider running an Evaluability Assessment ([see Section 2: Deciding to do an evaluation](#)) before launching a resource-intensive humanitarian impact evaluation.

What are you aiming to demonstrate or learn through an impact evaluation?

Being clear about this will help to determine the scope of the impact evaluation. For example, is the focus on whether the intervention had an impact on the affected population as intended by programme planners and funders? Or is it on the wider impact of the humanitarian intervention, perhaps to locate its significance in relation to what people have done for themselves?

When in the programme cycle is it appropriate to launch an impact evaluation?

The answer to this question will, in part, depend on the answer to the question above. If you are interested in the immediate effect of an intervention, you may want to launch your impact evaluation a few months into the implementation phase – for example, to track and understand the immediate impact of an emergency shelter programme. If you are interested in the longer-term consequences and impact of humanitarian action, you may want to launch your impact evaluation several years after implementation began – for example, to explore the impact of a programme designed to build resilience at the community and household level.

Is this best done as a single agency or a joint evaluation?

[Joint impact evaluations](#) can help to ensure that the overall impact of humanitarian action is explored rather than artificially attempting to isolate a single agency’s work (Beck, 2009: 1). They also allow comparison of different

approaches, which may help to evaluate the relative impact of different interventions. In the humanitarian sector, many agencies carrying out impact evaluations are doing so as a joint exercise.

What are the appropriate evaluation questions?

Once it is clear what the impact evaluation intends to find out or learn it is then possible to articulate the relevant evaluation questions. As mentioned in [Section 6: Choosing evaluation questions](#), it is important to identify a few high-level questions designed to give you the answers you are looking for. Remember that your evaluation questions should drive the method and design choice, not vice versa.

What are the best design, methods and tools available that will answer these questions?

In order to answer this question, pay attention to the programme attributes. For example, does the programme have a theory of change that can be tested? Is it possible to create a counterfactual? See [Section 18.4 opposite](#).

What kind of evaluation skills are needed?

Once the scope of the evaluation and the evaluation questions have been established, and there is some idea of the most appropriate methods, the next step is to identify the necessary evaluation skills – in particular, whether conventional skills are adequate or whether there is a need for evaluators with qualitative and/or quantitative research skills. Proudlock et al. suggest that ‘the design and implementation of impact assessments requires skills available only through investment in long-term partnerships between academics, donors, governments, practitioners and targeted recipients’ (2009: 7).

What are the budgetary and resource implications?

As thorough impact evaluations are likely to require longer and more intensive fieldwork than other types of EHA, and perhaps greater skills, they are also likely to require larger budgets than a conventional EHA.



Tip

Agencies should prioritise a small number of high-quality and strategic humanitarian impact evaluations rather than undertaking numerous impact evaluations that may be poorly resourced and therefore at risk of producing low-quality results.

18.4 Approaches and designs

The following approaches and methods are appropriate for humanitarian impact evaluation to infer causation. See [Section 11](#) for more detail on each of these methods. The list is not exhaustive and other possible methods could be considered. More detailed guidance is available through initiatives led by InterAction and UNICEF's Office of Research – Innocenti.³

Approach/ design	Potential	Constraints	Tips
RCTs or comparison group design	When feasible, can help to establish causation by providing credible comparative data and analysis	Rarely feasible for practical and ethical reasons in humanitarian contexts. Pre-test rarely feasible, only post-test	Consider opportunistic comparison groups, e.g. camps that have received different levels or types of assistance
Theory-based approaches	Test the underlying theory of change of a programme intervention, e.g. testing assumptions	Theory of change may not have been articulated by the programme planners or implementers	
Case-based approaches	Can use qualitative methods to explain how an intervention could be responsible for particular changes, based on a specific case	Difficulties of generalising from case studies in a diverse and rapidly changing context	Cases should be selected purposively
Longitudinal study	Research over a period of time to capture changes in the lives of the affected population	Likely to be resource-intensive and may not be feasible in contexts where population groups are highly mobile because of continued conflict, or where access is intermittent	It may be possible to use data from an existing longitudinal study – see Khoo (2010) for an example
'Contribution to change' approach (Few et al., 2014)	Assesses changes over time in the lives of affected people, the extent to which their livelihoods and well-being have recovered, and the role that interventions appear to have played in that recovery process	Assumes that changes in people's wellbeing and livelihoods can be most clearly identified at a household level. Designed mainly to be used following rapid-onset natural hazards such as flash floods, storms, landslides, earthquakes, tsunamis and volcanic eruptions	

A mixed-method approach is particularly appropriate to humanitarian impact evaluation, especially when mixed methods support each other in a synergistic way (Adato, 2011). Often it is only mixed methods that can deal with the large number of potentially confounding factors found in the typical context of humanitarian action. The Feinstein Center has pioneered a participatory approach to impact assessment that creates a learning partnership between the donor and implementing partners (Catley et al., 2009). Good practice example on pg 365 demonstrates how qualitative PRA methods can be used to explore impact with affected communities in participatory ways, even in a challenging conflict environment.



Tip

In selecting your impact evaluation design, consider what level of certainty and precision the evaluation must deliver in answering particular questions in order to be considered credible. This will help to inform the sampling approach, and any trade-offs in terms of budget and methods.



Tip

Extensive and in-depth consultation with the affected population is critical in impact evaluation in order to gain their perspective. Humanitarian impact evaluation almost always depends, to some extent, on value judgement, which means that the judgement of the affected population should be given high priority in terms of causation.



Good practice example: Joint WFP and UNHCR Impact Evaluation series on the Contribution of Food Assistance to Durable Solutions in Protracted Refugee Situations

WFP and UNHCR jointly commissioned and conducted a series of mixed-method impact evaluations in Bangladesh, Chad, Ethiopia and Rwanda, to assess the contribution of food assistance to durable solutions in protracted refugee situations (WFP and UNHCR, 2012). Overall, the impact evaluation series used a theory-based approach as the basis to infer causation. The four evaluations used the same theoretical framework and approach, but with details adapted to each context. Some of the key features of this impact evaluation series are its use of:

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

Source: Alexander and Bonino (2015: 11).



Good practice example: Participatory Impact Assessments

For a good example of a Participatory Impact Assessment, see FAO's distribution of emergency livelihood kits in South Sudan – April-May (2015).

As part of the evaluation of FAO's programme in South Sudan (focusing on the response to the Level 3 emergency), the FAO Office of Evaluation conducted a participatory impact assessment in some of the areas most severely affected by the conflict in 2014, when FAO had responded by distributing livelihoods kits containing staple crops and vegetable seeds, tools and fishing equipment.

The assessment sought to collect first-hand information from targeted communities on the positive, negative, intended and unintended impacts of distribution in order to improve future emergency interventions and to hear feedback from those who had been affected. Participatory research methods were used, in particular qualitative PRA tools. Village resource maps and timelines were drawn up by separate focus groups of men and women in order to understand the changes in communities' assets and vulnerability. Matrices on income, expenditure and coping strategies, and daily activity clocks were used to assess the impact of distribution at the household level as well as nutrition issues. Three different geographical areas were visited in order to include locations under both government and opposition control.

A team of national staff was selected with diverse ethnic backgrounds that allowed them to access the different areas. They were then trained in PRA methods and asked to carry out a trial of the methodologies in a 'safe' location. They then developed a training manual in order to train local enumerators in PRA. The results of the assessment have been incorporated in the 2016 FAO emergency programme in South Sudan

Source: (FAO, mimeo).

See also [Good practice example on pg 290 in Section 15: Constrained access](#) for a brief description of an impact study carried out in Somalia as part of the 2012/13 evaluation of FAO's cooperation in Somalia between 2007 and 2012.

Endnotes

- 1.** See Alexander and Bonino (2015) and Morra Imas and Rist (2009).
- 2.** From Bonino (2015), which is summarised in the protection guide, based on Rogers (2012) and Chigas, Church and Corlazzoli (2014).
- 3.** InterAction developed an Impact Evaluation Guidance Note and Webinar Series: www.interaction.org/impact-evaluation-notes.

The UNICEF Office of Research – Innocenti has collaborated with RMIT University, BetterEvaluation and the International Initiative for Impact Evaluation (3ie) to produce a series of methodological briefs and videos on impact evaluation: www.unicef-irc.org/KM/IE/impact.php.

Glossary

Accountability / 27

The means through which power is used responsibly. It is a process of taking into account the views of, and being held accountable by, different stakeholders, and primarily the people affected by authority or power.

After-action review / 265

A structured discussion of a humanitarian intervention that enables a team to consider and reflect on what happened, why it happened, and how to sustain strengths and improve on weaknesses.

Attribution / 29

The ascription of a causal link between observed (or expected to be observed) changes and a specific intervention.

Availability sampling / 217

A sampling procedure where the selection is based on their availability, researcher convenience, or selfselection.

Before and after comparison / 204

Compares the situation of a group before and after the intervention.

Case studies / 200

Intensive descriptions and analysis of one or more cases (which can range from individuals to states) to draw general conclusions about intervention.

Causal inference / 302

The establishing of a relationship between a cause and an effect.

Cluster evaluation / 80

Evaluation of multiple projects within a larger programme, OR evaluation related to the UN Cluster Coordination System.

Cluster sampling / 225

Sampling where a number of locations are sampled, each with a cluster of a particular number of cases.

Coding / 304

Assigns categories to particular pieces of evidence.

Comparison group / 205

Comparison group designs compare the assisted group with a selected comparison group.

Conclusion / 319

An inductive statement based on one or more findings.

Confidence interval / 227

The range within which we expect the value in the population as a whole to fall.

Confidence level / 227

The probability that the value in the population as a whole to fall within the confidence interval.

Content analysis / 180

Content analysis is analysis of textual information in a standardised way that allows evaluators to make inferences about the information.

Contribution / 29

Analysing contribution in evaluation refers to finding credible ways of showing that an intervention played some part in bringing about results. Contribution analysis is a kind of evaluative analysis that recognises that several causes might contribute to a result, even if individually they may not be necessary or sufficient to create impact.

Crowd-sourcing / 286

Uses a large number of volunteers either to collect data or to analyse imagery data, usually through indirect means. This type of data is called crowd-sourced data.

Data saturation / 217

Data saturation occurs when new cases no longer add new knowledge.

Descriptive statistics / 312

Statistics used to summarise key aspects of a population.

Design effect / 225

The factor by which you have to modify your sample size when you depart from simple random sampling.

Differences in difference / 204

This design estimates the effect of assistance by comparing the average

change over time in the outcome of interest between the assisted group and a comparison group.

Dissemination

The process of communicating information to specific audiences for the purpose of extending knowledge and with a view to modifying policies and practices.

Effect size / 228

The proportionate difference between the variable of interest in the treated and control group.

Evaluation of Humanitarian Action / 27

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.

Evaluand / 210

The subject of an evaluation, typically a programme or system rather than a person.

Evaluation synthesis / 349

Follows a systematic procedure to organise and summarise the findings from a set of (possibly quite disparate) evaluations. This can be an opportunity to ask particular questions about the set of evaluations, and/or identify common trends in the evaluation findings.

Evaluative reasoning / 297

The analytical process by which evaluators answer evaluative questions.

Evaluative rubric / 297

A table that describes what the evidence should look like at different levels of performance, on some criterion of interest or for the intervention overall.

Evidence / 293

The available body of facts or information that can support a particular proposition or belief.

Ex-ante evaluation / 83

An evaluation performed before an intervention begins.

Experimental designs / 199

Experimental designs are where units of analysis are randomly assigned to the assisted or the control group. Each element e.g. a person, family, or community) has an equal chance of being assigned to either the assisted or the control group.

Ex-post evaluation / 83

An evaluation performed after an intervention has been completed.

External or independent evaluations / 85

An evaluation carried out by evaluators who are outside the implementing team.

Finding / 319

A factual statement based on evidence.

Humanitarian action / 24

The objectives of humanitarian action are to save lives, alleviate suffering and maintain human dignity during and in the aftermath of crises and natural disasters, as well as to prevent and strengthen preparedness for the occurrence of such situations.

Humanitarian portfolio**evaluation / 81**

An evaluation of the whole humanitarian portfolio of an agency.

Impact / 29

Looks at the wider effects of the programme – social, economic, technical and environmental – on individuals, gender, age-groups, communities and institutions. Impacts can be intended and unintended, positive and negative, macro (sector) and micro (household, individual), short or long term.

Impact evaluation / 81

An evaluation that focuses on the wider effects of the humanitarian programme, including intended and unintended impact, positive and negative impact, macro (sector) and micro (household, individual) impact.

Inception phase / 135

The inception phase of the evaluation goes from the selection of the evaluation team up to approval of the inception report.

Inferential statistics / 312

Statistics used either to make inferences about a population from a sample, or to make inferences about hypotheses.

Inputs / 28

The financial, human and material resources used in the humanitarian action.

Institutional evaluation / 86

Evaluation of the internal dynamics of implementing organisations, their policy instruments, service delivery mechanisms and management practices, and the linkages among these.

Interrupted time series / 203

Provides an estimate of the impact of an intervention by examining a time series of data before and after an intervention.

Joint evaluation / 84

An evaluation carried out by two or more agencies, evaluating the work of two or more agencies.

Keyword analysis / 181

Keyword analysis is a form of content analysis that examines the frequency of occurrence of categories to highlight trends over time in a single document set or to compare two document sets.

Learning / 27

The process through which experience and reflection lead

to changes in behaviour or the acquisition of new abilities.

Logic model / 92

A table or diagram presenting the programme theory (the way in which inputs are expected to contribute to the overall goal) for an intervention.

Longitudinal study / 206

Study making repeated measurements of the same population over years.

Management group / 150

A group that manages the evaluation on a day-to-day basis, including drafting the ToR, contracting and managing the evaluation team, and managing the review and finalisation of the evaluation report.

Management response matrix / 347

A record of management's response to each evaluation recommendation and the steps managers plan to take to address it, with a target date and responsible party for each step.

Meta-evaluation / 87

An overarching evaluation designed to aggregate findings from a series of evaluations, usually against a set of benchmarks. It can also be used to denote the 'evaluation of an evaluation', in other words to judge the quality of a set of evaluations.

Mid-term evaluation / 83

An evaluation performed towards the middle of an intervention.

Monitoring / 30

A continuing function that uses systematic collection of data on specified indicators to provide management and the main stakeholders of an ongoing humanitarian intervention with indications of the extent of progress, achievement of objectives and progress in the use of allocated funds.

They can be desired (positive) or unwanted (negative). They can encompass behaviour change (actions, relations, policies, practices) of individuals, groups, communities, organisations, institutions or other social actors.

Natural experimental design / 208

Design that where a comparison between an assisted group and a similar group that, by chance, has not been assisted.

Outcome review / 202

Compares outcomes with planned outcomes.

Outputs / 28

The products, goods and services which result from an intervention.

Non-experimental designs / 195

Designs where there is no comparison, either between assisted and non-assisted populations, or for those who have received assistance over time.

Participatory design / 202

Participatory design involves all stakeholders throughout all phases of the evaluation, from the initial planning to the implementation of the recommendations

Non-random sampling / 214

Selects the sample based on some property of the sample.

Participatory evaluation / 85

An evaluation in which stakeholders, including the affected population, work together to design, carry out, and interpret an evaluation.

Normative evaluation / 82

An evaluation that compares what is being implemented with what was planned or with specific standards.

Participatory Rapid Appraisal / 276

Participatory Rapid Appraisal or Participatory Rural Appraisal refers to a group of methods enabling local people to enhance, analyse, and share their knowledge and learning in ways that outside evaluators can readily understand.

Ongoing evaluation / 83

A series of evaluations designed to run throughout an intervention.

Partner evaluation / 80

Evaluation of a set of interventions implemented by a single partner.

Outcomes / 28

Intended or unintended changes or shifts in conditions due directly or indirectly to an intervention.

Peer-review group / 151

A group that advises on quality issues, usually made up of evaluators and other specialists chosen for their knowledge of evaluation, the region, or the type of intervention being evaluated.

Policy evaluation / 86

An evaluation that examines the understandings, beliefs, and assumptions that make individual projects possible as well as desirable. It may evaluate both the efficacy of the policy itself and how that policy has been implemented.

Primary data / 293

Data collected for the purpose of the evaluation.

Process evaluation / 82

An evaluation that focuses on the processes by which inputs are converted into outputs; may also examine the intervention as a whole.

Process review / 201

Compares how processes function with how they were planned to function.

Programme evaluation / 79

Evaluation of a set of interventions with a unifying humanitarian objective.

Project evaluation / 79

Evaluation of a single humanitarian intervention with specific objectives, resources, and implementation schedule, which often exists within the framework of a broader programme.

Propensity score matching / 198

A statistical matching technique that attempts to match the comparison group to the control group through selecting one with the same probability of being assisted based on the group's characteristics.

Protection / 25

Comprises 'all activities aimed at obtaining full respect for the rights of the individual in accordance with the letter and the spirit of the relevant bodies of law' (IASC, 2011).

Pseudo-random sampling / 224

Sampling where there is no sampling frame. Typically, the first instance is randomly selected from a purposively selected starting point, and subsequent instances are selected using some rule.

Purposive sampling / 218

Purposive sampling selects the sample based purposively so that the sampled elements can provide the most information for the study.

Quasi-experimental designs / 197

Designs using a comparison where the comparison group is not randomly selected.

Quota sampling / 219

Divides the population into mutually exclusive subcategories, and then collects information for a previously established sample size or proportion for each category.

Randomised control trial / 207

Compares two randomly selected groups, one of which is given assistance while the other (the control group) receives none.

Random sampling / 215

Draws a sample from a population where each member of the population has an equal chance of being selected.

Recommendation / 312

A course of action the evaluators suggest as a way to address one or more conclusions.

Real-time evaluation / 82

An evaluation of an ongoing humanitarian operation as it unfolds.

Reference group / 150

A group made up of primary stakeholders familiar with the local environment who can advise on practical issues associated with the evaluation and on the feasibility of the resulting recommendations.

Regression Discontinuity**Design / 205**

Compares the regression lines for the variable of interest against the score on which the intervention was based.

Rubric / 178

A scoring tool used to assess a document against a set of criteria in a consistent way.

Sampling / 214

The selection of a subset of a population for inclusion in a study instead of the entire population.

Secondary data / 293

Secondary data is data collected for other purposes but is used by the evaluation.

Sector evaluation / 80

Evaluation of a group of interventions in a sector, all of which contribute to the achievement of a specific humanitarian goal. The evaluation can cover part of a country, one country, or multiple countries (UNICEF, 2013).

Self-evaluation / 85

An evaluation carried out by those who design and deliver an intervention, in other words an internal evaluation.

Single-agency evaluation / 84

An evaluation carried out by the agency that implemented the intervention.

Standard deviation / 228

A measure of the variability of a parameter.

Statistical power / 228

Statistical power is the probability that a negative result is not a false negative (1 minus the risk of a false negative result).

Steering group / 149

A group established to steer an evaluation through key stages such as establishing the ToR, writing the inception report, and drafting the final report.

Stratified sampling / 226

A sampling approach where the population is first divided into mutually exclusive segments and a simple random sample is taken from each one.

Survey instrument / 233

The questionnaire used by the interviewer during formal survey interviews.

System-wide evaluation / 84

An evaluation of the international humanitarian system's response to a humanitarian crisis, open to all actors in the system.

Technology evaluation / 86

An evaluation of a specific technique or technology.

Thematic evaluation / 81

An evaluation of a selection of interventions that all address a specific humanitarian priority that cuts across countries, regions, and possibly agencies and sectors.

Thematic review / 350

A thematic review draws out the findings and learning from a series of evaluations of humanitarian action in

a particular sub-sector, e.g. protection, shelter. These can be commissioned to answer/address a particular set of overarching questions.

Theory of change / 97

A description of the central mechanism by which change comes about for individuals, groups, and communities.

Terms of Reference / 118

The ToR presents 'an overview of the requirements and expectations of the evaluation. It provides an explicit statement of the objectives of the evaluation, roles and responsibilities of the evaluators and the evaluation client, and resources available for the evaluation' (Roberts et al., 2011: 2).

Treatment discontinuity**comparison / 206**

Compares the group just below the cut-off point for assistance with the group just below.

Triangulation / 211

Comparing data from different sources to see whether they support the same finding. (Discussed in section 13).

Utilisation / 62

An evaluation has been utilised if users with the intention and potential to act have given serious, active consideration to its findings, identifying meaningful uses according to their own interests and needs (Sandison, 2006: 100-101).



Bibliography

The following publications can also be accessed via the Humanitarian Evaluation and Learning Portal (HELP): www.alnap.org/resources/EHA-2016

Abate, G. T., de Brauw, A., Minot, N. and Bernard, T. (2014) *The impact of the use of new technologies on farmers wheat yield in Ethiopia: Evidence from a randomized controlled trial.* Washington, DC: REAP. (www.alnap.org/resource/22846.aspx).

Abdullah, M. Y. H., Bakar, N. R., Sulehan, J., Awang, A. H. and Liu, O. P. (2012) 'Participatory rural appraisal (PRA): An analysis of experience in Darmareja village, Sukabumi District, West Java, Indonesia'. *Akademika*, 82(1): 15–19. (www.alnap.org/resource/22847.aspx).

ACAPS. (2011) *Technical brief: Direct observation and key informant interview techniques for primary data collection during rapid assessments.* Geneva: ACAPS. (www.alnap.org/resource/22848.aspx).

ACAPS. (2013) *How sure are you? Judging quality and usability of data collected during rapid needs assessments.* Geneva: ACAPS. (www.alnap.org/resource/11437.aspx).

ACF. (2014) *ACF Learning Review 2014.* Paris: ACF. (www.alnap.org/resource/20231.aspx).

Adato, M. (2011) *Combining quantitative and qualitative methods for program monitoring and evaluation: Why are mixed-method designs best?* Washington, DC: The World Bank. (www.alnap.org/resource/8069.aspx).

Adinolfi, C., Bassiouni, D., Lauritszen, H. and Williams, H. (2005) *Humanitarian Response Review.* New York: UN OCHA. (www.alnap.org/resource/3284.aspx).

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

Ager, A., Stark, S. and Potts, A. (2010) *Participative ranking methodology: A brief guide: Version 1.1.* New York: Columbia University. (www.alnap.org/resource/8070.aspx).

Ager, A., Stark, L., Sparling, T. and Ager, W. (2011) *Rapid appraisal in humanitarian emergencies using participatory ranking methodology (PRM)*. New York: Columbia University. (www.alnap.org/resource/22849.aspx).

Ahmed, L., Malik, Z. and Nawab, B. (2012) *Humanitarian support to conflict and flood-affected populations in Pakistan*. Paris: ACF. (www.alnap.org/resource/21080.aspx).

Alexander, J. (2014) *Improving accuracy in humanitarian evaluations*. London: ALNAP. (www.alnap.org/resource/12636.aspx).

Alexander, J. and Bonino, F. (2014) *Ensuring quality of evidence generated through participatory evaluation in humanitarian context*. London: ALNAP. (www.alnap.org/resource/19163.aspx).

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

Alexander, J. and Cosgrave, J. (2014b) *Representative sampling in humanitarian evaluation*. London: ALNAP. (www.alnap.org/resource/10389.aspx).

Alexander, J., Darcy, J. and Kiani, M. (2013) *The 2013 humanitarian accountability report*. Geneva: HAP. (www.alnap.org/resource/8758.aspx).

ALNAP. (2002) *ALNAP annual review 2002: Humanitarian action: improving performance through improved learning*. London: ALNAP. (www.alnap.org/resource/5189.aspx).

ALNAP. (2005) *Assessing the quality of humanitarian evaluations. The ALNAP quality proforma 2005*. London: ALNAP. (www.alnap.org/resource/5320.aspx).

ALNAP. (2007a) *An introduction to evaluation of humanitarian action (EHA) course manual*. London: ALNAP. (www.alnap.org/resource/22850.aspx).

ALNAP. (2007b) *Slow-onset disasters: drought and food and livelihoods insecurity—learning from previous relief and recovery responses*. London: ALNAP. (www.alnap.org/resource/5243.aspx).

ALNAP. (2009) *Participation by crisis-affected populations in humanitarian action. A handbook for practitioners*. London: ALNAP. (www.alnap.org/resource/5271.aspx).

ALNAP. (2015) *The state of the humanitarian system*. London: ALNAP.
(www.alnap.org/resource/21036.aspx).

ALNAP and DEC. (2013) *Workshop Summary. Syria Crisis: Evaluators' learning exchange*. London: ALNAP. (www.alnap.org/resource/9300.aspx).

ALNAP and Groupe URD. (2009) *Participation handbook for humanitarian field workers: Involving crisis-affected people in humanitarian response*. London and Plaisians: ALNAP and Groupe URD. (www.alnap.org/resource/8531.aspx).

ALNAP, OECD-DAC, and UNEG. (2010) *Supporting Evaluation in Haiti Mission Report*. London: ALNAP. (www.alnap.org/resource/22851.aspx).

Amazon Fund. (2010) *Logical framework*. Amazon Fund.
(www.alnap.org/resource/22852.aspx).

AEA. (2003) *American Evaluation Association response to US Department of Education*. Washington, DC: AEA. (www.alnap.org/resource/22855.aspx).

AEA. (2004) *Guiding principles for evaluators*. Washington, DC: AEA.
(www.alnap.org/resource/22853.aspx).

Anderson, A.A. (2005) *The community builder's approach to theory of change: a practical guide to theory development*. New York: The Aspen Institute.
(www.alnap.org/resource/22854.aspx).

Ashdown, P. (2011) *Humanitarian emergency response review*. London: DFID.
(www.alnap.org/resource/6355).

Audsley, B., Halme, R. and Balzer, N. (2010) 'Comparing cash and food transfers: a cost-benefit analysis from rural Malawi', in Were Omamo, S., Gentilini, U. and Sandström, S. (eds) *Revolution: From food aid to food assistance: Innovations in overcoming hunger*. Rome: WFP. (www.alnap.org/resource/22856.aspx).

AusAID. (2013) *Shelter sector response evaluation typhoon Pablo December 2012 in Mindanao, Philippines - shelter cluster report*. Canberra: AusAID.
(www.alnap.org/resource/10074.aspx).

Baez, J.E. and Santos, I.V. (2007) *Children's vulnerability to weather shocks: A natural disaster as a natural experiment*. New York: Social Science Research Network. (www.alnap.org/resource/8078.aspx).

Bailey, S. (2013) *Evaluation of Concern Worldwide's emergency response in Masisi, North Kivu, DRC (2012-2013).* London: Concern.
(www.alnap.org/resource/22857.aspx).

Baker, J., Chantal, S., Hidalgo, S., Kayungura, G., Posada, S., Tasikasereka, M. and Vinas, M. (2013) *External evaluation of the rapid response to population movements (RRMP) program in the Democratic Republic of Congo.* Madrid: DARA.
(www.alnap.org/resource/12494.aspx).

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

Bamberger, M., Carden, F. and Rugh, J. (2009) *Alternatives to the conventional counterfactual: summary of session 713 think tank.* American Evaluation Association, Orlando 2009. Washington, DC: AEA. (www.alnap.org/resource/8222.aspx).

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

Bamberger, M., Rugh, J. and Mabry, L. (2011) *RealWorld evaluation. Working under budget, time, data, and political constraints: A condensed summary overview.* 2nd ed. London: SAGE. (www.alnap.org/resource/8076.aspx).

Bamberger, M. and Segone, M. (2011) *How to design and manage equity-focused evaluations.* New York: UNICEF. (www.alnap.org/resource/8080.aspx).

Barahona, C. (2010) *Randomised control trials for the impact evaluation of development initiatives: a statistician's point of view.* Rome: ILAC.
(www.alnap.org/resource/22858.aspx).

Barakat, S. (2006) *Mid-term evaluation report of the national solidarity programme (NSP), Afghanistan.* York: University of York. (www.alnap.org/resource/22859.aspx).

Barakat, S., Hardman, F., Connolly, D., Sundaram, V. and Zyck, S. A. (2010) *Programme review and evaluability study (PRES): UNICEF's education in emergencies and post-crisis transition (EEPCT) programme.* New York: UNICEF.
(www.alnap.org/resource/8081.aspx).

Barron, I. G., Abdallah, G. and Smith, P. (2012) 'Randomized control trial of a CBT trauma recovery program in Palestinian schools'. *Journal of Loss and Trauma*, 2013(18): 306-321. (www.alnap.org/resource/22860.aspx).

Baugh, J. B., Hallcom, A. S. and Harris, M. E. (2010) 'Computer assisted qualitative data analysis software: A practical perspective for applied research'. *Revista del Instituto Internacional de Costos*.
(www.alnap.org/resource/22861.aspx).

Beall, J. and Schütte, S. (2006) *Urban livelihoods in Afghanistan*. Kabul: AREU.
(www.alnap.org/resource/22862.aspx).

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

Beck, T. (2009) *Joint humanitarian impact evaluation: options paper*. London: ALNAP. (www.alnap.org/resource/5741.aspx).

Beck, T. (2011) *Joint humanitarian impact evaluation: report on consultations: Report for the inter-agency working group on joint humanitarian impact evaluation*. London: ALNAP. (www.alnap.org/resource/6046.aspx).

Beck, T. and Buchanan-Smith, M. (2008) 'Joint evaluations coming of age? The quality and future scope of joint evaluations', in *ALNAP 7th Review of Humanitarian Action*. ALNAP Review. London: ALNAP.
(www.alnap.org/resource/5232.aspx).

Benatar, S. R. (2002) 'Reflections and recommendations on research ethics in developing countries'. *Social Science & Medicine*, 54(7).
(www.alnap.org/resource/8084.aspx).

Benham, C., Cascioli Sharp, R., Chamberlain, P., Frederick, J., Gorgonio, R., Jensen, K., Rivera, J. and Veso, Y. (2014) *Real-time evaluation of World Vision's response to typhoon Haiyan*. Monrovia, CA: World Vision.
(www.alnap.org/resource/20190.aspx).

Benini, A. (2009) *Text analysis under time pressure: Tools for humanitarian and development workers*. (www.alnap.org/resource/8189.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*. Gatineau: CIDA.
(www.alnap.org/resource/8287.aspx).

Bentley, M. E., Boot, M., T., Gittelsohn, J. and Stallings, R. Y. (1994)

The use of structured observations in the study of health behaviour.

The Hague: IRC International Water and Sanitation Centre.

(www.alnap.org/resource/22863.aspx).

Berenson, M. L., Levine, D. M. and Szabat, K. A. (2013) 'Estimation and

sample size determination for finite populations', in *Basic business*

statistics: global edition. Harlow: Pearson Education Limited.

(www.alnap.org/resource/22864.aspx).

Berger, R., Pat-Horenczyk, R. and Gelkopf, M. (2007) 'School-based intervention

for prevention and treatment of elementary-students' terror-related distress

in Israel: A quasi-randomized controlled trial'. *Journal of Traumatic Stress*, 20(4):

541-551. (www.alnap.org/resource/22865.aspx).

Berk, R.A. (1983) 'An introduction to sample selection bias

in sociological data'. *American Sociological Review*, 48(3): 386–398.

(www.alnap.org/resource/22866.aspx).

Berlemann, M., Steinhardt, M. F. and Tutt, J. (2015) *Do natural disasters stimulate*

individual saving? Evidence from a natural experiment in a highly developed country.

Bonn: Institute for the Study of Labour. (www.alnap.org/resource/22867.aspx).

BetterEvaluation. (2013) *Understand causes of outcomes and impacts.*

Melbourne: BetterEvaluation. (www.alnap.org/resource/19077.aspx).

Bhattacharjee, A., Jacob, P. M., Sumasundaram, M., Ramachandran, R.,

Thileepan, S., Kumar, S., Srikantharajah, S. and Suppiah, L. (2007) *Final evaluation*

of CARE Australia supported tsunami response in Trincomalee and Batticaloa districts

of Sri Lanka. Canberra: CARE Australia. (www.alnap.org/resource/6029.aspx).

Bhattacharjee, A., Sida, L. and Reddick, M. (2010) *Evaluation of DFID-Unicef*

programme of cooperation: Investing in humanitarian action phase III 2006-2009.

New York: UNICEF. (www.alnap.org/resource/5895.aspx).

Bhattacharjee, A., Wynter, A., Baker, J., Varghese M. and Lowery, C. (2011)

Independent review of UNICEF's operational response to the January 2010

earthquake in Haiti. New York: UNICEF. (www.alnap.org/resource/6353.aspx).

Bilukha, O. (2008) 'Old and new cluster designs in emergency field surveys:

in search of a one-fits-all solution'. *Emerging Themes in Epidemiology*, 5(1).

(www.alnap.org/resource/8085.aspx).

Black, R. E., Allen, L. H., Bhutta, Z. A., Caulfield, L. E., de Onis, M., Ezzati, M., Mathers, C., and Rivera, J. (2008) 'Maternal and child undernutrition: global and regional exposures and health consequences'. *The Lancet*. (www.alnap.org/resource/8086.aspx).

Blake, C. F. (1981) 'Graffiti and racial insults: the archaeology of ethnic relations in Hawaii', in Gould R. A. and Schiffer, M. B. (eds) *Modern material culture: The archaeology of us*. New York: Academic Press. (www.alnap.org/resource/22868.aspx).

Bliss, D. and Campbell, J. (2007) *Recovering from the Java earthquake: Perceptions of the affected*. San Francisco, CA: Fritz Institute. (www.alnap.org/resource/8087.aspx).

Bliss, D. and Larsen, L. (2006) *Surviving the Pakistan earthquake: Perceptions of the affected one year later*. San Francisco, CA: Fritz Institute. (www.alnap.org/resource/8088.aspx).

Boller, K., Buek, K., Burwick, A., Chatterji, M., Paulsell, D., Amin, S., Borkum, E., Campuzano, L., Jacobson, J., Sattar, S., Kadel, S., Pholy, S., Rutajwaha, A., and Sabaa, S. (2011) *Evaluation of UNICEF's early childhood development programme with focus on government of Netherlands funding (2008-2010): Global synthesis report*. New York: UNICEF. (www.alnap.org/resource/8089.aspx).

Bonbright, D. (2012) *Use of impact evaluation results*. Washington, DC: InterAction. (www.alnap.org/resource/8534.aspx).

Bond. (2013) *Principles and checklist for assessing the quality of evidence*. London: Bond. (www.alnap.org/resource/22807.aspx).

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

Bonino, F., Jean, I. and Knox Clarke, P. (2014) *Closing the loop: Effective feedback in humanitarian contexts*. London: ALNAP. (www.alnap.org/resource/10676.aspx).

Boonstra, E., Lindbaek, M., Fidzani, B. and Bruusgaard, D. (2001) 'Cattle eradication and malnutrition in under fives: a natural experiment in Botswana'. *Public Health Nutrition*, 4(4): 877-882. (www.alnap.org/resource/22869.aspx).

Boothby, N., Crawford, J. and Halperin, J. (2006) 'Mozambique child soldier life outcome study: lessons learned in rehabilitation and reintegration efforts'. *Glob Public Health*, 1(1): 87–107. (www.alnap.org/resource/22870.aspx).

Borton, J. (ed.) (1994) *Code of conduct for the International Red Cross and Red Crescent movement and NGOs in disaster relief*. Geneva: IFRC. (www.alnap.org/resource/8091.aspx).

Borton, J., Brusset, E. and Hallam, A. (1996) *The international response to conflict and genocide: Lessons from the Rwanda experience: Humanitarian aid and effects (JEEAR)*. London: ODI. (www.alnap.org/resource/2517.aspx).

Borton, J., Millwood, D. (1996) *The international response to conflict and genocide : Lessons from the Rwanda experience: Study 3: Humanitarian aid and effects*. London: ODI. (www.alnap.org/resource/9976.aspx).

Box, G. E. P. and Tiao, G. C. (1975) 'Intervention analysis with applications to economic and environmental problems'. *Journal of the American Statistical Association*, 70(349): 70-79. (www.alnap.org/resource/22871.aspx).

Brancati, D. (2007) 'Political aftershocks: the impact of earthquakes on intrastate conflict'. *Journal of Conflict Resolution*, 51(5). (www.alnap.org/resource/8092.aspx).

Breier, H. (2005) *Joint evaluations: Recent experiences, lessons learned and options for the future*. Paris: OECD. (www.alnap.org/resource/8093.aspx).

British Parachute Association. (2015) How safe? (www.alnap.org/resource/22872.aspx).

Brookhart, M. A. Wyss, R., Layton, J. B. and Stürmer, T. (2013) 'Propensity score methods for confounding control in nonexperimental research'. *Circulation: Cardiovascular Quality and Outcomes*. (www.alnap.org/resource/22873.aspx).

Broughton, B., Maguire, S. and David-Toweh, K. (2006) *Inter-agency real-time evaluation of the humanitarian response to the Darfur crisis*. New York: OCHA. (www.alnap.org/resource/3245.aspx).

Brown, D. and Donini, A. (2014a) *Rhetoric or reality? Putting affected people at the centre of humanitarian action*. London: ALNAP. (www.alnap.org/resource/12859.aspx).

Brown, D., Donini, A., and Knox Clarke, P. (2014b) *Engagement of crisis-affected people in humanitarian action. ALNAP Background paper*. London: ALNAP. (www.alnap.org/resource/10439.aspx).

Brown, M., Neves, J., Sandison, P., Buchanan-Smith, M., and Wiles, P. (2005)
Evaluation of DFID-UNICEF programme of cooperation to strengthen UNICEF programming as it applies to humanitarian response, 2000-2005. New York: UNICEF.
(www.alnap.org/resource/3355.aspx).

Brusset, E., Bhatt, M., Bjornestad, K., Cosgrave, J., Davies, A., Deshmukh, Y., Haleem, J., Hidalgo, S., Immajati, Y., Jayasundere, R., Mattsson, A., Muhamin, N., Polastro, R. and Wu, T. (2009) *A ripple in development? Long term perspectives on the response to the Indian Ocean tsunami 2004.* Stockholm: SIDA.
(www.alnap.org/resource/5679.aspx).

Brusset, E., Pramana, W., Davies, A., Deshmukh, Y. and Pedersen, S. (2006)
Links between relief, rehabilitation and development in the tsunami response: Indonesia Case Study. London: TEC. (www.alnap.org/resource/5421.aspx).

Bryman, A. (2004) 'Triangulation', in Lewis-Beck, M. S., Bryman, A. and Liao, T. F. (eds) *Encyclopedia of social science research methods*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22874.aspx).

Buchanan-Smith, M., Leyland, T., Aklilu, Y., Noor, M., Ahmed, S., Dini, S. And Robinson, I. (2013) *Evaluation of FAO's cooperation in Somalia: 2007 to 2012.* Rome: FAO. (www.alnap.org/resource/8820.aspx).

Buchanan-Smith, M. et al. (2005) 'How the Sphere Project came into being: A case study of policy making in the humanitarian-aid sector and the relative influence of research', in Court, J., Hovland, I. and Young, J. (eds). *Bridging research and policy in development: Evidence and the change process*. London: ODI. (www.alnap.org/resource/23454.aspx).

Buchanan-Smith, M., Ong, J. C. and Routley, S. (2015) *Who's listening? Accountability to affected people in the Haiyan response.* Woking: Plan International. (www.alnap.org/resource/20632.aspx).

Burns, M., Rowland, M., N'Guessan, R., Carneiro, I., Beeche, A., Sesler Ruiz, S., Kamara, S., Takken, W., Carnevale, P. and Allan R. (2012) 'Insecticide-treated plastic sheeting for emergency malaria prevention and shelter among displaced populations: An observational cohort study in a refugee setting in Sierra Leone'. *The American Journal of Tropical Medicine and Hygiene*, 87(2): 242–250.
(www.alnap.org/resource/22875.aspx).

Bush, J. and Ati, H.A. (2007) *Oxfam's cash transfers in the Red Sea state.* Oxford: Oxfam. (www.alnap.org/resource/8187.aspx).

Bush, K. and Duggan, C. (2013) 'Evaluation in conflict zones: Methodological and ethical challenges'. *Journal of Peacebuilding & Development*, 8(2): 5–25. (www.alnap.org/resource/22876.aspx).

Buttenheim, A. (2009) *Impact evaluations in the post-disaster setting: A conceptual discussion in the context of the 2005 Pakistan earthquake*. New Delhi: 3ie. (www.alnap.org/resource/8206.aspx).

Caliendo, M. and Kopeinig, S. (2005) *Some practical guidance for the implementation of propensity score matching*. Bonn: IZA. (www.alnap.org/resource/22877.aspx).

Cameron, L. A. and Shah, M. (2012) *Risk-taking behavior in the wake of natural disasters*. Bonn: IZA. (www.alnap.org/resource/22878.aspx).

Campo, R. (2006) "Anecdotal evidence": why narratives matter to medical practice'. *PLoS Medicine*, 3(10), 1677-1678. (www.alnap.org/resource/22879.aspx).

Canteli, C., Morris, T. and Steen, N. (2012) *Synthesis of mixed method impact evaluations of the contribution of food assistance to durable solutions in protracted refugee situations*. Rome: WFP. (www.alnap.org/resource/19902.aspx).

Carden, F. (2009) *Knowledge to policy: making the most of development research*. New Delhi: SAGE. (www.alnap.org/resource/22881.aspx).

Carlsson, J., Eriksson-Baaz, M., Fallenius, A. and Lovgren, E. (1999) *Are evaluations useful? Cases from Swedish Development co-operation*. Stockholm: SIDA. (www.alnap.org/resource/12371.aspx).

Catalano, R., Bruckner, T., Gould, J., Eskenazi, B., and Anderson, E. (2005) 'Sex ratios in California following the terrorist attacks of September 11, 2001'. *Human Reproduction*, 20(5): 1221–1227. (www.alnap.org/resource/22882.aspx).

CRS. (2012) *CRS Haiti accountability framework*. Baltimore, MD: CRS. (www.alnap.org/resource/10549.aspx).

Catley, A., Burns, J., Abebe, D. and Suji, O. (2008) *Participatory impact assessment: A guide for practitioners*. Somerville, MA: Feinstein International Centre. (www.alnap.org/resource/8094.aspx).

Catley, A., Burns, J., Abebe, D. and Suji, O. (2013) *Participatory impact assessment: A design guide*. Somerville, MA: Feinstein International Centre. (www.alnap.org/resource/10811.aspx).

CDC. (2009) *Data collection methods for program evaluation: interviews*. Atlanta, GA: CDC. (www.alnap.org/resource/19260.aspx).

Channel Research. (2011) *5-year evaluation of the Central Emergency Response Fund*. London: Channel Research. (www.alnap.org/resource/6357.aspx).

Chapman, N. and Vaillant, C. (2010) *Synthesis of country programme evaluations conducted in fragile states*. London: DFID. (www.alnap.org/resource/6358.aspx).

Chatham House. (2007) *Chatham house rules*. London: Chatham House. (www.alnap.org/resource/8097.aspx).

Checchi, F., Roberts, B. and Morgan, O. (2009) *A new method to estimate mortality in crisis-affected populations: Validation and feasibility study*. Washington, DC: USAID. (www.alnap.org/resource/22883.aspx).

Christensen, G. (2016) *Manual of best practices in transparent social science research*. Berkeley, CA: Berkeley Initiative for Transparency in the Social Sciences. (www.alnap.org/resource/22884.aspx).

Christoplos, I. (2006) *Links between relief, rehabilitation and development in the tsunami response*. London: TEC. (www.alnap.org/resource/3533.aspx).

CHS Alliance. (2014) *Core humanitarian standard on quality and accountability*. Geneva: CHS Alliance. (www.alnap.org/resource/22799.aspx).

CHS Alliance. (2015) *Core Humanitarian CHS Guidance Notes and Indicators*. Geneva: CHS Alliance. (www.alnap.org/resource/22885.aspx).

CIDA. (2000) *How to perform evaluations: Model ToR*. Gatineau: CIDA. (www.alnap.org/resource/22886.aspx).

CIDA. (2005) *Additional resources: Accompanying document of the environment handbook for community development initiatives*. Gatineau: CIDA. (www.alnap.org/resource/8098.aspx).

Clarke, N., Loveless, J., Ojok, B., Routley, S. and Vaux, T. (2015) *Report of the inter-agency humanitarian evaluation (IAHE) of the response to conflict in South Sudan*. New York: OCHA. (www.alnap.org/resource/22828.aspx).

Collier, D. and Mahoney, J. (1996) 'Insights and pitfalls: Selection bias in qualitative research'. *World Politics*, 49(1): 56–91. (www.alnap.org/resource/22887.aspx).

Collinson, S. and Elhawary, S. (2012) *Humanitarian space: a review of trends and issues*. London: ODI. (www.alnap.org/resource/22888.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).

Corlazzoli, V. and White, J. (2013) *Practical approaches to theories of change in conflict, security, and justice. Part II: Using theories of change in monitoring and evaluation*. London: DFID. (www.alnap.org/resource/22890.aspx).

Cornwall, A. (2014) *Using participatory process evaluation to understand the dynamics of change in a nutrition education programme*. Brighton: IDS. (www.alnap.org/resource/22891.aspx).

Cosgrave, J. (2004) *Danish assistance to internally displaced persons in Angola 1999-2003*. London: Channel Research. (www.alnap.org/resource/22892.aspx).

Cosgrave, J. (2008) *Responding to earthquakes 2008: Learning from earthquake relief and recovery operations*. London: ALNAP. (www.alnap.org/resource/5239.aspx).

Cosgrave, J. (2010) *Evaluability assessment of a proposed evaluation of humanitarian interventions in South and Central Somalia: Commissioned by the IASC for Somalia: overall report*. Geneva: IASC. (www.alnap.org/resource/22893.aspx).

Cosgrave, J., Crawford, N. and Mosel, I. (2015) *10 Things to know about refugees and displacement*. London: HPG/ODI. (www.alnap.org/resource/22895.aspx).

Cosgrave, J., Goncalves, C., Martyris, D., Polastro, R. and Sikumba-Dils, M. (2007) *Inter-agency real-time evaluation of the response to the February 2007 floods and cyclone in Mozambique*. Madrid: DARA. (www.alnap.org/resource/3457.aspx).

Cosgrave, J., Polastro, R. and Zafar, F. (2010) *Inter-agency real time evaluation (IA RTE) of the humanitarian response to Pakistan's 2009 displacement crisis*. Geneva: IASC. (www.alnap.org/resource/11896.aspx).

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

Cosgrave, J., Selvester, K., Fidalgo, L., Hallam, A. and Taimo, N. (2001) *Independent evaluation of DEC Mozambique floods appeal funds March 2000 - December 2000 : Volume One: Main Findings*. London: DEC. (www.alnap.org/resource/2999.aspx).

Cosgrave, J., Wata, H., Ntata, P., Immajati, Y. And Bhatt, M. (2010) *Programme evaluation of disaster risk reduction commissioned by Cordaid overall report*. London: Channel Research. (www.alnap.org/resource/12523.aspx).

Cossée, O., Belli, L., Bultemeier, B. and Carrugi, C. (2010) *Evaluation of FAO interventions funded by the CERF: Final report*. Rome: FAO. (www.alnap.org/resource/8099.aspx).

Court, J. and Young, J. (2003) *Bridging research and policy: Insights from 50 case studies*. London: ODI. (www.alnap.org/resource/8100.aspx).

CP MERG. (2012) *Ethical principles, dilemmas and risks in collecting data on violence against children: A review of available literature*. CP MERG. (www.alnap.org/resource/22327.aspx).

Crawford, N., Cosgrave, J., Haysom, S., Walicki, N. (2015) *Protracted displacement: uncertain paths to self-reliance in exile*. London: HPG/ODI. (www.alnap.org/resource/21304.aspx).

Crawford, P., Bysouth, K., Nichols, D. and Thompon, F. (2006) *CAER cluster evaluation: Pakistan earthquake*. Canberra: AusAID. (www.alnap.org/resource/8190.aspx).

Crawford, P. and Eagles, J. (2007) *ANCP Pacific cluster evaluation report*. Version 2.0. Canberra: AusAID. (www.alnap.org/resource/8191.aspx).

Creswell, J.W. (2002) *Educational research: planning, conducting, and evaluating quantitative and qualitative research*. 4th ed. Upper Saddle River: Pearson. (www.alnap.org/resource/22896.aspx).

Creswell, J. W. and Plano Clark, V. L. (2011) *Designing and conducting mixed methods research*. 2nd ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22897.aspx).

Crewe, E. and Young, J. (2002) *Bridging research and policy: Context, evidence and links*. London: ODI. (www.alnap.org/resource/22898.aspx).

Crisp, J., Garras, G., McAvoy, J., Schenkenberg, E., Speigel, P. and Voon, F. (2013) *From slow boil to breaking point: A real-time evaluation of UNHCR's response to the Syrian refugee emergency*. Geneva: UNHCR. (www.alnap.org/resource/8848.aspx).

Curtis, V. and Kanki, B. (1999) *A manual on hygiene promotion*. New York: UNICEF. (www.alnap.org/resource/22899.aspx).

Dabelstein, N. (1996) 'Evaluating the international humanitarian system: rationale, process and management of the joint evaluation of the international response to the Rwanda genocide. *Disasters*, 20(4). (www.alnap.org/resource/8101.aspx).

Dahlgren, A.-L., DeRoo, L. A., Avril, J. and Loutan, L. (2009) 'Health risks and risk-taking behaviors among International Committee of the Red Cross (ICRC) expatriates returning from humanitarian missions. *Journal of Travel Medicine*, 16(6): 382–390. (www.alnap.org/resource/22900.aspx).

Daito, H., Suzuki, M., Shiihara, J., Kilgore, P. E., Ohtomo, H., Morimoto, K., Ishida, M., Kamigaki, T., Oshitani, H., Hashizume, M., Endo, W., Hagiwara, K., Ariyoshi, K. and Okinaga, S. (2013) 'Impact of the Tohoku earthquake and tsunami on pneumonia hospitalisations and mortality among adults in northern Miyagi, Japan: a multicentre observational study'. *Thorax*, 68(6): 544–550. (www.alnap.org/resource/22901.aspx).

DANIDA. (2012) *Evaluation of Danish development support to Afghanistan*. Copenhagen: DANIDA. (www.alnap.org/resource/22903.aspx).

Daniel, J. (2013) *Sampling essentials: Practical guidelines for making sampling choices*. Los Angeles, CA: SAGE. (www.alnap.org/resource/9161.aspx).

Das, N. C. and Shams, R. (2011) *Asset transfer programme for the ultra poor: A randomized control trial evaluation*. Dhaka: BRAC. (www.alnap.org/resource/22904.aspx).

Davidson, E. J. (2007) 'Unlearning some of our social scientist habits'. *Journal of MultiDisciplinary Evaluation*, 4(8): iii–vi. (www.alnap.org/resource/22906.aspx).

Davidson, E. J. (2009) *Causal inference: Nuts and bolts: A mini workshop for the anzea Wellington branch*. Wellington: Davidson Consulting.
(www.alnap.org/resource/22908.aspx).

Davidson, E. J. (2010a) '9 golden rules for commissioning a waste-of-money evaluation', in in Rogers, P. J. and Davidson, E. J. (eds) *Genuine Evaluation*.
(www.alnap.org/resource/22909.aspx).

Davidson, E. J. (2010b) 'Commissioning XGEMs – the sequel', in Rogers, P. J. and Davidson, E. J. (eds) *Genuine Evaluation*. (www.alnap.org/resource/22910.aspx).

Davidson, E. J. (2010c) 'Extreme genuine evaluation makeovers (XGEMs) for commissioning', in Rogers, P. J. and Davidson, E. J. (eds) *Genuine Evaluation*.
(www.alnap.org/resource/22912.aspx).

Davidson, E. J. (2012a) '6 questions that cut to the chase when choosing the right evaluation team', in Rogers, P. J. and Davidson, E. J. (eds) *Genuine Evaluation*.
(www.alnap.org/resource/22905.aspx).

Davidson, E. J. (2012b) '9 hot tips for commissioning, managing (and doing!) actionable evaluation' in Rogers, P. J. and Davidson, E. J. (eds) *Genuine Evaluation*.
(www.alnap.org/resource/22913.aspx).

Davidson, E. J. (2014a) *Evaluative Reasoning*. Florence: UNICEF.
(www.alnap.org/resource/22331.aspx).

Davidson, E. J. (2014b) "Minirubrics" – 7 hot tips for using this cool tool to focus evaluative conversations', in Rogers, P. J. and Davidson, E. J. (eds) *Genuine Evaluation*. (www.alnap.org/resource/22914.aspx).

Davies, R. and Dart, J. (2005) *The "most significant change" (MSC) technique: A guide to its use*. Atlanta, GA: CARE International.
(www.alnap.org/resource/8102.aspx).

Davin, E., Gonzalez, V. and Majidi, N. (2009) *UNHCR's voluntary repatriation program: Evaluation of the impact of the cash grant*. Geneva: UNHCR.
(www.alnap.org/resource/22915.aspx).

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H. and Resnicow, K. (2010) 'Interviewer effects in public health surveys'. *Health Education Research*, 25(1): 14–26. (www.alnap.org/resource/8103.aspx).

Day, S. J. and Altman, D. G. (2000) 'Blinding in clinical trials and other studies'. *BMJ*, 321(8): 504. (www.alnap.org/resource/22917.aspx).

DEC. (2013) *DEC Approach to Learning*. London: DEC. (www.alnap.org/resource/8649.aspx).

de Nicola, F. and Giné, X. (2011) *How accurate are recall data? Evidence from coastal India*. Washington, DC: World Bank. (www.alnap.org/resource/8104.aspx).

Denzin, N. K. (2010) 'Moments, mixed methods, and paradigm dialogs. *Qualitative Inquiry*, 16(6): 419–427. (www.alnap.org/resource/8105.aspx).

de Ville de Goyet, C. and Morinière, L. (2006) *The role of needs assessment in the tsunami response*. London: TEC. (www.alnap.org/resource/12406.aspx).

de Ville de Goyet, C., Scheuermann, P., Reed, S. B., Al Wehaidy, R., and Termil, A. (2011) *SDC humanitarian aid: Emergency relief*. Bern: Swiss Agency for Development and Cooperation. (www.alnap.org/resource/6142.aspx).

de Ville de Goyet, C., Buckle, P., levers, J., Parlan, H. P. and Viandrito, J. (2013) *Evaluation of the European Commission's humanitarian and disaster risk reduction activities (DIPECHO) in Indonesia*. Brussels: ECHO. (www.alnap.org/resource/11739.aspx).

DFID. (2003) *Tools for development: A handbook for those engaged in development activity*. London: DFID. (www.alnap.org/resource/11859.aspx).

DFID, (2005) *Guidance and evaluation and review for DFID staff*. London: DFID. (www.alnap.org/resource/22918.aspx).

DFID. (2010) *Evaluation study Terms of Reference (template)*. London: DFID. (www.alnap.org/resource/22920.aspx).

DFID. (2014) *Assessing the strength of evidence*. London: DFID. (www.alnap.org/resource/10949.aspx).

DFID and UKAID. (2011) *Multilateral aid review: ensuring maximum value for money for UK aid through multilateral organisations*. London: DFID and UKAID. (www.alnap.org/resource/8106.aspx).

- Dillman, D. A., Smyth, J. D. and Christian, L. M. (2009)** *Internet, mail and mixed-mode surveys: The tailored design method*. 3rd ed. Hoboken, NJ: Wiley. (www.alnap.org/resource/8107.aspx).
- Disch, A., Haarberg, K., Ghartey, A. B. and Lunøe, B. (2008)** *Synthesis study on best practices and innovative approaches to capacity development in low-income African countries*. Oslo: NORAD. (www.alnap.org/resource/10132.aspx).
- Dohrenwend, B. S., Colombotos, J. and Dohrenwend, B. P. (1968)** 'Social distance and interviewer effects'. *Public Opinion Quarterly*, 32(3): 410–422. (www.alnap.org/resource/8109.aspx).
- Drucker, P. F. (1985)** *Innovation and entrepreneurship*. Oxford: Butterworth-Heinemann. (www.alnap.org/resource/22921.aspx).
- Dryden-Peterson, S. (2011)** *Refugee education: A global review*. Geneva: UNHCR. (www.alnap.org/resource/8110.aspx).
- Duncalf, J. (2013)** *A real-time evaluation of ACF International's response to typhoon Haiyan / Yolanda in the Philippines*. Paris: ACF. (www.alnap.org/resource/20994.aspx).
- du Preez, M., Conroy, R. M., Wright, J. A., Moyo, S., Potgieter, N. and Gundry, S. W. (2008)** 'Use of ceramic water filtration in the prevention of diarrheal disease: A randomized controlled trial in rural South Africa and Zimbabwe'. *The American Journal of Tropical Medicine and Hygiene*, 79(5): 696–701. (www.alnap.org/resource/22922.aspx).
- Durkin, M. S., Khan, N., Davidson, L. L., Zaman, S. S. and Stein, Z. A. (1993)** 'The effects of a natural disaster on child behavior: evidence for posttraumatic stress'. *American Journal of Public Health*, 83(11): 1549–1553. (www.alnap.org/resource/12530.aspx).
- Ebbinghaus, B. (2005)** 'When less is more: Selection problems in large- N and small- N cross-national comparisons'. *International Sociology*, 20(2): 133–152. (www.alnap.org/resource/22924.aspx).
- ECHO. (2002)** *Evaluating humanitarian action funded by the humanitarian aid office of the European Commission*. Brussels: ECHO. (www.alnap.org/resource/22925.aspx).

Edmondson, A. C. (2004) 'Learning from failure in health care: frequent opportunities, pervasive barriers'. *Quality and Safety in Health Care*, 13(2):ii3–ii9. (www.alnap.org/resource/8111.aspx).

ECB. (2011a) *What we know about joint evaluations of humanitarian action*. London: ECB. (www.alnap.org/resource/19286.aspx).

ECB. (2011b) *What we know about joint evaluations of humanitarian action. Section 1 of 3: the guide*. London: ECB. (www.alnap.org/resource/9536.aspx).

ECB. (2011c) *What we know about joint evaluations of humanitarian action. Section 2 of 3: the stories*. London: ECB. (www.alnap.org/resource/9537.aspx).

ECB. (2011d) *What we know about joint evaluations of humanitarian action. Section 3 of 3: the tools*. London: ECB. (www.alnap.org/resource/9538.aspx).

ECB. (2012) *Joint evaluations. Guidelines from the ECB project*. London: ECB. (www.alnap.org/resource/19413.aspx).

ECB. (2013) *What we know about collaboration: the ECB country consortium experience*. London: ECB. (www.alnap.org/resource/10552.aspx).

EES. (2007) *EES statement: The importance of a methodologically diverse approach to impact evaluation - specifically with respect to development aid and development interventions*. Prague: EES. (www.alnap.org/resource/11119.aspx).

EES. (2015) *The EES capabilities framework*. Prague: EES. (www.alnap.org/resource/22927.aspx).

EPOC. (2013) *Interrupted time series (ITS) analyses*. Oslo: Norwegian Knowledge Centre for the Health Services. (www.alnap.org/resource/22928.aspx).

Egeland, J., Harmer, A. and Stoddard, A. (2011) *To stay and deliver: Good practice for humanitarians in complex security environments*. New York: OCHA. (www.alnap.org/resource/6364.aspx).

Eisinga, R., te Grotenhuis, M., Larsen, J. K., Pelzer, B. and van Strien, T. (2011) BMI of interviewer effects. *International Journal of Public Opinion Research*, 23(4): 530–543. (www.alnap.org/resource/8112.aspx).

FDFA. (2014) *Humanitarian access in situations of armed conflict*. Bern: FDFA. (www.alnap.org/resource/21733.aspx).

Fearon, J., Humphreys, M. and Weinstein, J. (2008) *Community driven reconstruction in Lofa County: Impact assessment*. New York: IRC.
(www.alnap.org/resource/8192.aspx).

Featherstone, A. (2012) *Evaluation of Somalia drought response 2011/2012: Using Oxfam GB's global humanitarian indicator tool*. Oxford: Oxfam.
(www.alnap.org/resource/7983.aspx).

Featherstone, A. (2013a) *Improving impact: do accountability mechanisms deliver results?* Geneva: HAP. (www.alnap.org/resource/8388.aspx).

Featherstone, A. (2013b) *Syria needs analysis project (SNAP): External mid-term review*. Geneva: ACAPS. (www.alnap.org/resource/10390.aspx).

Featherstone, A. (2014) *A review of UNHCR's utilisation of the Central Emergency Response Fund*. Geneva: UNHCR. (www.alnap.org/resource/12430.aspx).

Ferf, A. and Fabbri, P. (2014) *Impact evaluation of Swiss solidarity Asian tsunami programme*. Geneva: Swiss Solidarity. (www.alnap.org/resource/22473.aspx).

Fergusson, D., Cranley Glass, K., Waring, D., Shapiro, S. (2004) 'Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials'. *BMJ*. (www.alnap.org/resource/22930.aspx).

Few, R., McAvoy, D., Tarazona, M. and Walden, V. (2014) *Contribution to Change: An approach to evaluating the role of intervention in disaster recovery*. Oxford: Oxfam. (www.alnap.org/resource/10305.aspx).

Flaherty, E. W., Barry, E. and Swift, M. (1978) 'Use of an unobtrusive measure for the evaluation of interagency coordination. *Evaluation Review*, 2(2): 261–273. (www.alnap.org/resource/22931.aspx).

Flores-Macias, F. and Lawson, C. (2008) 'Effects of interviewer gender on survey responses: Findings from a household survey in Mexico'. *International Journal of Public Opinion Research*, 20(1): 100–110. (www.alnap.org/resource/8113.aspx).

Foley, P. (2009) *Participatory evaluation of the 2008 farmer field school programme, Lira, Uganda*. Kampala: ACF. (www.alnap.org/resource/20908.aspx).

Fortune, V. and Rasal, P. (2010) *British Red Cross - mass sanitation module - 2010 Haiti earthquake response - post deployment learning evaluation*. London: British Red Cross. (www.alnap.org/resource/6038.aspx).

Foubert, V. and Eskandar, H.(2009) *Taking the initiative – Exploring quality and accountability in the humanitarian sector: an introduction to eight initiatives*. Geneva: Sphere Project. (www.alnap.org/resource/5673.aspx).

Fowler, F. J. (2009) *Survey research methods*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/8115.aspx).

Fritz Institute. (2005) *Recipient perceptions of aid effectiveness: rescue, relief and rehabilitation in tsunami affected Indonesia, India and Sri Lanka*. San Francisco: CA: Fritz Institute. (www.alnap.org/resource/5459.aspx).

Fritz Institute. (2006) *Hurricane Katrina: Perceptions of the affected*. San Francisco, CA: Fritz Institute. (www.alnap.org/resource/8116.aspx).

Frangillo, E. A. (2012) *Analysis of complex surveys*. Cornell: Cornell Statistical Consulting Unit. (www.alnap.org/resource/22932.aspx).

Funnell, S. C. and Rogers, P. (2011) *Purposeful program theory: Effective use of theories of change and logic models*. Hoboken, NJ: Wiley. (www.alnap.org/resource/8193.aspx).

Gamarra, T., Reed, S. B. and Wilding, J. (2005) *Final evaluation of hurricanes operation 2004*. Geneva: IFRC. (www.alnap.org/resource/12536.aspx).

Garvin, D. A., Edmondson, A. C. and Gino, F. (2008) 'Is yours a learning organization?' *Harvard Business Review*, 86(3). (www.alnap.org/resource/8117.aspx).

Gerring, J. (2006) 'Techniques for choosing cases', in *Case study research: principles and practices*. New York: Cambridge University Press. (www.alnap.org/resource/9014.aspx).

Gilmour, S., Degenhardt, L., Hall, W. and Day, C. (2006) 'Using intervention time series analyses to assess the effects of imperfectly identifiable natural events: a general method and example'. *BMC Medical Research Methodology*, 6(1): 1–9. (www.alnap.org/resource/22933.aspx).

Given, L. M. (2008) *The SAGE encyclopedia of qualitative research methods*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22934.aspx).

Glass, G. V. (1997) *Interrupted time series quasi-experiments*. Tempe, AZ: Arizona State University. (www.alnap.org/resource/22935.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).

Godden, B. (2004) *Sample size formulas.* Chicago: Appraisal institute. (www.alnap.org/resource/22936.aspx).

Goertz, G. (2012) *A tale of two cultures: Qualitative and quantitative research in the social sciences.* Princeton: Princeton University Press. (www.alnap.org/resource/13029.aspx).

Goertz, G. and Mahoney, J. (2012) 'Methodological Rorschach tests: Contrasting interpretations in qualitative and quantitative research. *Comparative Political Studies*, 46(2): 236–251. (www.alnap.org/resource/22937.aspx).

GHD. (2003a.) 'Good practices in donor financing, management and accountability', in *23 principles and good practice of humanitarian donorship*. (www.alnap.org/resource/10184.aspx).

GHD. (2003b) *International meeting on good humanitarian donorship.* Stockholm: GHD. (www.alnap.org/resource/22940.aspx).

Goodrick, D. (2014) *Comparative case studies.* Florence: UNICEF. (www.alnap.org/resource/22941.aspx).

Gover, A. R., Macdonald, J. M. and Alpert, G. P. (2003) 'Combating domestic violence: Findings from an evaluation of a local domestic violence court'. *Criminology & Public Policy*, 3(1): 109–132. (www.alnap.org/resource/22942.aspx).

Goyder, H. (2010) *Evaluation of DEC-funded shelter projects following the 2009 Indonesian earthquake.* Teddington: Tearfund. (www.alnap.org/resource/5958.aspx).

Graig, E. (2011) *Hiring an evaluation consultant.* Riverdale: Usable Knowledge. (www.alnap.org/resource/22944.aspx).

Graig, E. (2013) 'An alternative to the evaluation RFP process', in Usable Knowledge. (www.alnap.org/resource/22943.aspx).

Grais, R. F., Luquero, F. J., Grellety, E., Pham, H., Coghlan, B. and Salignon, P. (2009) 'Learning lessons from field surveys in humanitarian contexts: a case study of field surveys conducted in North Kivu, DRC 2006-2008'. *Conflict and Health*, 3(8). (www.alnap.org/resource/22946.aspx).

Grasso, P. G. Imas, L. M. and Fostvedt, N. (2013) *Use of evaluation in the Norwegian development cooperation system*. Oslo: Norad. (www.alnap.org/resource/22947.aspx).

Gray, D.E. (2014) *Doing research in the real world*. 3rd ed. Los Angeles: SAGE. (www.alnap.org/resource/22948.aspx).

Green, D. (2012) *Creating killer facts and graphics*. Oxford: Oxfam. (www.alnap.org/resource/22949.aspx).

Groupe URD. (2009) *Quality Compas Companion Book*. Plaisians: Groupe URD. (www.alnap.org/resource/8541.aspx).

Grünewald, F., Binder, A. and Georges, Y. (2010) *Inter-agency real-time evaluation in Haiti: 3 months after the earthquake*. Plaisians: Groupe URD. (www.alnap.org/resource/9962.aspx).

Grünewald, F., Kauffmann, D., Boyer, B. and Patinet, J. (2011) *Real-time evaluation of humanitarian action supported by DG ECHO in Haiti: 2009-2011: November 2010-April 2011*. Plaisians: Groupe URD. (www.alnap.org/resource/22950.aspx).

Gubbels, P. and Bousquet, C. (2013) *Independent evaluation of CARE's response to the 2011-2012 Sahel humanitarian crisis*. Geneva: CARE International. (www.alnap.org/resource/20768.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).

Guest, G., Bunce, A. and Johnson, L. (2006) How many interviews are enough? *Field methods*, 18(1): 59–82. (www.alnap.org/resource/22952.aspx).

Guijt, I. (2014) *Participatory Approaches*, Florence: UNICEF. (www.alnap.org/resource/22798.aspx).

Hagens, C. and Ishida, L. (2010) *Real time evaluation of CRS's flood response in Pakistan: KPK and Baluchistan*. Baltimore, MD: CRS. (www.alnap.org/resource/5956.aspx).

Hahn, J., Todd, P. and Van der Klaauw, W. (2001) 'Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1): 201–209. (www.alnap.org/resource/22954.aspx).

Hallam, A. (2011) *Harnessing the power of evaluation in humanitarian action: An initiative to improve understanding and use of evaluation*. London: ALNAP. (www.alnap.org/resource/6123.aspx).

Hallam, A. and Bonino, F. (2013) *Using evaluation for a change: Insights from humanitarian practitioners*. London: ALNAP. (www.alnap.org/resource/8980.aspx).

Hammersley, M. (1992) *What's wrong with ethnography?* London: Routledge. (www.alnap.org/resource/9017.aspx).

Hanley, T., Binas, R., Murray, J. and Tribunalo, B. (2014) *IASC Inter-agency humanitarian evaluation of the typhoon Haiyan response*. Geneva: IASC. (www.alnap.org/resource/19318.aspx).

HAP. (2010) *The 2010 HAP standard in accountability and quality management*. Geneva: HAP. (www.alnap.org/resource/8125.aspx).

Hart, C. (1998) *Doing a literature review: releasing the social science research imagination*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/8126.aspx).

Hassan, P., Freese, A. and Guallar, M. (2013) *Desk review of UNDAFs commencing in 2013*. New York: UNDP. (www.alnap.org/resource/22811.aspx).

Haver, K., Hatungimana, F. and Tennant, V. (2009) *Money matters: An evaluation of the use of cash grants in UNHCR's voluntary repatriation and reintegration programme in Burundi*. Geneva: UNHCR. (www.alnap.org/resource/5710.aspx).

HelpAge International. (2010) *A study of humanitarian funding for older people*. London: HelpAge International. (www.alnap.org/resource/8127.aspx).

Hidalgo, S., LaGuardia, D., Trudi, G., Sole, R., Moussa, Z., van Dijk, J., Merckx, P. and Zimmer, L. (2015) *Beyond humanitarian assistance? UNHCR and the response to Syrian refugees in Jordan and Lebanon: January 2013 – April 2014*. Brussels: Transtec. (www.alnap.org/resource/20222.aspx).

Hidalgo, S. and Théodate, M. P. (2011) *Inter-agency real-time evaluation of the humanitarian response to the earthquake in Haiti: 20 months after*. Geneva: IASC. (www.alnap.org/resource/6330.aspx).

Hill, M. E. (2002) 'Race of the interviewer and perception of skin color: Evidence from the multi-city study of urban inequality'. *American Sociological Review*, 67(1): 99–108. (www.alnap.org/resource/8123.aspx).

Horton, K. and Roche, C. (eds) (2010) *Ethical questions and international NGOs: An exchange between philosophers and NGOs*. Houten: Springer. (www.alnap.org/resource/22955.aspx).

Hróbjartsson, A., Forfang, E., Haahr, M. T., Als-Nielsen, B. and Brorson, S. (2007) 'Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding'. *International Journal of Epidemiology*, 36(3): 654–663. (www.alnap.org/resource/22956.aspx).

Humanitarian Initiatives UK. (2001) *Independent evaluation: The DEC response to the earthquake in Gujarat January - October 2001*. London: Humanitarian Initiatives UK. (www.alnap.org/resource/3432.aspx).

Hund, L., Bedrick, E. J. and Pagano, M. (2015) 'Choosing a cluster sampling design for lot quality assurance sampling surveys. *PLoS ONE*, 10(6). (www.alnap.org/resource/22958.aspx).

Hunt, D. (2014) *Photo Story: Risk Analysis in Tcharow, Goz Beida, Sila Region, Chad*. Concern Worldwide.

Hyatt, S. and Auten, S. (2011) *Evaluation of the Palestinian youth empowerment program (RUWWAD)*. Washington, DC: USAID. (www.alnap.org/resource/22959.aspx).

IASC. (2010) *Gender marker: How-to-code tip sheet*. Geneva: IASC. (www.alnap.org/resource/8128.aspx).

IASC. (2010b) *Guidance note for clusters to implement the IASC gender marker: Creating gender-responsive projects and tracking gender-related allocations in humanitarian appeals and funding mechanisms*. Geneva: IASC. (www.alnap.org/resource/8129).

IASC. (2010c) *IASC gender marker fact sheet*. Geneva: IASC. (www.alnap.org/resource/22800.aspx).

IASC. (2010d) *Inter-agency real-time evaluation (IA RTE) of the humanitarian response to [disaster XYZ In country XYZ]: Terms of Reference*. Geneva: IASC. (www.alnap.org/resource/22960.aspx).

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

IASC. (2013) *The centrality of protection in humanitarian action*. Geneva: IASC. (www.alnap.org/resource/19093.aspx).

ICAI. (2013) *DFID's support for Palestine refugees through UNRWA*. London: ICAI. (www.alnap.org/resource/10111.aspx).

ICRC. (1977) *Protocol additional to the Geneva conventions of 12 August 1949, and relating to the protection of victims of non-international armed conflicts (Protocol II)*. Geneva: ICRC. (www.alnap.org/resource/8122.aspx).

IFAD. (2011) *Evaluating the impact of participatory mapping activities*. Rome: IFAD. (www.alnap.org/resource/22962.aspx).

IFAD. (2014) *Calculating the sample size*. Rome: IFAD. (www.alnap.org/resource/22961.aspx).

IFRC (n.d.) *Types of disasters*. Geneva: IFRC. (www.alnap.org/resource/22963.aspx).

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

IFRC. (2010) *World disasters report 2010: Focus on urban risk*. Geneva: IFRC. (www.alnap.org/resource/9272.aspx).

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

IFRC and ICRC. (1994) *Code of conduct for the International Red Cross and Red Crescent movement and non-governmental organizations (NGOs) in disaster relief*. Geneva: IFRC/ICRC. (www.alnap.org/resource/19291.aspx).

3ie. (2008a) *3ie impact evaluation practice: a guide for grantees*. New Delhi: 3ie (www.alnap.org/resource/8130.aspx).

3ie. (2008b) *Principles for impact evaluation*. New Delhi: 3ie. (www.alnap.org/resource/8131.aspx).

ILO. (2014a) *Checklist 5 preparing the evaluation report*. Geneva: ILO. (www.alnap.org/resource/22964.aspx).

ILO. (2014b) *Checklist 6 rating the quality of evaluation*. Geneva: ILO. (www.alnap.org/resource/22965.aspx).

Imbens, G. W. and Lemieux, T. (2007) 'Regression discontinuity designs: A guide to practice'. *Journal of Econometrics*, 142(2): 615–635. (www.alnap.org/resource/22966.aspx).

INEE. (2004) *Minimum standards for education in emergencies, chronic crises and early reconstruction*. New York: INEE. (www.alnap.org/resource/7860.aspx).

Innovation Network. (2005) *Logic model workbook*. Washington, DC: Innovation Network. (www.alnap.org/resource/22967.aspx).

Israel, G. D. (2009) *Determining sample size*. Gainesville: IFAS University of Florida. (www.alnap.org/resource/22968.aspx).

Jadad, A. R., Moore, A., Carroll, D., Jenkinson, C., Reynolds, J. M., Gavaghan, D. J. and McQuay, H. J. (1996) 'Assessing the quality of reports of randomized clinical trials: is blinding necessary?' *Controlled Clinical Trials*, 17(1): 1–12. (www.alnap.org/resource/22969.aspx).

Jalan, J. and Ravallion, M. (2001) 'Does piped water reduce diarrhea for children in rural India?' *Journal of Econometrics*, 112(1): 153-173. (www.alnap.org/resource/22970.aspx).

Jaspars, S. and Young, H. (1995) *General food distribution in emergencies*, London: ODI. (www.alnap.org/resource/2725.aspx).

Johnson, A. and Head, D. (2011) *CRS Haiti real time evaluation of the 2010 earthquake response : findings, recommendations, and suggested follow up*. Baltimore, MD: CRS. (www.alnap.org/resource/6074.aspx).

Johri, M., Ridde, V., Heinmüller, R. and Haddad, S. (2014) *Estimation of maternal and child mortality one year after user-fee elimination: an impact evaluation and modelling study in Burkina Faso*. Geneva: WHO. (www.alnap.org/resource/22971.aspx).

Jones, N., Jones, H., Steer, L. and Datta, A. (2009) *Improving impact evaluation production and use*. London: ODI. (www.alnap.org/resource/8225.aspx).

Journal of Peacebuilding & Development. (2013) 'Evaluation in violently divided societies: Politics, ethics and methods'. *Journal of Peacebuilding & Development*, 8(2): 1–4. (www.alnap.org/resource/23078.aspx).

Kahneman, D., Knetsch, J. L. and Thaler, R. H. (1991) 'The endowment effect, loss aversion, and status quo bias'. *The Journal of Economic Perspectives*, 5(1): 193–206. (www.alnap.org/resource/22972.aspx).

Kaiser, R., Woodruff, B. A., Bilukha, O., Spiegel, P. B. and Salama, P. (2006) 'Using design effects from previous cluster surveys to guide sample size calculation in emergency settings'. *Disasters*, 30(2): 199–211. (www.alnap.org/resource/22973.aspx).

Kalantari, M., Yule, W., Dyregrov, A., Neshatdoost, H. and Ahmadi, S. J. (2012) 'Efficacy of writing for recovery on traumatic grief symptoms of Afghani refugee bereaved adolescents: A randomized control trial. *OMEGA - Journal of Death and Dying*, 65(2): 139–150. (www.alnap.org/resource/22974.aspx).

Kane, E. W. and Macaulay, L. J. (1993) 'Interviewer gender and gender attitudes'. *Public Opinion Quarterly*, 57(1): 1–28. (www.alnap.org/resource/8132.aspx).

Kaplan, A. M. and Haenlein, M. (2010) 'Users of the world, unite! The challenges and opportunities of social media'. *Business Horizons*, 53(1): 59–68. (www.alnap.org/resource/22975.aspx).

Kennedy, M. (2006) *Interview Probes*. (www.alnap.org/resource/22976.aspx).

Kevlihan, R. (2010) Productivity and cash-for-work in Niger: GOAL's experience. *Humanitarian Exchange* 48: 29-30. London: HPN/ODI. (www.alnap.org/resource/8135.aspx).

Khandker, S. R., Koolwal, G. B. and Samad, H. A. (2010) *Handbook on impact evaluation: quantitative methods and practices*. Washington, DC: World Bank. (www.alnap.org/resource/22977.aspx).

Khoo, S.-E. (2010) 'Health and humanitarian migrants' economic participation. *Journal of Immigrant and Minority Health*, 12(3): 327–339. (www.alnap.org/resource/22978.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).

Kirkby, J., Saeed, A. and Zogopoulos, A. (2006) *Independent evaluation of CARE international's earthquake response in Northern Pakistan*. Atlanta, GA: CARE International. (www.alnap.org/resource/3489.aspx).

Knox Clarke, P. and Darcy, J. (2014) *Insufficient evidence? The quality and use of evidence in humanitarian action*. ALNAP Study. London: ALNAP/ODI. (www.alnap.org/resource/10441.aspx).

Koenig, T. (2004) CAQDAS - A primer. (www.alnap.org/resource/22979.aspx).

Kohrt, B. A., Burkey, M., Stuart, E., A. and Koirala, S. (2015) 'Alternative approaches for studying humanitarian interventions: propensity score methods to evaluate reintegration packages impact on depression, PTSD, and function impairment among child soldiers in Nepal'. *Global Mental Health*. (www.alnap.org/resource/22980.aspx).

Kosko, J., Klassen, T. P., Bishop, T. and Hartling, L. (2006) 'Evidence-based medicine and the anecdote: Uneasy bedfellows or ideal couple?' *Paediatrics & Child Health*, 11(10): 665–668. (www.alnap.org/resource/22981.aspx).

Krippendorff, K. (2004) *Content analysis: an introduction to its methodology*. 2nd ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/8223.aspx).

Krueger, R. A. and Casey, M. A. (2009) *Focus groups: A practical guide for applied research*. Thousand Oaks, CA: Sage. (www.alnap.org/resource/8136.aspx).

Krueger, S. and Sagmeister, E. (2014) *Real-time evaluation of humanitarian assistance revisited: Lessons learned and the way forward*. Berlin: GPPi. (www.alnap.org/resource/12886.aspx).

Kumar, S. (2002) *Methods for community participation: a complete guide for practitioners*. Rugby: ITDG Publishing. (www.alnap.org/resource/8137.aspx).

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

Ky Luu, J. D., Nkwake, A. M., Morris, M. F., Hansch, S., Gago, M., Corzine, E. and Marcello, J. (2014) *A decade of learning: Lessons from an evaluation of the emergency capacity building (ECB) project*. New Orleans, LA: DRG. (www.alnap.org/resource/12981.aspx).

LaGuardia, D. and Van den Toorn, W. (2011) *Evaluation of UNRWA's organizational development (OD)*. Amman: UNRWA. (www.alnap.org/resource/22986.aspx).

Lanphear, B. P., Vorhees, C. V. and Bellinger, D. C. (2005)
'Protecting children from environmental toxins'. *PLoS Medicine*, 2(3): 203-208. (www.alnap.org/resource/22987.aspx).

Larcom, S., Rauch, F. and Willems, T. (2015) *The benefits of forced experimentation: Striking evidence from the London Underground network*. Oxford: University of Oxford. (www.alnap.org/resource/22988.aspx).

Lautze, S. and Raven-Roberts, A. (2003) *Food security in complex emergencies: building policy frameworks to address longer-term programming challenges*. Rome: FAO. (www.alnap.org/resource/7785.aspx).

Lawday, A. (2014) *Evaluation of HelpAge International's programme "assistance to specific vulnerable groups affected by the Syrian crisis"* (2013-2014). London: Andrew Lawday Evaluations. (www.alnap.org/resource/12857.aspx).

Lawday, A. (2015) *Inter-agency humanitarian evaluation (IAHE) of the response to the Central African Republic's crisis 2013-2015*. Washington: Konterra Group. (www.alnap.org/resource/22989.aspx).

Laybourn, C. and Obrecht, A. (2014) *DEC accountability self-assessment validation 2013/14*. London: One World Trust. (www.alnap.org/resource/22990.aspx).

Leader, N. (2000) *The politics of principle: the principles of humanitarian action in practice*. London: ODI/HPG. (www.alnap.org/resource/8139.aspx).

Lee, R. M. (2000) *Unobtrusive methods in social research*. Buckingham: Open University Press. (www.alnap.org/resource/22991.aspx).

Lemeshow, S. and Taber, S. (1991) 'Lot quality assurance sampling: single-and double-sampling plans'. *World Health Stat Q*, 44(3): 115–132. (www.alnap.org/resource/22992.aspx).

Leturque, H., Beaujeu, R., Majeed, Y. and Saleem, S. (2012)
Evaluation of the CBHA early recovery programme in Pakistan. London: CBHA.
(www.alnap.org/resource/6370.aspx).

Lieberson, S. (1991) 'Small N's and big conclusions: An examination of the reasoning in comparative studies based on a small number of cases. *Social Forces*, 70(2): 307–320. (www.alnap.org/resource/22994.aspx).

Lijmer, J. G., Mol, B. W., Heisterkamp, S., Bonsel, G. J., Prins, M. H. van der Meulen, J. H. P. and Bossuyt, P. M. M. (1999) 'Empirical evidence of design-related bias in studies of diagnostic tests'. *JAMA*, 282(11): 1061–1066. (www.alnap.org/resource/22995.aspx).

Lindgren, D., Matondang, M. and Putri, D. (2005) *Tsunami relief study: Effectiveness of the tsunami relief effort*. London: TEC. (www.alnap.org/resource/22997.aspx).

Lister, S., Anderson, S., Mokbel Genequand, M., Gordon, A., Sandford, J. and Turner, S. (2011) *WFP's school feeding policy: a policy evaluation. Vol. I full report*. Rome: WFP. (www.alnap.org/resource/22998.aspx).

Longo, M., Canetti, D. and Hite-Rubin, N. (2014) 'A checkpoint effect? Evidence from a natural experiment on travel restrictions in the West Bank'. *American Journal of Political Science*, 58(4): 1006-1023.
(www.alnap.org/resource/23000.aspx).

Lysy, C. (2014) *More data, shorter reports: three ways to respond*. Fresh Spectrum (www.alnap.org/resource/23001.aspx).

MacDonald, B. and Gedeon, H. (2012) *Banking with mobile phone: A report on a T-cash pilot project*, Baltimore, MD: CRS. (www.alnap.org/resource/9882.aspx).

Mackay, K., Clark, M., Sartorius, R., Bamberger, M. (2004)
Monitoring & evaluation: Some tools, methods & approaches. Washington, DC: World Bank. (www.alnap.org/resource/23003.aspx).

Mahoney, J. (2000) 'Strategies of causal inference in small-N analysis'. *Sociological Methods & Research*, 28(4): 387–424.
(www.alnap.org/resource/23004.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).

Majewski, B., Boulet-Desbareau, P., Slezak, M., De Meulder, F. and Wilson, K. (2012) *Joint evaluation of the global logistics cluster*. (www.alnap.org/resource/7904.aspx).

Malki, E. (2008) *A benchmarking model for measuring the efficiency of a humanitarian aid program: a case study of an international NGO*. Munich: MPRA. (www.alnap.org/resource/23006.aspx).

Manesh, A. O., Sheldon, T., A., Pickett, K. E., and Carr-Hill, R. (2008) 'Accuracy of child morbidity data in demographic and health surveys'. *International Journal of Epidemiology*, 37(1). (www.alnap.org/resource/8140.aspx).

Manning, R. (2009) *The quality of DFID's evaluations and evaluation management systems: How do they compare with other agencies?* London: IACDI. (www.alnap.org/resource/23007.aspx).

UN OIOS. (2014) *Inspection and evaluation manual*. New York: UN OIOS. (www.alnap.org/resource/19243.aspx).

Marshall, M. N. (1996) 'Sampling for qualitative research'. *Family Practice*, 13(6): 522–526. (www.alnap.org/resource/23008.aspx).

Martin, E., Petty, C. and Acidri, J. (2009) *Livelihoods in crisis: a longitudinal study in Pader, Uganda*. London: HPG/ODI. (www.alnap.org/resource/23010.aspx).

Martinez, A. (2009) *A participatory programme review of the international federation of the Red Cross and Red Crescent societies response to the Peru earthquake of 15th August 2007*. Geneva: IFRC. (www.alnap.org/resource/8141.aspx).

Martinez, D. E. (2011) *The logical framework approach in non-governmental organizations*. Edmonton: University of Alberta. (www.alnap.org/resource/23011.aspx).

Mays, N. and Pope, C. (2000) 'Assessing quality in qualitative research'. *BMJ*. (www.alnap.org/resource/8142.aspx).

Maystadt, J. F. (2011) *Poverty reduction in a refugee-hosting economy: A natural experiment*. Washington, DC: IFPRI. (www.alnap.org/resource/8143.aspx).

Mazurana, D., Benelli, P., Gupta, H. and Walker, P. (2011) *Sex and age matter: Improving humanitarian response in emergencies*. Atlanta, GA: CARE International. (www.alnap.org/resource/8144.aspx).

McDonald, B. and Rogers, P. (2014) *Interviewing*. Florence: UNICEF. (www.alnap.org/resource/23012.aspx).

McGearty, S., O'Hagan, P. and Montinard, M. (2012) *An independent final evaluation of the action of churches together alliance Haiti appeal HTI-101 (Jan 2010 – Dec 2011)*. London: Channel Research. (www.alnap.org/resource/6339.aspx).

McRAM Team. (2009) *Multi-cluster rapid assessment mechanism (McRAM) in Pakistan*. Islamabad: McRAM. (www.alnap.org/resource/8145.aspx).

Hanif, A. (2008) *HAP initiatives in Concern Worldwide: A case study of Bangladesh* (2008). London: Concern Worldwide. (www.alnap.org/resource/23013.aspx).

Mendoza, U. and Thomas, V. (2009) *UNHCR's AGDM evaluation: A participatory evaluation of AGDM results in four Colombian communities*. Geneva: UNHCR. (www.alnap.org/node/24503.aspx).

Meng, X. and Qian, N. (2009) *The long term consequences of famine on survivors: Evidence from a unique natural experiment using China's great famine*. Cambridge, MA: NBER. (www.alnap.org/resource/23015.aspx).

Meyer, A.-M. (2007) *The causes of malnutrition in children under 3 in the Somali Region of Ethiopia related to household caring practices*. London: Save the Children. (www.alnap.org/resource/3591.aspx).

Mezuk, B., Larkin, G. L., Prescott, M. R., Tracy, M., Vlahov, D., Tardiff, K. and Galea, S. (2009) 'The influence of a major disaster on suicide risk in the population'. *Journal of Traumatic Stress*, 22(6): 481-488. (www.alnap.org/resource/23017.aspx).

Miles, M. B., Huberman, A. M. and Saldaña, J. (2013) *Qualitative data analysis: A methods sourcebook*. 3rd ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/23018.aspx).

Miller, J. and Stoltz, K. (2008) *Moving beyond rhetoric: Consultation and participation with populations displaced by conflict or natural disasters*. Washington, DC: The Brookings Institution. (www.alnap.org/resource/8751.aspx).

- Milojevic, A., Armstrong, B., Hashizume, M., McAllister, K., Faruque, A., Yunus, M., Kim Streatfield, P. Moji, K. and Wilkinson, P. (2012)** 'Health effects of flooding in rural Bangladesh'. *Epidemiology*, 23(1): 107–115. (www.alnap.org/resource/23019.aspx).
- Milojevic, A., Armstrong, B., Kovats, S., Butler, B., Hayes, E., Leonardi, G., Murray, V. and Wilkinson, P. (2011)** 'Long-term effects of flooding on mortality in England and Wales, 1994-2005: controlled interrupted time-series analysis'. *Environmental Health*. (www.alnap.org/resource/23020.aspx).
- Miranda, R. (2009)** *Eva the Evaluator*. Learningham Press. (www.alnap.org/resource/23028.aspx).
- Mirano, S. (2012)** *Learning from the Urban Transitional Shelter Response in Haiti*. Baltimore, MD: CRS. (www.alnap.org/resource/6539.aspx).
- Molund, S. and Schill, G. (2007)** *Looking back, moving forward*. 2nd ed. Stockholm: SIDA. (www.alnap.org/resource/7902.aspx).
- Molyneux, S., Mulupi, S., Mbaabu, L. and Marsh, V. (2012)** 'Benefits and payments for research participants: Experiences and views from a research centre on the Kenyan coast'. *BMC Medical Ethics*, 13(1). (www.alnap.org/resource/8147.aspx).
- Montbiot, E. (2006)** *World Vision: Lessons by Sectors*. Monrovia, CA: World Vision. (www.alnap.org/resource/23029.aspx).
- Morán, J. L. Á. (2012)** *External evaluation: Emergency nutrition programme in Sindh Province, Pakistan*. London: Action Against Hunger. (www.alnap.org/resource/20934.aspx).
- Morel, D. and Hagens, C. (2012)** *Monitoring, evaluation, accountability and learning in emergencies - a resource pack for simple and strong MEAL*. Baltimore, MD: CRS. (www.alnap.org/resource/9200.aspx).
- Morgan, D. L. (1997)** *Focus groups as qualitative research*. 2nd ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/23030.aspx).
- Morinière, L. (2011a)** *Evaluation of the ERF component of the Haiti emergency response (ERF) fund*. New York: OCHA. (www.alnap.org/resource/6054.aspx).
- Morinière, L. (2011b)** *External evaluation of the Haiti emergency relief and response fund (ERRF), 2008-2011*. New York: OCHA. (www.alnap.org/resource/8149.aspx).

Morra-Imas, L .G. and Rist, R. C. (2009) *The road to results: Designing and conducting effective development evaluations*. Washington, DC: World Bank. (www.alnap.org/resource/8470.aspx).

Morris, M. F. and Shaughnessy, D. E. (2007) *Final evaluation report: Emergency capacity building project*. Atlanta, GA: CARE International. (www.alnap.org/resource/10636.aspx).

Morris, T., Steen, N. and Canteli, C. (2013) *Synthesis of mixed method impact evaluations of the contribution of food assistance to durable solutions in protracted refugee situations*. Geneva: UNHCR. (www.alnap.org/resource/19902.aspx).

Moscoe, E., Bor, J. and Bärnighausen, T. (2015) 'Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice'. *Journal of Clinical Epidemiology*, 68(2): 132–143. (www.alnap.org/resource/23031.aspx).

Mowjee, T., Fleming, D. and Toft, E. (2015) *Evaluation of the strategy for Danish humanitarian action 2010-2015*. Copenhagen: DANIDA. (www.alnap.org/resource/20738.aspx).

MSF, 2010. *Summary of findings of emergency response evaluations*. Geneva: MSF. (www.alnap.org/resource/23032.aspx).

MSF. (2012) *A handbook for initiating, managing and conducting evaluations in MSF*. Geneva: MSF. (www.alnap.org/resource/10004.aspx).

Must, A., Phillips, S., Naumova, E., Blum, M., Harris, S., Dawson-Hughes, B. and Rand, W. (2002) 'Recall of early menstrual history and menarcheal body size: After 30 years, how well do women remember?' *American Journal of Epidemiology*, 155(7): 672–679. (www.alnap.org/resource/8150.aspx).

La Pelle, N. R. (2004) 'Simplifying qualitative data analysis using general purpose software tools'. *Field Methods*, 16(1): 85–108. (www.alnap.org/resource/23033.aspx).

Naidu, V. J. (2010) *A thematic evaluation of livelihood and community based disaster preparedness projects*. Atlanta, GA: CARE International. (www.alnap.org/resource/10252.aspx).

NAO. (2000) *A practical guide to sampling*. London: NAO. (www.alnap.org/resource/10228.aspx).

NORAD. (1998) *Results management in Norwegian development cooperation : A practical guide*. Oslo: NORAD. (www.alnap.org/resource/8052.aspx).

NORAD. (1999) *Logical framework approach: handbook for objectives-oriented planning*. Oslo: NORAD. (www.alnap.org/resource/23034.aspx).

Norman, B. (2012) *Monitoring and accountability practices for remotely managed projects implemented in volatile operating environments*. Teddington: Tearfund. (www.alnap.org/resource/7956.aspx).

NRC. (2013) *Evaluation of five humanitarian programmes of the Norwegian Refugee Council (NRC) and of the standby roster NORCAP. Case country report – Somalia*. Oslo: NRC. (www.alnap.org/resource/12227.aspx).

O'Hagan, P., Love, K. and Rouse, A. (2010) *An independent joint evaluation of the Haiti earthquake humanitarian response*. Atlanta, GA: CARE International. (www.alnap.org/resource/5969.aspx).

O'Neil, G. (2012) *7 new ways to present evaluation findings*. 3-5 October, Helsinki. (www.alnap.org/resource/23035.aspx).

Oakden, J. (2013a) *Evaluation rubrics: how to ensure transparent and clear assessment that respects diverse lines of evidence*. Melbourne: Better Evaluation. (www.alnap.org/resource/23036.aspx).

Oakden, J., 2013b. *Guest blog: Why rubrics are useful in evaluations*. Melbourne: Better Evaluation. (www.alnap.org/resource/23037.aspx).

Obrecht, A., Laybourne, C., Hammer, M. and Ray, S. (2012)
The 2011/12 DEC accountability framework assessment: findings from the external evaluation and peer review process. London: One World Trust. (www.alnap.org/resource/23038.aspx).

OCHA. (2004) *Guiding principles on internal displacement*. New York: OCHA. (www.alnap.org/resource/8054.aspx).

OCHA. (2009) 'OCHA on message : Humanitarian access case study'. *OCHA annual report 2008*. New York: OCHA. (www.alnap.org/resource/11252.aspx).

OCHA. (2011) *Inter-agency RTE for the 2011 Pakistan floods*. New York: OCHA. (www.alnap.org/resource/23040.aspx).

OCHA. (2012) *OCHA on message: humanitarian principles*. New York: OCHA. (www.alnap.org/resource/19270.aspx).

OCHA. (2013) *United Nations Disaster Assessment and Coordination UNDAC field handbook*. New York: OCHA. (www.alnap.org/resource/23041.aspx).

ODI. (2013) *Research and policy in development (RAPID)*. London: ODI. (www.alnap.org/resource/23048.aspx).

OECD. (2010) *Managing Joint Evaluations*. Paris: OECD. (www.alnap.org/resource/23039.aspx).

OECD/DAC. (1991) *Principles for evaluation of development assistance*. Paris: OECD/DAC. (www.alnap.org/resource/20830.aspx).

OECD/DAC. (1999) *Guidance for evaluating humanitarian assistance in complex emergencies*. Paris: OECD/DAC. (www.alnap.org/resource/8221.aspx).

OECD/DAC. (2010) *Glossary of key terms in evaluation and results based management*. Paris: OECD/DAC. (www.alnap.org/resource/8489.aspx).

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

OIOS. (2014) *Inspection and evaluation manual*. New York: OIOS. (www.alnap.org/resource/19243.aspx).

Ojha, G. P. (2010) *Appreciative inquiry approach to evaluation practices in South Asia*. Iasi: Lumen Publishing House. (www.alnap.org/resource/23042.aspx).

Oliver, M. L. (2007) *CARE's humanitarian operations: Review of CARE's use of evaluations and after action reviews in decision-making*. Atlanta, GA: CARE International. (www.alnap.org/resource/8058.aspx).

Oliver, M. L. (2009) 'Metaevaluation as a means of examining evaluation influence'. *Journal of MultiDisciplinary Evaluation*, 6(11): 32–37. (www.alnap.org/resource/8285.aspx).

Ong, J. C., Flores, J. M. and Combinido, P. (2015) *Obliged to be grateful*. Woking: Plan International. (www.alnap.org/resource/20633.aspx).

Onwuegbuzie, A. J. and Collins, K. M. T. (2007) 'A typology of mixed methods sampling designs in social science research'. *The Qualitative Report*, 12(2): 281–316. (www.alnap.org/resource/23044.aspx).

Onwuegbuzie, A. J. and Leech, N. (2007) 'A call for qualitative power analyses. *Quality & Quantity*, 41(1): 105–121. (www.alnap.org/resource/23045.aspx).

Ott, E., Krystalli, R. C., Stites, E., Timmins, N., Walden, V., Gillingham, E., Bushby, K. (2015) *Humanitarian evidence synthesis and communication programme: Abridged inception report*. Oxford: Oxfam and Feinstein International Center. (www.alnap.org/resource/23046.aspx).

Otto, R. (2015) *Story in 5: Ralf Otto, Momologue and Ebaix*. [YouTube] London: ALNAP. (www.alnap.org/resource/23047.aspx).

Otto, R., Conrad, C., Brusset, E. and Stone, L. (2006) *Organisational learning review of Caritas Internationalis' response to the Tsunami emergency*. London: Channel Research. (www.alnap.org/resource/8059.aspx).

Oxfam. (2011) *How are effectiveness reviews carried out?* Oxford: Oxfam. (www.alnap.org/resource/8473.aspx).

Oxfam. (2012) *Project effectiveness reviews: summary of 2011/12 findings and lessons learned*. Oxford: Oxfam. (www.alnap.org/resource/8472.aspx).

Page, L., Savage, D. and Torgler, B. (2012) *Variation in risk seeking behavior in a natural experiment on large losses induced by a natural disaster*. Zürich: CREMA. (www.alnap.org/resource/23050.aspx).

Pandya, M., Pathak, V. and Oza, S. (2012) *Future of M&E in humanitarian sector: Possible list for discussion*. Gujarat: AIDMI. (www.alnap.org/resource/22894.aspx).

Pantera, G. (2012) *AGIRE humanitarian response to the East Africa drought*. Rome: AGIRE. (www.alnap.org/resource/6316.aspx).

Patrick, J. (2011) *Haiti earthquake response. Emerging evaluation lessons*. London: ALNAP. (www.alnap.org/resource/6125.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. (2008) *Utilization-focused evaluation*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/10060.aspx).

Patton, M. Q. and Cochran, M. (2002) *A guide to using qualitative research methodology*. Geneva: MSF. (www.alnap.org/resource/13024.aspx).

Pawson, R. and Tilley, N. (2004) *Realist evaluation*. Mount Torrens: Community Matters. (www.alnap.org/resource/8195.aspx).

Peersman, G. (2014a) *Evaluative criteria*. Florence: UNICEF. (www.alnap.org/resource/23051.aspx).

Peersman, G. (2014b) *Overview: Data collection and analysis methods in impact evaluation*. Florence: UNICEF. (www.alnap.org/resource/22654.aspx).

Penfold, R. B. and Zhang, F. (2013) 'Use of interrupted time series analysis in evaluating health care quality improvements'. *Academic Pediatrics*, 13(6): S38–S44. (www.alnap.org/resource/23052.aspx).

People in Aid. (2003) *Code of good practice in the management and support of aid personnel*. London: People in Aid. (www.alnap.org/resource/8061.aspx).

Perrin, B. (2012) Linking Monitoring and Evaluation to Impact Evaluation. Washington, DC: InterAction. (www.alnap.org/resource/19267.aspx).

Petticrew, M., Cummins, S., Ferrell, C., Findlay, A., Higgins, C., Hoy, C., Kearns, A. and Sparks, L. (2005) 'Natural experiments: an underused tool for public health?' *Public Health*, 119(9): 751-757. (www.alnap.org/resource/23053.aspx).

Pfister, R. (2011) 'Wardrobe malfunctions and the measurement of internet behaviour'. *Psychology*, 2(3): 266–268. (www.alnap.org/resource/23054.aspx).

Picciotto, R. (2012) 'Experimentalism and development evaluation: Will the bubble burst?' *Evaluation*, 18(2): 213–229. (www.alnap.org/resource/8062.aspx).

Rawlins, R., Pimkina, S., Barrett, C. B., Pedersen, S. and Wydick, B. (2013) 'Got milk? The impact of Heifer International's livestock donation programs in Rwanda'. *Food Policy*, 44(2): 202–213. (www.alnap.org/resource/23055.aspx).

Polastro, R. (2014) *Evaluating humanitarian action in real time: Recent practices, challenges, and innovations*. Sheffield: IOD Parc. (www.alnap.org/resource/12928.aspx).

Polastro, R., Khalif, M. A., van Ebyen, M., Posada, S., Salah, A., Steen, N. and Toft, E. (2011) *IASC evaluation of the humanitarian response in South Central Somalia 2005-2010*. Madrid: DARA. (www.alnap.org/resource/9301.aspx).

Polastro, R., Nagrah, A., Steen, N. and Zafar, F. (2011) *Inter-agency real time evaluation of the humanitarian response to Pakistan's 2010 Flood Crisis*. Madrid: DARA. (www.alnap.org/resource/6087.aspx).

Poulos, C., Pattanayak, S. K. and Jones, K. (2006) *A guide to water and sanitation sector impact evaluations*. Washington, DC: World Bank. (www.alnap.org/resource/8063.aspx).

Poulsen, L., Stacy, R., Bell, L. and Kumar Range, S. (2009) *Joint thematic evaluation of FAO and WFP support to information systems for food security*. Rome: WFP. (www.alnap.org/resource/5845.aspx).

Preskill, H. S. and Catsambas, T. T. (2006) *Reframing evaluation through appreciative inquiry*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/8224.aspx).

Proudlock, K., Ramalingam, B. and Sandison, P. (2009) 'Improving humanitarian impact assessment: bridging theory and practice', in *ALNAP's 8th review of humanitarian action: Performance, Impact and Innovation*. ALNAP Review. London: ALNAP. (www.alnap.org/resource/5663.aspx).

Puoane, T., Alexander, L. and Hutton, B. (2011) *The revolving door: child malnutrition in Mount Frere, Eastern Cape Province of South Africa*. Cape Town: University of the Western Cape. (www.alnap.org/resource/23056.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).

Qasim, Q., 3le and Stevens, K. (2014) *Regression Discontinuity*. Melbourne: Better Evaluation. (www.alnap.org/resource/23057.aspx).

Rabbani, M., Prakash, V. A. and Sulaiman, M. (2006) *Impact assessment of CFPR/TUP: a descriptive analysis based on 2002-2005 panel data*. Dhaka: BRAC. (www.alnap.org/resource/23058.aspx).

Rademaker, L. L., Grace, E. J. and Curda, S. K. (2012) 'Using computer-assisted qualitative data analysis software (CAQDAS) to re-examine traditionally analyzed data: Expanding our understanding of the data and of ourselves as scholars'. *The Qualitative Report*, 17(22): 1–11. (www.alnap.org/resource/23059.aspx).

Ramalingam, B. (2011) *Learning how to learn: eight lessons for impact evaluations that make a difference*. London: ODI. (www.alnap.org/resource/8458.aspx).

Ramalingam, B. and Mitchell, J. (2014) *Responding to changing needs?* London: ALNAP. (www.alnap.org/resource/19246.aspx).

Raphael, K. (1987) 'Recall bias: A proposal for assessment and control'. *International Journal of Epidemiology*, 16(2): 167–170. (www.alnap.org/resource/23060.aspx).

Raworth, K. (2016) *Reviewing the existing literature*. Oxford: Oxfam. (www.alnap.org/resource/23049.aspx).

Reddit Inc. (2013) *Reddit: the front page of the internet*. (www.alnap.org/resource/23061.aspx).

Reid, E. and Novak, P. (1975) 'Personal space: An unobtrusive measures study'. *Bulletin of the Psychonomic Society*, 5(3): 265–266. (www.alnap.org/resource/23062.aspx).

Reid, S., Stibbe, D. and Lowery, C. (2010) *Review of the global education cluster co-leadership arrangement*. Oxford: The Partnering Initiative. (www.alnap.org/resource/6354.aspx).

Remler, D. K. and Van Ryzin, G. G. (2014) *Research methods in practice: Strategies for description and causation*. 2nd ed. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/23063.aspx).

Ridde, V., Haddad, S. and Heinmüller, R. (2013) 'Improving equity by removing healthcare fees for children in Burkina Faso'. *Journal of Epidemiology and Community Health*. 2013(0): 1-7. (www.alnap.org/resource/23064.aspx).

Riddell, R. C. (2009) *The quality of DFID's evaluation reports and assurance systems: A report for IACDI based on the quality review undertaken by consultants Burt Perrin and Richard Manning*. London: IACDI. (www.alnap.org/resource/8496.aspx).

Riessman, C. K. (1977) *The effect of interviewer status and respondent sex on symptom reporting.* (www.alnap.org/resource/8064.aspx).

Rist, R. C., Boily, M.-H. and Martin, F. (eds) (2011) *Influencing change: building evaluation capacity to strengthen governance.* Washington, DC: World Bank. (www.alnap.org/resource/23065.aspx).

Ritter, T. and Estoesta, J. (2015) *Ethical challenges in the evaluation of a Pacific multi- country training project by an Australian NGO.* Deakin, ACT: ACFID. (www.alnap.org/resource/23066.aspx).

Robert, S. and Valon, A. (2014) *Quality and Accountability for Project Cycle Management.*

Roche, C. (1999) *Impact assessment for development agencies: Learning to value change.* Oxford: Oxfam. (www.alnap.org/resource/8065.aspx).

Roche, C. (2000) *Impact assessment: Seeing the wood and the trees.* Oxford: Oxfam. (www.alnap.org/resource/11135.aspx).

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

Rogers, P. (2013) *52 weeks of BetterEvaluation: Week 11: Using rubrics.* Melbourne: Better Evaluation. (www.alnap.org/resource/23067.aspx).

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

Rogers, P. (2014b) *Overview: Strategies for causal attribution.* Florence: UNICEF. (www.alnap.org/resource/22674.aspx).

Rogers, P. (2014c) *Theory of change.* Florence: UNICEF. (www.alnap.org/resource/22673.aspx).

Rosenberg, L .J., Posner, L. D. and Hanley, E. J. (1970) *Project evaluation and the project appraisal reporting system: Volume one: Summary.* Washington: USAID. (www.alnap.org/resource/23068.aspx).

Rouse, A. (2012) *After action review (AAR) report.* Atlanta, GA: CARE International. (www.alnap.org/resource/8495.aspx).

Ryan, M. E. (2009) 'Making visible the coding process: Using qualitative data software in a post-structural study'. *Issues in Educational Research*, 19(2): 142–161. (www.alnap.org/resource/23069.aspx).

Sackett, D. L. (1979) 'Bias in analytic research'. *Journal of Chronic Diseases*, 32(1–2): 51–63. (www.alnap.org/resource/23070.aspx).

Sackett, D. L. (2007) 'Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't?' *International Journal of Epidemiology*, 36(3): 664–665. (www.alnap.org/resource/23071.aspx).

Salama, P., Assefa, F., Talley, L., Spiegel, P., van der Veen, A. and Gotway, C. A. (2001) 'Malnutrition, measles, mortality, and the humanitarian response during a famine in Ethiopia'. *JAMA*, 286(5): 563–571. (www.alnap.org/resource/23072.aspx).

Sandison, P. (2006) '*The utilisation of evaluations*', in ALNAP review of humanitarian action in 2005: Evaluation utilisation. ALNAP Review. London: ALNAP. (www.alnap.org/resource/5225.aspx).

Sandström, S. and Tchatchua, L. (2010) 'Do cash transfers improve food security in emergencies? Evidence from Sri Lanka', in Omamo, S. W., Gentilini, U. and Sandström, S. (eds), *Revolution: From food aid to food assistance: Innovations in overcoming hunger*. Rome: WFP. (www.alnap.org/resource/8077.aspx).

Saunders, M. (2011) *Evaluation use and usability: theoretical foundations and current state of international thinking*. 8 July, Gdansk. (www.alnap.org/resource/23073.aspx).

Savage, K., Delesques, L., Martin, E. and Ulfat, G. P. (2007) *Corruption perceptions and risks in humanitarian assistance: An Afghanistan case study*. London: HPG/ODI. (www.alnap.org/resource/22801.aspx).

Sawada, Y. and Shimizutani, S. (2008) 'How do people cope with natural disasters? Evidence from the great Hanshin-Awaji (Kobe) earthquake in 1995'. *Journal of Money, Credit and Banking*, 40(2-3): 463–488. (www.alnap.org/resource/23074.aspx).

Schmidt, W.-P., Arnold, B. F., Boisson, S., Genser, B., Luby, S. P., Barreto, M. L., Clasen, T. and Cairncross, S. (2011) 'Epidemiological methods in diarrhoea studies—an update'. *International Journal of Epidemiology*, 40(6): 1678–1692. (www.alnap.org/resource/23075.aspx).

Schulz, K. F., Chalmers, I., Hayes, R. J. and Altman, D. G. (1995) 'Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials'. *JAMA*, 273(5): 408–412. (www.alnap.org/resource/23076.aspx).

Schwartz, L., Sinding, C., Hunt, M., Elit, L., Redwood-Campbell, L., Adelson, N., Luther, L., Ranford, J. and DeLaat, S. (2010) 'Ethics in humanitarian aid work: Learning from the narratives of humanitarian health workers'. *AJOB Primary Research*, 1(3): 45–54. (www.alnap.org/resource/23077.aspx).

Scriven, M. (1991) *Evaluation thesaurus*. Newbury Park, CA: SAGE. (www.alnap.org/resource/8480.aspx).

Scriven, M. (2008) 'A summative evaluation of RCT methodology: and an alternative approach to causal research'. *Journal of MultiDisciplinary Evaluation*, 5(9): 14. (www.alnap.org/resource/8153.aspx).

Scriven, M. (2011) *Evaluating evaluations: a meta-evaluation checklist*. (www.alnap.org/resource/8154.aspx).

Seebregts, C. J., Zwarenstein, M., Mathews, C., Fairall, L., Flisher, A. J., Seebregts, C., Mukoma, W. and Klepp, K. I. (2009) 'Handheld computers for survey and trial data collection in resource-poor settings: Development and evaluation of PDACT, a Palm Pilot interviewing system'. *International journal of medical informatics*. (www.alnap.org/resource/8155.aspx).

Serrat, O. (2009) *The most significant change technique*. Manila: Asian Development Bank. (www.alnap.org/resource/22802.aspx).

Shackman, G. (2001) *Sample size and design effect*, Albany: American Statistical Association. (www.alnap.org/resource/23080.aspx).

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) 'Experiments and generalized causal inference and a critical assessment of our assumptions', in *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company. (www.alnap.org/resource/22889.aspx).

Shah, R. (2014) *Evaluation of the Norwegian Refugee Council's Palestine education programme 2010-2014*. Oslo: NRC. (www.alnap.org/resource/12820.aspx).

Shanks, L., Ariti, C., Siddiqui, M. R., Pintaldi, G., Venis, S. de Jong, K. and Denault, M. (2013) 'Counselling in humanitarian settings: a retrospective analysis of 18 individual-focused non-specialised counselling programmes'. *Conflict and Health*, 7(19). (www.alnap.org/resource/23081.aspx).

Shannon, H. S., Hutson, R., Kolbe, A., Stringer, B., Haines, T., Nelson, N., Yang, L., Reilly, A., Hardin, J. and Hartley, D. (2012) 'Choosing a survey sample when data on the population are limited'. *Emerging Themes in Epidemiology*. (www.alnap.org/resource/8156.aspx).

Shaw, I., Greene, J. C. and Mark, M. M. (2006) *The SAGE handbook of evaluation: policies, programs and practices*. London: SAGE. (www.alnap.org/resource/23082.aspx).

Shirima, K., Mukasa, O., Armstrong Schellenberg, J., Manzi, F., John, D., Mushi, A., Mrisho, M., Tanner, M., Mshinda, H. and Schellenberg, D. (2007) 'The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania'. *Emerging themes in epidemiology 2007*. (www.alnap.org/resource/8157.aspx).

SIDA. (2011) *Strategic Evaluation Plan 2012*. Stockholm: SIDA. (www.alnap.org/resource/23083.aspx).

Sida, L. (2013) *Syria crisis: Evaluators learning exchange on 13 September 2013*. London: ALNAP. (www.alnap.org/resource/9300.aspx).

Sida, L. (2014) *Remote monitoring, evaluation and accountability in the Syria response*. London: ALNAP. (www.alnap.org/resource/12803.aspx).

Sida, L., Gray, B. and Asmare, E. (2012) *IASC real time evaluation (IASC RTE) of the humanitarian response to the Horn of Africa drought crisis – Ethiopia*. Geneva: IASC. (www.alnap.org/resource/7505.aspx).

Silver, C. and Lewins, A. (2014) *Using software in qualitative research: a step by step guide*. 2nd ed. Los Angeles, CA: SAGE. (www.alnap.org/resource/23084.aspx).

Slim, H. and Trombetta, L. (2014) *Syria crisis common context analysis*, New York: OCHA. (www.alnap.org/resource/12718.aspx).

DANIDA. (2015) *Evaluation of the Danish engagement in Palestine*. Copenhagen: DANIDA. (www.alnap.org/resource/23085.aspx).

SLRC. (2012) *Making systematic reviews work for international development*. London: SLRC/ODI. (www.alnap.org/resource/2079.aspx).

Small, M .L. (2011) 'How to conduct a mixed methods study: Recent trends in a rapidly growing literature'. *Annual Review of Sociology*, 37(1): 57–86. (www.alnap.org/resource/23086.aspx).

SMART, 2012. *Sampling methods and sample size calculation for the SMART methodology*. Toronto: SMART. (www.alnap.org/resource/23087.aspx).

Smytlo, T. (2001) *Crouching impact, hidden attribution: Overcoming threats to learning in development programs*. Ottawa: International Development Research Centre. (www.alnap.org/resource/19253.aspx).

Sphere Project. (2011) *Humanitarian charter and minimum standards in humanitarian response*. Geneva: Sphere Project. (www.alnap.org/resource/8161.aspx).

Start Fund. (2014) *South Sudan crisis response summary dashboard crisis response*. London: Start Fund. (www.alnap.org/resource/19411.aspx).

Start, D. and Hovland, I. (2004) *Tools for policy impact: A handbook for researchers*. London: ODI. (www.alnap.org/resource/8479.aspx).

Steets, J., Darcy, J., Weingartner, L. and Leguene, P. (2014) *Strategic evaluation FAO/WFP joint evaluation of food security cluster coordination in humanitarian action*. Rome: WFP. (www.alnap.org/resource/19699.aspx).

Steets, J. and Dubai, K. (2010) *Real-time evaluation of UNICEF's response to the Sa'ada conflict in Northern Yemen*. Berlin: GPPi. (www.alnap.org/resource/8896.aspx).

Steets, J., Grünwald, F., Binder, A., de Geoffroy, V., Kauffmann, D., Krüger, S., Meier, C., and Sokpoh, B. (2010) *Cluster approach evaluation 2 Synthesis Report*. Berlin: GPPi. (www.alnap.org/resource/5986.aspx).

Stein, D. and Valters, C. (2012) *Understanding theory of change in international development: a review of existing knowledge*. London: London School of Economics. (www.alnap.org/resource/23088.aspx).

- Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., and Befani, B. (2012)**
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).
- Stetson, V. (2008)** *Communicating and reporting on an evaluation. Guidelines and tools*. Washington, DC: American Red Cross. (www.alnap.org/resource/9888.aspx).
- Stoddard, A., Harmer, A., Haver, K., Salomons, D. and Wheeler, V. (2007)**
Cluster approach evaluation. New York: OCHA. (www.alnap.org/resource/3641.aspx).
- Stokke, K. (2007)** *Humanitarian response to natural disasters: a synthesis of evaluation findings*. Oslo: NORAD. (www.alnap.org/resource/3551.aspx).
- Street, A. (2009)** *Synthesis report: Review of the engagement of NGOs with the humanitarian reform process*. (www.alnap.org/resource/12647.aspx).
- Suchman, L. and Jordan, B. (1990)** 'Interactional troubles in face-to-face survey interviews'. *Journal of the American Statistical Association*, 85(409): 232–241. (www.alnap.org/resource/23089.aspx).
- Sutter, P., Downen, J., Walters, T., Izabiriza, B., Kagendo, R. S., Langworthy, M., Sagara, B. and Mueller, M. (2012)** *The contribution of food assistance to Durable Solutions in protracted refugee situations: its impact and role in Rwanda (2007–2011)*. Rome: WFP. (www.alnap.org/resource/19832.aspx).
- TANGO International. (2007)** *End of project study of tsunami impacted communities in Southern India*. Tucson, AZ: TANGO International. (www.alnap.org/resource/3536.aspx).
- Taplin, D. H. and Heléne, C. (2012)** *Theory of change basics: A primer on theory of change*. New York: ActKnowledge. (www.alnap.org/resource/22902.aspx).
- Taplin, D. H., Clark, H., Collins, E. and Colby, D. C. (2013)** *Theory of change: Technical papers: a series of papers to support development of theories of change based on practice in the field*. New York: ActKnowledge. (www.alnap.org/resource/23090.aspx).

Taylor-Powell, E. and Henert, E. (2008) *Developing a logic model: Teaching and training guide*. Madison, WI: University of Wisconsin. (www.alnap.org/resource/23091.aspx).

Tearfund. (2015) *Impact and learning report 2015: Inspiring change*. Teddington: Tearfund. (www.alnap.org/resource/23092.aspx).

Teddlie, C. and Yu, F. (2007) 'Mixed methods sampling: A typology with examples'. *Journal of Mixed Methods Research*, 1(1): 77–100. (www.alnap.org/resource/23093.aspx).

Telford, J. (1997) *Counting and identification of beneficiary populations in emergencies: registration and its alternatives*. London: ODI. (www.alnap.org/resource/11010.aspx).

Telford, J. (2009) *Review of joint evaluations and the future of inter agency evaluations*. New York: OCHA. (www.alnap.org/resource/5772.aspx).

Telford, J. and Cosgrave, J. (2007) *The international humanitarian system and the 2004 Indian Ocean earthquake and tsunamis*. (www.alnap.org/resource/8162.aspx).

Telford, J., Cosgrave, J. and Houghton, R. (2006) *Joint evaluation of the international response to the Indian Ocean tsunami: Synthesis report*. London: TEC. (www.alnap.org/resource/3535.aspx).

Telyukov, A. and Paterson, M. (2008) *Impact evaluation of PRM humanitarian assistance to the repatriation and reintegration of Burundi refugees (2003-08)*. Washington, DC: USAID. (www.alnap.org/resource/5661.aspx).

Ternström, B. (2013) *Evaluation of five humanitarian programmes of the Norwegian Refugee Council and of the standby roster: NORCAP*. Oslo: NORAD. (www.alnap.org/resource/10101.aspx).

Tessitore, S. (2013) "Like a good trip to town without selling your animals": *A study of FAO Somalia's cash for work programme*. Rome: FAO. (www.alnap.org/resource/8821.aspx).

Theis, J. and Grady, H. (1991) *Participatory rapid appraisal for community development: A training manual based on experiences in the Middle East and North Africa*. London: IIED. (www.alnap.org/resource/8163.aspx).

Thizy, D. (2013) *Federation-wide livelihood program evaluation*. Geneva: IFRC.
(www.alnap.org/resource/8864.aspx).

Thomas, A. and Mohnan, G. (2007) *Research skills for policy and development: how to find out fast*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/8164.aspx).

Thomas, V. and Beck, T. (2010) *Changing the way UNHCR does business? An evaluation of the age, gender and diversity mainstreaming strategy, 2004-2009*. Geneva: UNHCR. (www.alnap.org/resource/5852.aspx).

Todd, D., Batchelor, C., Brouder, S., Coccia, F., Castiel, E. F. and Soussan, J. (2015) Evaluation of the CGIAR research program on water, land and ecosystems. Montpellier: CGIAR. (www.alnap.org/resource/22803.aspx).

Tolani-Brown, N., Mortensen, J. and Davis, J. (2010) *Evaluation of the UNICEF education programme in Timor Leste 2003 – 2009*. New York: UNICEF. (www.alnap.org/resource/6383.aspx).

Tomlinson, M., Solomon, W., Singh, Y., Doherty, T., Chopra, M., Ijumba, P., Tsai, A. and Jackson, D. (2009) *The use of mobile phones as a data collection tool: A report from a household survey in South Africa*. London: BioMed Central. (www.alnap.org/resource/8165.aspx).

Beck, T. (2011) *Joint humanitarian impact evaluation: Report for the inter-agency working group on joint humanitarian impact evaluation*. London: TEC. (www.alnap.org/resource/6046.aspx).

tools4dev. (2014) *How to pretest and pilot a survey questionnaire*. (www.alnap.org/resource/22804.aspx).

Tufte, E. R. (1990) *Envisioning information*. Graphics Press.
(www.alnap.org/resource/8166.aspx).

Tufte, E. R. (1997) *Visual explanations: images and quantities, evidence and narrative*. Graphics Press. (www.alnap.org/resource/22805.aspx).

Tufte, E. R. and Graves-Morris, P. R. (2001) *The visual display of quantitative information*. Graphics Press. (www.alnap.org/resource/8167.aspx).

Tulane University. (2012) *Haiti humanitarian assistance evaluation from a resilience perspective*. New Orleans, LA: Tulane University.
(www.alnap.org/resource/6377.aspx).

Tuttle, A. H., Tohyama, S., Ramsay, T., Kimmelman, J., Schweinhardt, P, Bennett, G., Mogil, J. (2015) 'Increasing placebo responses over time in U.S. clinical trials of neuropathic pain'. *Pain*, 156(12): 2616-2626. (www.alnap.org/resource/22806.aspx).

Ubels, J., Acquaye-Baddoo, N. A. and Fowler, A. (2010) *Capacity development in practice*. London: Earthscan. (www.alnap.org/resource/22808.aspx).

UN General Assembly. (1989) *Convention on the rights of the child*. New York: UN General Assembly. (www.alnap.org/resource/22809.aspx).

UN General Assembly. (1951) *Convention relating to the status of refugees*. New York: UN General Assembly. (www.alnap.org/resource/22810.aspx).

UNDP. (2009) *Handbook on planning, monitoring and evaluating for development results*. New York: UNDP. (www.alnap.org/resource/22812.aspx).

UNDP. (2011) *Updated guidance on evaluation in the handbook on planning, monitoring and evaluation for development results (2009)*. New York: UNDP. (www.alnap.org/resource/22813.aspx).

UNEG. (2005a) *Norms for evaluation in the UN system*. New York: UNEG. (www.alnap.org/resource/8215.aspx).

UNEG. (2005b) *Standards for evaluation in the UN system*. New York: UNEG. (www.alnap.org/resource/19410.aspx).

UNEG. (2008a) *Core competencies for evaluators of the UN System*, New York: UNEG. (www.alnap.org/resource/22814.aspx).

UNEG. (2008b) *Evaluability assessments of the programme country pilots: Delivering as one UN*. New York: UNEG. (www.alnap.org/resource/22815.aspx).

UNEG. (2010a). *Good practice guidelines for follow up to evaluations*. New York: UNEG. (www.alnap.org/resource/8446.aspx).

UNEG. (2010b) *UNEG quality checklist for evaluation reports*. New York: UNEG. (www.alnap.org/resource/22816.aspx).

UNEG. (2011) *Integrating human rights and gender equality in evaluation - Towards UNEG guidance*. New York: UNEG. (www.alnap.org/resource/8171.aspx).

UNEG. (2013a) *Resource pack on joint evaluations*. New York: UNEG.
(www.alnap.org/resource/12613.aspx).

UNEG. (2013b) *UNEG handbook for conducting evaluations of normative work in the UN system*. New York: UNEG. (www.alnap.org/resource/22817.aspx).

UNEG. (2014) *Integrating human rights and gender equality in evaluation - Towards UNEG guidance*. New York: UNEG. (www.alnap.org/resource/19283.aspx).

UNEG HEIG. (2016) *Reflecting Humanitarian Principles in Evaluation*.
New York: UNEG. (<http://www.alnap.org/resource/23385.aspx>).

UNHCR. (1997) *Commodity distribution*, Geneva: UNHCR.
(www.alnap.org/resource/22818.aspx).

UNHCR. (2007) *Handbook for emergencies*. New York: UNHCR.
(www.alnap.org/resource/8172.aspx).

UNHCR and WFP. (2013) *Synthesis report of the joint WFP and UNHCR impact evaluations on the contribution of food assistance to durable solutions in protracted refugee situations*. Geneva and Rome: UNHCR/WFP.
(www.alnap.org/resource/12260.aspx).

UNICEF. (1998) *The state of the world's children*. New York: UNICEF.
(www.alnap.org/resource/19280.aspx).

UNICEF. (2002) *Children participating in research, monitoring And evaluation - Ethics and your responsibilities as a manager*. New York: UNICEF.
(www.alnap.org/resource/8174.aspx).

UNICEF. (2005) *Emergency field handbook*. New York: UNICEF.
(www.alnap.org/resource/8173.aspx).

UNICEF. (2010) *Evaluation of the UNICEF education programme in Timor Leste 2003 – 2009*. New York: UNICEF. (www.alnap.org/resource/6383.aspx).

UNICEF. (2013a) *Evaluability assessment of the peacebuilding, education and advocacy programme*. New York: UNICEF. (www.alnap.org/resource/12679.aspx).

UNICEF. (2013b) *Evaluation of UNICEF programmes to protect children in emergencies: Pakistan country case study*. New York: UNICEF.
(www.alnap.org/resource/22820.aspx).

- UNICEF. (2013c)** *Evaluation of UNICEF programmes to protect children in emergencies. Synthesis report*. New York: UNICEF. (www.alnap.org/resource/12491.aspx).
- UNICEF. (2013d)** *Global evaluation reports oversight system (GEROS)*, New York: UNICEF. (www.alnap.org/resource/7948.aspx).
- UNICEF. (2013e)** *Thematic synthesis report on evaluation of humanitarian action*. New York: UNICEF. (www.alnap.org/resource/8442.aspx).
- UNICEF. (2013f)** *Improving Child Nutrition: The achievable imperative for global progress*. New York: UNICEF. (<http://www.alnap.org/resource/23386.aspx>).
- UNICEF. (2014)** *Evaluation of UNICEF's cluster lead agency role in humanitarian action (CLARE)*. New York: UNICEF. (www.alnap.org/resource/19327.aspx).
- UNICEF and Save the Children. (2010)** *Management response to the review of the global education cluster co-leadership arrangement*. New York and London: UNICEF/Save the Children. (www.alnap.org/resource/6354.aspx).
- UNIFEM. (2011)** *UNIFEM strategic plan 2008-2011: Evaluability assessment*. New York: UNIFEM. (www.alnap.org/resource/22821.aspx).
- Universalia. (2013)** *GEROS – Global meta-evaluation report 2012: final report*, New York: UNICEF. (www.alnap.org/resource/22822.aspx).
- UNODC. (2012)** *Evaluability assessment template*. Vienna: UNODC. (www.alnap.org/resource/22823.aspx).
- USAID. (1996)** *Preparing an evaluation scope of work*. Washington: USAID. (www.alnap.org/resource/22825.aspx).
- USAID. (2010)** *Performance monitoring and evaluation tips using rapid appraisal methods*. Washington, DC: USAID. (www.alnap.org/resource/19281.aspx).
- USAID. 2012.** *How-to note: Preparing evaluation reports*, Washington, DC: USAID. (www.alnap.org/resource/22824.aspx).

USAID. (2013) *After-action review guidance*. Washington, DC: USAID. (www.alnap.org/resource/19257.aspx).

USAID. (2015) *USAID office of food for peace food security desk review for Katanga, North Kivu and South Kivu, Democratic Republic of Congo*. Washington, DC: USAID. (www.alnap.org/resource/22826.aspx).

Valters, C. (2014) *Theories of change in international development: Communication, learning, or accountability?* London: London School of Economics. (www.alnap.org/resource/19265.aspx).

Van Bruaene, M., Dumélie, R., Kunze, M., Pankhurst, M. and Potter, J. (2010) *Review concerning the establishment of a European voluntary humanitarian aid corps*. Brussels: Prolog Consult. (www.alnap.org/resource/3633.aspx).

Varanoa, S. P., Schaferb, J. A., Cancinoc, J. M., Deckerd, S., H. and Greenee, J. R. (2010) 'A tale of three cities: Crime and displacement after Hurricane Katrina'. *Journal of Criminal Justice*, 38(1): 42-50. (www.alnap.org/resource/22827.aspx).

Wagner, A. K., Soumerai, S. B., Zhang, F. and Ross-Degnan, D. (2002) 'Segmented regression analysis of interrupted time series studies in medication use research'. *Journal of Clinical Pharmacy and Therapeutics*, 2002(27): 299-309. (www.alnap.org/resource/22830.aspx).

Walden, V. (2013) *A quick guide to monitoring, evaluation, accountability, and learning in fragile contexts*. Oxford: Oxfam. (www.alnap.org/resource/10593.aspx).

Warner, A. (2014) *EHA, let's do some stretches! ALNAP EHA Practice Note*. Shared in ALNAP Humanitarian Evaluation Community of Practice in October.

Warner, A. (2014) *Evaluation of humanitarian action discussion series*. *Repeat after me: communicate, disseminate and support take-up!* London: ALNAP. (www.alnap.org/resource/12920.aspx).

Webb, E. J., Campbell, D. T., Schwartz, R. D. and Sechrest, L. (1966) *Unobtrusive measures: nonreactive research in the social sciences*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22831.aspx).

Webster, C. (1996) 'Hispanic and Anglo interviewer and respondent ethnicity and gender: The impact on survey response quality. *Journal of Marketing Research*, 33(1): 62-72. (www.alnap.org/resource/8176.aspx).

Weingärtner, L., Otto, R. and Hoerz, T. (2011) *Die deutsche humanitäre Hilfe im Ausland. Band I: Hauptbericht.* Bonn: BMZ. (www.alnap.org/resource/22832.aspx).

Weiss, C. H. (1997) *Evaluation methods for studying programs and policies.* Upper Saddle River, NJ: Prentice Hall. (www.alnap.org/resource/8540.aspx).

Wesonga, D. (2013) *Accelerated primary education support (APES) project: Final evaluation report.* London: Concern Worldwide. (www.alnap.org/resource/12232.aspx).

WFP. (2002) *Emergency field operations pocketbook.* Rome: WFP. (www.alnap.org/resource/8177.aspx).

WFP. (2010) *Evaluation Matrix.* Rome: WFP. (www.alnap.org/resource/22833.aspx).

WFP. (2012) *Mixed-method impact evaluation - The contribution of food assistance to durable solutions in protracted refugee situations: its impact and role in Bangladesh.* Rome: WFP. (www.alnap.org/resource/19830.aspx).

WFP. (2013) *Synthesis of the series of joint UNHCR-WFP impact evaluations of food assistance to refugees in protracted situations.* Rome: WFP. (www.alnap.org/resource/19903).

WFP. (2014) *Impact evaluations I. guidance for process and content.* Rome: WFP. (www.alnap.org/resource/22834.aspx).

WFP. (2015) *Evaluation quality assurance system (EQAS) guidelines for operation evaluations.* Rome: WFP. (www.alnap.org/resource/22835.aspx).

White, H. (2009a) *Some reflections on current debates in impact evaluation.* New Delhi: 3ie. (www.alnap.org/resource/8178.aspx).

White, H. (2009b) *Theory-based impact evaluation: principles and practice.* New Delhi: 3ie. (www.alnap.org/resource/8179.aspx).

White, H., 2014. *The counterfactual. Howard White International Initiative for Impact Evaluation.* New Delhi: 3ie. (www.alnap.org/resource/22836.aspx).

White, H. and Sabarwal, S. (2014a) *Developing and selecting measures of child well-being.* Florence: UNICEF. (www.alnap.org/resource/22837.aspx).

- White, H. and Sabarwal, S. (2014b)** *Modelling*. Florence: UNICEF.
(www.alnap.org/resource/22838.aspx).
- White, H. and Sabarwal, S. (2014c)** *Quasi-experimental design and methods*,. Florence: UNICEF. (www.alnap.org/resource/22839.aspx).
- White, H., Sabarwal, S. and de Hoop, T. (2014)** *Randomized controlled trials (RCTs)*. Florence: UNICEF. (www.alnap.org/resource/22840.aspx).
- van der Wijk, J. and Kazan, L. (2010)** *Evaluation of the DEC-funded CAFOD health and WASH project in the DRC*. London: CAFOD. (www.alnap.org/resource/5849.aspx).
- Wilding, J., Swift, J. and Hartung, H. (2009)** *Mid term evaluation of DG ECHO's regional drought decision in the Greater Horn of Africa: March - May 2009*. Brussels: ECHO. (www.alnap.org/resource/5919.aspx).
- Wiles, P., Chan, L. and Horwood, C. (1999)** *Evaluation of Danish humanitarian assistance to Afghanistan 1992-98*. Copenhagen: DANIDA. (www.alnap.org/resource/2827.aspx).
- Wilson, P. and Reilly, D. (2007)** *Joint evaluation of their responses to the Yogyakarta earthquake July 2007*. Atlanta, GA: CARE International.
(www.alnap.org/resource/3611.aspx).
- W.K. Kellogg Foundation. (1998)** *W.K. Kellogg Foundation evaluation handbook*. Albuquerque, NM: (www.alnap.org/resource/22841.aspx).
- World Bank. (2004)** *Citizen report card surveys - A note on the concept and methodology*. Washington, DC: World Bank. (www.alnap.org/resource/8182.aspx).
- World Bank. (2009)** *Making smart policy: using impact evaluation for policy making*. Washington, DC: World Bank. (www.alnap.org/resource/8434.aspx).
- World Bank. (2007a)** 'Participatory tools for micro-level poverty and social impact analysis', in *Tools for institutional, political, and social analysis of policy reform*. Washington, DC: World Bank. (www.alnap.org/resource/22842.aspx).
- World Bank. (2007b)** *Sourcebook for Evaluating Global and Regional Partnership Programs: Indicative Principles and Standards*, Washington: The Independent Evaluation Group of the World Bank. (www.alnap.org/resource/22843.aspx).

Wortel, E. (2009) *Humanitarians and their moral stance in war: the underlying values*. Geneva: ICRC. (www.alnap.org/resource/8183.aspx).

Yang, C. H. Xirasagar, S., Chung, H. C., Huang, Y. T. and Lin, H. C. (2005)
'Suicide trends following the Taiwan earthquake of 1999: empirical evidence and policy implications'. *Acta Psychiatrica Scandinavica*, 112: 442–448. (www.alnap.org/resource/8184.aspx).

Yarbrough, D., Shulha, L., Hopson, R. and Caruthers, F. *The program evaluation standards*. Thousand Oaks, CA: SAGE. (www.alnap.org/resource/22725.aspx).

Zaiontz, C. (2015) *Performing real statistical analysis using excel*. (www.alnap.org/resource/22844.aspx).