Missing the point? Reflections on current practice in evaluating humanitarian action

Discussion paper

James Darcy and Neil Dillon

Contents

Acknowledgements:

The authors would like to thank those who generously gave their time to be interviewed in preparation for writing this paper (their names are listed in the Annex). In particular they would like to thank Alice Obrecht and Alistair Hallam for their helpful feedback on an earlier draft of the paper, and Hannah Caddick and Maria Gili for their judicious editing. Responsibility for the views expressed here lies solely with the authors.

1. Introduction

Humanitarian evaluation is more prolific than ever before: more evaluations are being conducted, by more agencies, covering more crises. The volume of evaluation output appears to have increased significantly over the past decade, as evidenced by a 100% increase in uploads to ALNAP's HELP library between 2008 and 2018. Over the same period, humanitarian agencies have put significant effort into updating and revising their evaluation norms and standards, placing more emphasis on the quality and use of evaluations in the sector. Yet the quality and usefulness of these evaluations seems to vary greatly. A review of recent humanitarian evaluations, conducted for ALNAP's State of the Humanitarian System (SOHS) 2018 report,[1] found a mismatch between this growth in effort and output on the one hand, and continuing concerns about the quality and utility of humanitarian evaluation on the other.[2]

Three issues in particular stood out. The first was the mixed quality of the evaluations concerned, which derived in part from a high degree of variability in methodological rigour. The second issue was the limited scope of the evaluations reviewed: the great majority deal with context-specific crisis responses by individual agencies; relatively few are concerned with system-wide performance or organisational performance across a range of different contexts. The third issue was a fundamental inconsistency in how and to what extent *effectiveness* is understood and assessed, or indeed properly evaluated at all.

These issues are not new. In 1999, the Organisation for Economic Co-operation and Development (OECD) described the variance in humanitarian evaluation as 'methodological anarchy'. There has since been extensive discussion on the subject of improving evaluation quality (e.g. World Bank IEG, 2009; Rogers, 2009; Stern et al., 2012). ALNAP and others have undertaken several initiatives to improve the professionalism and quality of evaluative practice in humanitarian settings, and many agencies have taken steps towards this by:

- revising evaluation policies (e.g. DFID, 2013; UNICEF, 2018)
- enhancing quality assurance systems (e.g. the World Food Programme's Evaluation Quality Assurance System and UNICEF's Global Evaluation Reports Oversight System)
- improving evaluation methods (e.g. Stern et al., 2012; Vogel, 2012)
- peer review of evaluation systems (Liverani and Lundgren, 2007; UNEG, 2011; 2012)
- agreeing to common norms and standards for evaluation (UNEG, 2016).

So, why do the quality and utility of humanitarian evaluations still continue to raise concerns? Given the apparent advances in methodological practice, the problem may lie deeper than tools and frameworks. The weaknesses seen in the 2018 SOHS evaluation review warrant a closer look at the foundations of humanitarian evaluation. In particular, how do we understand the purpose and function of evaluation in this sector? What kinds of evidence do evaluations generate? How do we evaluate effectiveness in humanitarian contexts? And how does variable organisational practice relate to questions of quality and scope?

In seeking to answer these questions, this paper addresses the main issues with current evaluation practice noted in the SOHS 2018 evaluation synthesis. It compares the purpose and function of humanitarian evaluations with the range of approaches available beyond the sector, and reviews the evaluation policies of major international humanitarian agencies to better understand how the evaluation function is currently understood.

The paper employed four main research methods:[3]

- A **review of the academic literature** on evaluation function, including 23 academic reports and grey literature on evaluation from inside and outside the humanitarian community.
- A **review of evaluation policies** and frameworks, using a purposive sample of 16 evaluation policies (including United Nations, non-governmental organisations (NGOs) and donor organisations). Each policy covered organisation-wide evaluation policy at the time of writing.

[1] A sample of 120 evaluations (chosen for their scope and relevance) out of 549 publicly available evaluative studies were reviewed and common themes identified, along with some common characteristics of the evaluations themselves.
[2] Informants to this study tended to echo earlier concerns over utility and utilisation of humanitarian evaluations raised by Hallam and Bonino (2013), building on Sandison (2006).
[3] A list of those interviewed and list of evaluation policies reviewed are provided in Annex A of this paper. See Technical Annex for a fuller account of the methods used.

- **Key stakeholder interviews**, including with a purposive sample of 21 stakeholders within the humanitarian evaluation community, primarily with heads of evaluation offices, senior evaluation managers and senior evaluators from across the ALNAP Membership.
- A **review of the evaluation synthesis** conducted for the 2018 SOHS. The authors used the sample of evaluations selected for this synthesis to assess evidential quality, its driving factors, evaluation scope and method.

This paper aims to stimulate further discussion both of current evaluation practice in the sector and, in proposing potential solutions to the various challenges identified, of possible future directions. It should be of interest to evaluators and those who commission evaluations, and to the wider field of humanitarian practitioners who are often the intended end users of evaluations.
Section 2 reviews the types of evaluative evidence generated by current evaluation practice in the sector as a prelude to a discussion of the purpose and function of evaluation (section 3). The authors draw here on the wider literature on development and public policy evaluation and consider in particular the ways in which the evaluation of effectiveness might be better understood in the humanitarian sector. The paper concludes with three specific suggestions as to how the sector might evolve evaluation practice to address some of the problems noted (section 4) and offers some final concluding thoughts (section 5).

2. Assessing the coverage and quality of evidence from current evaluation practice

The evaluation synthesis conducted for the SOHS 2018 report found that the quality of evidence generated by humanitarian evaluation was highly variable. Analysis of this variability for the current paper suggested that quality was driven primarily by organisational factors – such as agency and evaluation type – together with the choice of evaluators, rather than the contextual challenges posed by conducting evaluations in different crises or regions. The review found that the scope of humanitarian evaluations is typically confined to single agency responses (projects or programmes) assessed in their own terms, and rarely looks at the links to contextual and system-wide issues. Finally, it found there remains a radical inconsistency in how *effectiveness* is understood and evaluated.

2.1 Quality and depth of evidence

Using the methodology outlined in the Technical Annex, the authors reviewed and assessed the quality of evidence provided in the evaluations that were selected for the SOHS 2018 report.[4] Based on this, the authors identified those factors most associated with high-quality evidence.

The review found that, on average, evidential quality across the evaluations was fairly good, but that **the quality of evidence from individual evaluations varied considerably** around this average.[5] Strikingly, evidential depth (broadly the weight and scope of evidence used) and evaluations' relevance for the SOHS report were rated more poorly than the quality of analysis.[6] This suggests that the evaluation system is generally better at setting the framework (clear questions and methods) for high-quality evidence than it is at providing in-depth answers to them.

When looking at the factors most associated with high-quality evidence, the review found **evidential quality was highly correlated with evaluation type, agency and sector-focus** – which notably goes a long way to determining the choice of evaluators themselves. The subject of the evaluation, evaluation criteria used, geographical region and crisis type had low correlation with quality. This suggests that quality variance is

[4] An evaluation of quality formed part of the evidence-weighting process for the SOHS evaluation synthesis. 'Quality' was understood primarily with reference to clarity and cogency of analysis. The question of quality, and the factors associated with higher or lower quality, were further analysed for the purposes of this paper – see Technical Annex for detail.
[5] The mean average quality score under the methodology adopted was 2.42 out of 3.00, or 81%. The standard deviation around the mean average was 0.63, and the coefficient of variance was 25%.
[6] 'Evidential depth' was understood both in terms of the weight of evidence adduced and the extent to which the evaluations considered a range of potential causal factors and evidence beyond the immediate context of the programme concerned. The average score for depth of evidence was 2.23 out of 3.00 (74%).

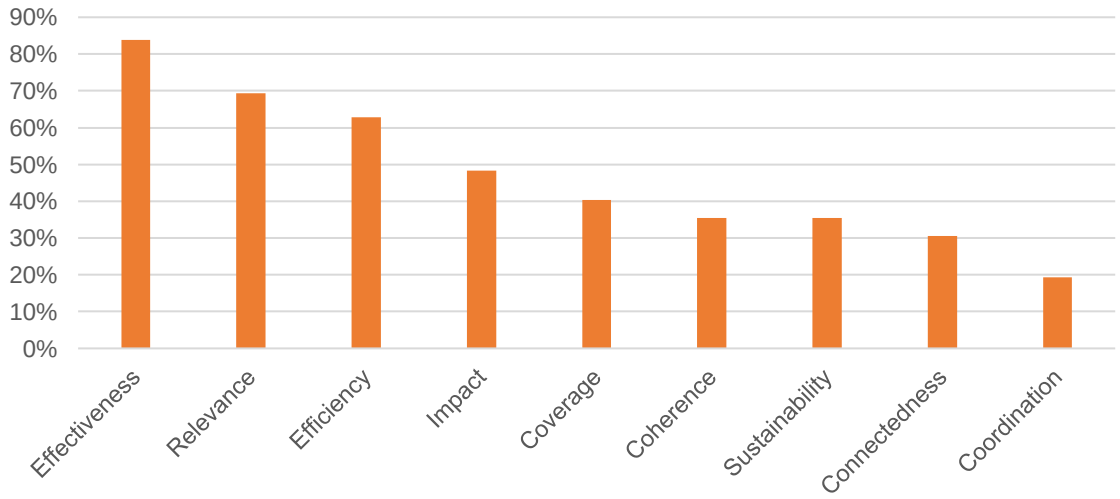
related more to the different institutional set-ups and evaluation processes than to the difficulty of doing evaluation in crisis settings.

Following this train of thought, the authors looked at the impact of the different evaluation methods used. Of 120 evaluations sampled, none used a control group, and none were purely quantitative. Rather, there was a 50:50 split between purely qualitative studies and mixed-method approaches.[7] The difference in evidential quality between these two approaches was noticeable, both in terms of mean average quality score and variance. Compared to the mixed-method evaluations, the **qualitative studies were more variable and, on average, lower quality**.[8]

2.2 Categories of evidence and evidence gaps

The SOHS process raised questions about the evaluation coverage of some of the key topics of concern for the sector. Figure 1 shows the extent to which the OECD Development Assistance Committee (DAC) Evaluation Criteria[9] – used almost universally as a reference point in framing the scope of evaluations – were covered in the evaluation sample reviewed for the 2018 SOHS report.

Figure 1 Evaluation coverage per the DAC Evaluation Criteria



Source: author's own

Here, and in a wider review of relevant studies for the SOHS report, the authors found a **marked absence of studies that looked at systemic or response-wide issues**, with most looking at project or single-agency level. Only 20% of the 549 studies identified in the 2015 to 2018 period were considered relevant to the system-wide scope of the SOHS report. Even within the sample of evaluations selected for the review, the focus of the stated evaluation frameworks tended to be on evaluation criteria that are more obviously related to the performance of single projects and programmes (relevance, effectiveness and efficiency) as opposed to those relating more to systemic and response-wide issues (connectedness and coordination).

It should be stressed that the evaluation sample did not show that effectiveness, relevance and efficiency were covered *well* by the evaluations in the sample, nor that impact had been properly addressed by almost 50% of the evaluations. Rather, it showed that the evaluation questions and frameworks explicitly sought answers to questions relating to these criteria (and generally did not ask questions about coordination or connectedness).

[7] The definition of 'mixed-method' employed here is that proposed by Bamberger (2012): to count as mixed-method, an evaluation had to include elements of both qualitative and quantitative primary data collection and analysis in the methodology defined by the final report.
[8] See Technical Annex for further detail. Qualitative evaluations had a mean average score of 2.25 out of 3.00 and a variance of 28%. This compares to 2.56 and 20% for the mixed methods studies.
[9] Evaluation criteria established by the OECD DAC, as modified for use in humanitarian contexts. See Beck (2006).

Those consulted for this paper also noted that, although the production of evaluations has risen over the past 20 years, clear gaps remain in the topics covered. Interviewees highlighted shortcomings in coverage related to issues loosely termed 'strategic', such as response-wide coverage of affected people; inter-agency coordination; inter-sectoral resilience initiatives and the humanitarian-development nexus; and, to a lesser extent, system-wide commitments such as the Grand Bargain. Likewise, informants noted that the number of joint or interagency evaluations has remained small in comparison to the volume of single-agency studies.

The emerging picture is one in which individual agencies have over the past decade increasingly evaluated their own activities – perhaps a reflection of increased public accountability pressures to demonstrate aid results. But there are few systems in place to evaluate the collective performance of humanitarian agencies in response to major crises or to locate interventions in a wider context, which makes it hard for humanitarian evaluations to provide a solid basis for collective learning or joint reflection. This gap is felt particularly keenly in areas where system-wide analyses are most important, including the humanitarian–development nexus and other strategic challenges noted in the previous paragraph. Each of these issues is critical for evolving the humanitarian system and the subject of growing policy attention in the wake of the 2016 World Humanitarian Summit.

3. Understanding evaluation's purpose and function

The observed weaknesses in the evidence of the evaluation sample (section 2) relate in part to different organisational approaches to evaluation and to methodological rigour. But interviews also suggested that these weakness and persistent concerns about evaluation usefulness may relate to differences in understanding about the purpose and function of evaluations within organisations and the humanitarian system.[10] One hypothesis arising from interviews for this paper is that the variance in quality output noted above is related to different understandings of what we expect evaluations to be: how we understand their intended purpose and actual function within organisations and within the humanitarian system.

This section considers the ways in which the purpose and function of evaluation is understood inside and outside the humanitarian sector. Specifically, it compares the conception of evaluation's function that emerges from the evaluation policies of the major humanitarian agencies with that found in the literature on development and public policy evaluation. The authors suggest ways in which that wider literature can help to inform future thinking about humanitarian evaluation. First, we consider current interpretations of the evaluation function in the humanitarian sector.

3.1 Current understanding in the humanitarian sector

A focus on learning and accountability

The purpose of evaluations in the humanitarian sector is typically described in terms of **learning** and **accountability**. In practice, these twin goals are rarely elucidated. Although published evaluation policies and frameworks consistently focused on the role of evaluation as a learning and accountability tool, our review found relatively little discussion of what these two goals mean in practice. Both learning and accountability imply a link to better-informed future decision-making, yet it is often unclear what decisions evaluation is intended to inform or how agencies understand the relationship between evaluation and other evidence tools, such as monitoring and needs assessment. Learning and accountability are broad categories within which evaluation is expected to deliver often unspecified results. This makes it difficult to determine what kind of evaluation process is appropriate in a given situation – and indeed whether other processes (such as peer review, technical or strategy review) may be more suitable.

In theory, the purpose of evaluation is to provide a sound assessment of the value of a project, programme or policy (Scriven, 1991). This translates into a number of sub-agendas. In the humanitarian context, evaluations are generally expected to assess the relevance and effectiveness of a given intervention or policy (among other criteria), while also addressing related questions of organisational performance. In assessing effectiveness, an evaluation is generally expected to shed light on the *reasons* behind success or failure. The evidence generated is thereby intended to inform future decisions about strategy, programme or policy; to contribute to

[10] This paper distinguishes between purpose (what evaluations are supposed to do) and function (what they actually do or how they are actually used); although in practice the terms are often used interchangeably.

organisational learning and performance improvement; and to satisfy internal or external accountability requirements.

Some forms of enquiry and assessment used in the humanitarian sector, such as applied research and programme monitoring, are designed to provide objective and essentially value-neutral evidence about the context and the related intervention. Evaluations, while expected to be grounded in a thorough review of evidence, involve making explicit value judgements. They are also intended to generate evidence of a certain kind and to contribute to learning, for example about 'what works' in addressing a particular humanitarian challenge. In this way, evaluation has both an evidential and an evaluative function: its role is both to build the knowledge base and to assess – and sometimes challenge – assumptions about a project's value or worth.

In the evaluation policies reviewed for this paper, evaluation purpose was most commonly described as objective results measurement. The United Nations Evaluation Group (UNEG) Norms and Standards define evaluation as the systematic and impartial assessment of results achieved (2016). Likewise, the Danish Refugee Council defines evaluation purpose as 'understanding the results of our work' (DRC, 2015: 4). While 'results' can mean a range of different things, this emphasis on objectivity and results measurement is broadly shared across the policies of the individual United Nations (UN) agencies and many international NGOs.[11]

The model of evaluation as central to a process of *evidence-based decision-making* also comes through in the evaluation policies. The World Food Programme's 2016–2021 policy says that evaluation aims to generate 'relevant recommendations for optimal use in evidence-based decision-making' (WFP, 2015). Similarly, the United Nations High Commissioner for Refugees suggests that:

> An evaluation should provide credible, useful evidence-based information that enables the timely incorporation of its findings, recommendations and lessons into the decision-making processes of organizations and stakeholders. (UNHCR, 2016: 4)

3.2 Wider understandings of evaluation's purpose and function

Beyond accountability and learning

Contemporary evaluation literature distinguishes a range of different functions, beyond accountability and learning, for evaluation within public policy-making. It identifies two types of value judgement in evaluative practice: those judgements based primarily on the measurement of results and those based on reflective dialogue between competing narratives and different stakeholders. These types are not mutually exclusive, but teasing out the differences between them helps to sheds light on the various possible functions of evaluation. A review of recent literature suggests the following non-exhaustive taxonomy of evaluation functions:

Measuring specific results. In a New Public Management context, evaluation can take a primarily technical function.[12] Under this conception, evaluation deploys objectively verifiable methods with the aim of providing a formal accountability structure to verify claims made by public policies and programmes against criteria such as efficacy, efficiency and long-term impact.

Understanding broadly defined change. Evaluation is seen primarily as a means of identifying whether an activity has contributed to a broadly defined cultural change – for example, in international development, whether public sector reform programmes have contributed to good governance.[13]

Facilitating organisational reflection. The core function of evaluation is to ensure a space for stakeholder dialogue and reflection – potentially including the full range of people affected by the programme.[14] Schwandt (2015) illustrates this type of evaluation using a hypothetical example of a social programme aimed at improving educational opportunities for pre-school-age children. A facilitative approach would consider not only progress towards stated objectives, such as the improved reading ability of the children, but also a range of other community perspectives on how well-adapted the programme was to their specific needs and priorities (ibid: 101).

[11] The most recent UN evaluation policies reviewed were UNICEF (2018), UNHCR (2016) and WFP (2015).
[12] Dahler-Larsen (2012) provides a useful overview of this conception of evaluation.
[13] Betts and Wedgewood (2011) give a detailed presentation of this type of analysis.
[14] See, for example, House and Howe, 1999.

Facilitating system-wide dialogue. Some of the more recent applications of systems-thinking to the evaluation field have identified three main tasks for evaluation.[15] The first is clarifying inter-relationships and power dynamics between people, things, ideas and contexts. The second task is engaging the perspectives of all stakeholders in the project and supporting the agency of people from the margins of the intervention to stimulate transformation (Stephens et al., 2018: 19). Finally, the third is reflecting on the boundaries between the programme, the broader organisational architecture, the crisis response and the crisis context itself.

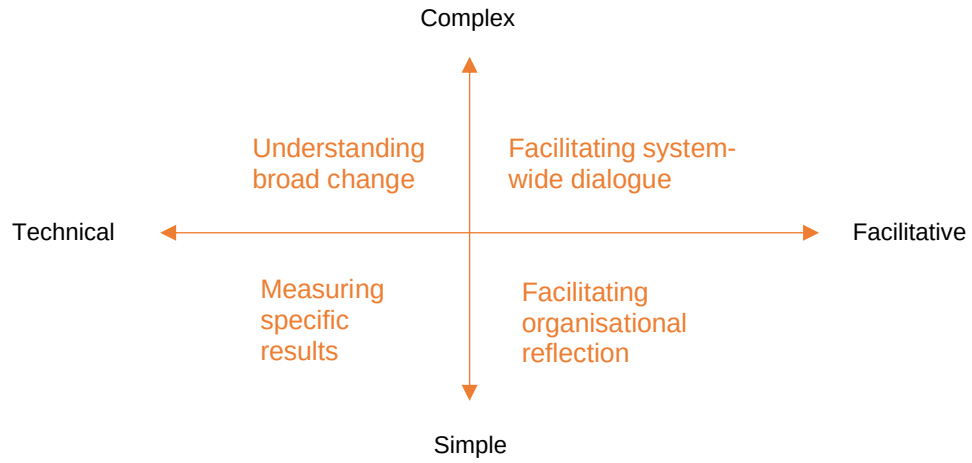
Classifying evaluation types according to function and purpose

While these four categories may be as much a description of the characteristics of good evaluation *process* as they are of their *functions*, they point to a way of classifying evaluation types across two dimensions (Figure 1):

- **Technical vs facilitative:** the degree to which evaluation is seen as the provision of value-neutral evidence versus facilitating stakeholder dialogue between competing narratives.
- **Simple vs complex:** the degree to which evaluation seeks to analyse linear relationships between programme activities and intended results, versus analysing the complex interactions between programmes, contexts, stakeholders, and hard-to-define or predict changes, often at a community or system-wide level.

By looking at evaluation functions in this way, we can develop a clearer idea of what to expect from an evaluation, beyond the simple overarching goals of learning and accountability.

Figure 2 Mapping evaluation types across two dimensions



Source: authors' own.

While the simple vs complex distinction is relevant to our discussion of the lack of evaluation coverage relating to systemic issues, our main focus here is on the technical vs facilitative distinction.

Schwandt (2015) provides a useful presentation of the different expectations arising from a purely technical or purely facilitative evaluation. **Purely technical evaluations** are primarily concerned with measuring performance against pre-identified and accepted performance metrics. They are based on the following beliefs and expectations:

- The primary function of evaluation is to provide objective and value-neutral evidence.
- Decision-making is best when it is 'evidence-based'. That is, when it follows from the impartial consideration of objective and value-neutral evidence.
- The relationship between an evaluator and a decision-maker is akin to that between an expert and a practitioner: the evaluator provides objective evidence and recommendations that the decision-maker must then apply in a complex and value-laden world.
  (Schwandt, 2015: 98–99)

[15] The clearest example of this approach is given in Williams (2016).

Impact evaluations can often be placed in this purely technical category, given their emphasis on the independent expertise of the evidence provider, value-neutral assessments of merit and worth, and an assumption that decision-making can be improved by an objective assessment of what worked and what did not. Organisational self-assessments against policies or standards would constitute a similarly 'technical' approach to performance assessment, given the focus on a singular understanding of what 'good' looks like for all programme stakeholders.

A more **facilitative evaluation** focuses on the evaluator as the enabler of dialogue and reflection among programme stakeholders:

> Evaluation is a form of practical argumentation that unfolds in complex, often highly political environments in which normative concerns and political choices cannot be neatly separated from the analytic, scientific process involved in determining the value of a program or policy. (Schwandt, 2015: 101)

It is important to note that this facilitation is more than enabling others to learn.[16] Rather, facilitative evaluation means instigating, mediating and guiding dialogue between different stakeholders' potentially competing visions and narratives of a project, programme or policy – including by contrasting the views of the programme team and the evaluator themself.[17]

In contrast to technical evaluation, facilitative evaluation is based on the following understanding:

- The primary function of evaluation is to facilitate organisational reflection based on dialogue between competing value-laden narratives.
- Decision-making is best when it is 'evidence-informed' (cp. 'evidence-based'). That is, it involves consideration of objective evidence alongside a range of other factors, including 'other elements of reasoning that differ from and can contradict scientific reasons', such as social factors (Prewitt et al., 2012).[18]
- The relationship between an evaluator and a decision-maker is like that between a practitioner and their external peers: the evaluator provides a judgement based on critical reflection of competing narratives, including their own experience of similar contexts, which the decision-maker takes into consideration alongside other factors.

Table 1 summarises the differences between technical and facilitative conceptions of evaluation.

Table 1 Underlying beliefs of technical and facilitative evaluation

	Technical evaluation	**Facilitative evaluation**
Approach to objectivity	Value-neutral	Value-based
Model of good decision-making	Evidence-based	Evidence-informed
Evaluator–decision-maker relationship	Expert–practitioner	Practitioner-to-practitioner

The distinction between technical and facilitative is not absolute: as illustrated in Figure 2, it is a spectrum, and evaluations may sit anywhere between the two extremes. Where an evaluation sits on this spectrum can, however, have wide-ranging implications. Expectations of both evaluation users and participants to the evaluation process can differ significantly, depending on the extent to which an evaluation is technical or facilitative, as can the means of conducting the evaluation itself (e.g. in terms of staffing, evaluator skill-sets, and the relationship between evaluators, commissioning units and decision-makers).

[16] As described in, for example, Engel et al. (2003).
[17] For this reason, Schwandt (2015) calls this an essentially 'argumentative' function. It is related to what is known as discourse analysis: 'Discourse analysis involves the recognition of the fact that "there is a plurality of values and arguments available for thinking about any specific policy issue. Analysis, therefore, has to be part of a process in which these several points of view are taken into account..."' (White, 1994, in Gasper and Apthorpe, 1996)
[18] Recent experience of public policy-making in response to the COVID-19 pandemic highlights the importance of this distinction.

Too much focus on the technical?

Much of the discussion around evaluation in public policy-making is primarily technical in nature. Evaluation is often described in value-neutral terms. The United Nations Evaluation Group defines evaluation as the systematic and impartial assessment of an activity, contributing to evidence-based decision-making (UNEG, 2016) – a definition that is echoed by the evaluation policies of individual UN agencies.

But evaluation theorists have questioned this conception. Schwandt (2015), Dahler-Larsen (2012) and Power (1997) have all argued the New Public Management school of public-sector reform has taught us to think of evaluation 'less like a critical voice weighing in on the value (or lack thereof) of public programs and policies and more like a technology that operates with well-defined procedures and indicators' (Schwandt, 2015: 96). If we think of humanitarian evaluation only as a technical endeavour, then we miss opportunities for humanitarian decision-makers and programme teams to reflect and think critically. In doing so, we risk creating a dangerous division of labour between 'expert' and 'practitioner' – and thereby effectively 'subcontracting' thought.

This is surely the opposite of the intended outcome. We should avoid thinking of 'accountability' as requiring a purely technical approach, just as we should not assume 'learning' requires a purely facilitative (non-technical) approach. To fully achieve these twin goals, we may need to take a mixed approach, though the balance will differ in each case.

3.3 Implications for humanitarian evaluation policy

None of the agency evaluation policies reviewed for this paper explicitly considered the distinction between technical and facilitative evaluation, or between simple and complex cause–effect relationships.[19] They do, however, imply a broadly technical conception of evaluation function, while emphasising simple linear relationships between activities and results. There are clear echoes here of Schwandt's description of the 'technical' evaluator as the provider of impartial, value-neutral evidence, unlike the practitioner engaged in a values-based discussion of a programme or policy's merits.

Most would agree that the evidential functions described in section 2 are essential aspects of evaluation, and a concern with generating evidence certainly does not exclude a more facilitative conception of evaluation. But there is a marked absence of any discussion of this. None of the evaluation policies reviewed articulate a role for evaluation in validating or challenging the fundamental value of the programmes and policies, or the prevailing organisational assumptions about a given intervention.[20] And only rarely do they explicitly consider the potential tensions between the perspectives of crisis-affected populations, programme teams, donors and partners. As a result, it is difficult to read the evaluation policies in their current form without concluding that humanitarian evaluations are set up primarily to assess project results, with little consideration of basic questions about value and how perceptions of value might differ between stakeholders.

This is reflected in our review of the evaluations themselves. Of the 120 evaluations selected for the SOHS 2018 evaluation synthesis, the majority provided evidence on overall achievement against objectives, very often focused on output-level achievements – for example whether planned activities were delivered, or the planned number of beneficiaries reached. Most evaluations provided limited discussion of perspectives from affected people and none explicitly sought to consider or incorporate competing narratives about programme performance.[21] In line with the technical orientation of the evaluation policies, it seems the evaluations themselves prioritise reporting against intended 'results' (often interpreted as outputs) rather than fostering dialogue about the value of a given intervention or policy.

[19] This should be contrasted with broader policies on development evaluation. The UK Department for International Development (DFID) for example explicitly 'recognises that development is complex and non-linear and thus evaluation must be designed to reflect this in terms of its ambition, design and application' (DFID, 2013). This of course reflects the department's conception of development rather than humanitarian response, but as we argue in section 3.4, the links between humanitarian action and results are too often wrongly assumed to be linear in nature.

[20] The evaluation of relevance and appropriateness, which might be expected to shed light on these questions, rarely appears to do so in practice – perhaps because of the narrow parameters within which these criteria are often evaluated. Since our study is limited to humanitarian evaluations, we draw no conclusions here as to whether an explicit validation function is similarly lacking in development evaluations.

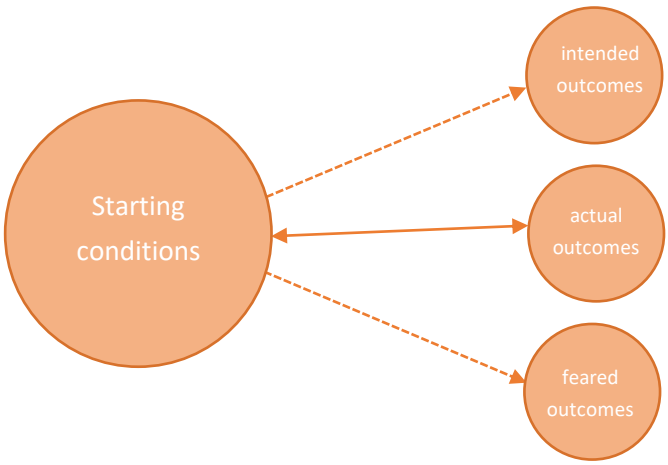
[21] There is an important set of related issues here about the (perceived) need to be seen to succeed, or not to fail, especially in published evaluation reports; and potential tensions between accountability and learning functions in this regard. Further discussion of this is beyond the scope of this paper.

External evaluation in particular offers the chance to validate and challenge prevailing internal narratives and to facilitate dialogue between different stakeholders and perspectives. But it is hard to see how this can be achieved if the focus is solely on measuring results, particularly when those results are often defined either self-referentially (in terms of an agency's outputs) or in terms of outcomes to which there is no clear connection with the intervention concerned. This may partially explain the lack of perceived utility of evaluations – and hence of utilisation.

3.4 On evaluating effectiveness

As noted in section 2, the review of evaluations conducted for the SOHS 2018 report highlighted inconsistencies in how 'effectiveness' is understood and evaluated. At first, the concept appears simple: effectiveness is a measure of the extent to which the objectives of a given intervention have been achieved in practice. But problems arise partly in the very different ways in which objectives are formulated; in uncertainty as to what constitutes an appropriate measure of achievement and causal attribution; and in the challenge of actually measuring the effects of interventions in potentially complex and fluid operating environments. Compared to the wider literature on development and public policy evaluation, much less attention has been given to this subject in the humanitarian sector – despite the very different nature of the goals and processes involved. [22]While the subject of effectiveness has received considerable policy attention recently in the humanitarian sector (particularly since the World Humanitarian Summit in 2016),[23] the concept itself and in particular the *evaluation* of effectiveness has not had the same level of focus.

The logic of humanitarian action



The figure shows three potential outcomes of the crisis (starting conditions), two of which (feared and intended outcomes) are hypothetical (shown by the dotted lines) and result from 'No intervention' and 'Intervention', respectively. According to this model, 'effectiveness' can be understood in terms of the relationship between a given intervention and the feared, intended and actual outcomes. Broadly speaking, the aim of humanitarian action is that actual outcomes should be as far as possible from the feared (adverse) outcomes and as close as possible to the intended (more positive) outcomes. An intervention is effective, in the broadest sense, to the extent that it helps to achieve this.

Source: author's own.

[22] In the development sector, much of this attention has been prompted by the growing emphasis on demonstrating results (often linked to the Millennium Development Goals and the Sustainable Development agendas) and more generally on aid effectiveness and the evaluation of aid programmes against commitments in the 2005 Paris Declaration on Aid Effectiveness.
[23] See for example OCHA (2016), and the Save the Children Humanitarian Effectiveness Project.

The logic of humanitarian action

Unpacking this question of evaluating effectiveness requires some consideration of the basic logic of humanitarian action. 'Humanitarian crisis' can be understood in terms of a set of conditions that are likely, without intervention, to result in catastrophic or highly adverse outcomes for large numbers of people. Although overly reductive, the 'core' humanitarian agenda is typically framed in terms of saving lives, preventing excess mortality and morbidity, preventing destitution and, increasingly, protecting civilians,

displaced people and vulnerable groups. The main feared outcomes relate to threats to life, health, livelihoods and security, and may have already materialised for some; humanitarian action is designed to prevent the escalation or continuance of these feared outcomes, and to mitigate or reverse their effects (Box 1).[24]

On this basis, the logic of humanitarian intervention, typically described in terms of meeting needs, may be better expressed in terms of risk and outcomes: of averting (or making less likely) certain feared outcomes and ensuring (or making more likely) more favourable outcomes. But a given humanitarian intervention is only one of many factors (and may be one of many interventions) that contribute to the outcomes in question, and the causal chain is rarely as simple or as linear as suggested by the figure in Box 1. Aid interventions often combine in complex ways with other contextual factors to affect behaviours and outcomes. This is particularly the case in complex political crises, where the key determinants of humanitarian outcomes may be political, military or socioeconomic factors over which the humanitarian system has limited influence. Evaluating the space for humanitarian action – and what use is made of that space – becomes key in such situations.

Working with real-world complexity

Consideration of real-world outcomes – at least proximate outcomes – is surely essential to any meaningful conception of humanitarian effectiveness. But effectiveness cannot simply be defined as the difference between the starting conditions and the actual outcomes after intervention. Situations evolve (negatively or positively) *without* such interventions and there are multiple potential factors involved, including affected communities' own coping strategies and behaviours. Rather, an intervention's effectiveness might be measured in terms of (its contribution to) the difference between the feared and the actual outcome, or in terms of the gap between the actual and intended outcomes.

Given this complexity, evaluating effectiveness is often difficult. And yet it lies at the heart of humanitarian evaluation. Sometimes the effectiveness of an intervention may be stated in terms of a reduction in known risk factors for certain outcomes – for example a reduction of insanitary conditions as a known risk factor for acute diarrhoeal disease. Proxy indicators of this kind are also likely to be used in the assessment of needs and vulnerabilities, pending more precise outcome data. The use of proxy indicators is often justified in chaotic and fast-moving humanitarian contexts, particularly where access is restricted. Yet they remain *proxies* – things that can be measured and which are assumed to be necessary factors in the achievement of a positive outcome, or adequate signifiers of the achievement of that outcome. Given the multiplicity of factors involved, this logic needs to be tested to the extent possible against *actual* outcomes, by measuring relevant outcome indicators through monitoring, surveys, community consultation, etc.[25] 'Effectiveness' is otherwise likely to remain subjective, a matter of opinion rather than something demonstrable with a reasonable degree of certainty.

[24] See Knox Clarke and Darcy (2014) on the underlying propositional logic involved. The related propositions in this case take the form 'If we do not intervene, the outcome will be X' (feared) and 'If we do intervene, the outcome will be Y' (intended). Much of the humanitarian evidence-gathering agenda is concerned with testing variants of these two kinds of proposition.

[25] This concern with testing the logic of interventions against real-world indicators cannot be confined to evaluation. It needs to happen in real time, to allow for adjustments to the intervention where the intended outcome is not being realised for whatever reason. This in turn depends on effective monitoring. On this subject, see further recent ALNAP papers from Dillon (2019); Dillon & Sundberg (2019); Sundberg (2019).

The challenges to evaluating effectiveness are considerable, and we should not burden evaluations with unrealistic expectations in this regard. The precise effect of a given humanitarian intervention (except perhaps for interventions like immunisation or basic commodity transfers) will often be indeterminable, as both points of reference – both feared and intended outcomes – are hypothetical and actual outcomes may be hard to measure.[26]Evaluators will usually aim to produce an assessment of effectiveness that is robust (i.e. as certain as it can be) but less than precise. What is possible and appropriate in this regard, given the context, has to be decided for each evaluation.

It is important to distinguish here between proximate or short-term effects and ultimate effects. Evaluating an intervention's ultimate effects (equivalent to 'impact' in standard evaluation terminology) is not generally part of the evaluation terms of reference for humanitarian evaluations. Yet the ambition for humanitarian intervention is often greater than meeting basic needs or averting critical risks, particularly in protracted crises lasting many years. In particular, an increasing ambition of multi-year programmes is that the crisis-affected population be made more *resilient* to future risks of this kind. Evaluating the effectiveness of resilience interventions is potentially an even more hypothetical exercise than that suggested by the figure in Box 1. As a minimum, it depends on the existence of a framework of sub-objectives and indicators to allow progress towards the ultimate goal to be assessed. While resilience in general may be hard to define, resilient water systems and even resilient livelihoods might be considered in quite tangible terms.

Reframing the effectiveness question

Effectiveness must have as its ultimate reference point something that is external to the intervention in question – that is, a measurable *effect* in the real world, relating to a change in people's situation. That said, a robust argument based on the (evaluated) logic of the intervention coupled with the analysis of relevant proxy indicators is still far preferable to what is often presented in evaluations. Too often, the effectiveness question is answered in essentially self-referential terms, based on the delivery of outputs, perhaps the application of quality standards, and *assumptions* about resultant effects based on a programme's own logic. And while 'effectiveness' may be understood in terms of achieving stated objectives, unless those objectives are themselves clearly defined in terms of external change (i.e. the intended outcomes of intervention), the evaluation of effectiveness is in danger of becoming a circular exercise.

Evaluations can go some way to answering questions about effectiveness more fully with reference to external change, but there are practical limitations to the evaluation process and the available (secondary) data on which it relies. With regard to effectiveness, the function of evaluation may therefore lie in playing a facilitative role as much as a technical one, asking the organisation concerned, 'How do *you* know whether the programme has been effective?' and prompting a discussion about programme monitoring and responsive practice.

4. Future directions

Bridging the gap between expectations and reality, and improving the utility (and use) of evaluations, will require continued effort on the part of evaluation offices and continued commitment from donors.

Evolving current practice in three specific ways could help significantly:

- Promote a fuller understanding of the functions of evaluation, beyond simply 'learning' and 'accountability', with greater emphasis on the facilitative and validation functions
- Strengthen the links (theoretical and practical) between the monitoring and evaluation functions in organisations and promoting a more externally-grounded idea of 'effectiveness'
- Collaboration to strengthen inter-comparability of evaluation results and the coverage of system-wide and 'whole-of-context' issues.

[26] Evaluating the validity of ex-post counter-factual arguments ('if we had not done this, the result would have been X') is notoriously difficult in the absence of adequate comparators. But we should be alert to the possibility that an adverse outcome that it is claimed has been averted was unlikely to have occurred in reality; or that a positive outcome would have occurred even without a given intervention. We should also be alert to (positive or negative) unintended as well as intended effects of intervention even where wider impact is not part of the evaluation scope. This implies a willingness to evaluate 'effect' rather than just effectiveness.

Achieving this requires both the evolution of policy and practice within individual agencies – perhaps even a culture shift – and stronger inter-agency collaboration on this agenda. Above all, it depends on creating an enabling culture within organisations (see below) and related dialogue between evaluation offices, monitoring and evaluation staff, and those responsible for humanitarian programming.

Here we briefly explore each of these three agendas in turn.

4.1 Promoting a fuller understanding the function of evaluation

The discussion in section 3 about evaluation function provokes several questions. Have we given enough thought to the variety of functions that evaluation can play within the humanitarian sector? Do evaluation producers and users alike give due consideration to the relative priority of facilitating dialogue versus measuring specific results; of investigating simple linear change processes versus probing broader systemic change? Can and should organisations do more to encourage a greater variety of evaluation types within their portfolios?

This paper shows that humanitarian evaluations have the potential to play a role that is not adequately captured by the dominant accountability and learning paradigm. As well as providing evidence for decision-makers, independent (external) evaluations can – and often do – provide an essential *validation* function, for which there is no adequate internal substitute.[27]

Essentially this function consists of testing and, where necessary, challenging the validity of the prevailing internal narrative – the stories we tell to and about ourselves – about a given intervention and the assumptions on which it is based. These shared narratives are often reflected in internal and external reports, and play an important role in uniting teams behind a common effort, helping create a sense of self-belief and worth. The importance of developing and having these shared narratives should not be underestimated; but the narratives themselves may be false or incomplete, distorted by the incentive to succeed and be seen to succeed. Evaluation has an important role to play in validating the claims and assumptions involved, not least from an accountability perspective.

A key part of the external evaluator's function is therefore to help organisations realise that alternative narratives may exist that may better reflect the reality of a situation and the role that the organisation has played or could play. This involves encouraging greater consideration of the reflective purpose of evaluation and means thinking of the evaluator as a peer practitioner rather than as a scientific or evaluation 'expert'. It means recognising that fostering critical thinking within programme teams is a legitimate (and arguably essential) evaluation product, above and beyond the final evaluation report. And it means valuing the exploration of competing perspectives and narratives on programme and project performance.

One implication of this view is that the key attribute of the external evaluator – rather than strict objectivity (all evaluators have their biases) – is their independence from the organisational policy and decision-making processes and their perspective as outsiders, free from the constraints of adhering to internal narratives. Impartiality, in the sense of being unattached to any particular narrative, is what is essential here.

Impartiality versus objectivity

While a dispassionate and unprejudiced review of evidence should be the cornerstone of any evaluation, subjectivity cannot be factored out of evaluations – just as it cannot from organisational decision making. Evaluators themselves will form judgements that are (implicitly) influenced by personal experience, individual preferences and perhaps cognitive biases. This is something that balanced evaluation teams, and the requirement for explicitly reasoned findings linked to evidence, are designed to counteract. The evidence itself may be largely qualitative and based to a significant

[27] This is not a new observation. See for example the 1999 OECD Guidance for Evaluating Humanitarian Assistance in Complex Emergencies, in its discussion of policy evaluation in the humanitarian context. 'Policy evaluations seek out the inherent tensions or contradictions in policy objectives, through tools such as discourse analysis and logic-of-argument analysis… [They involve] a process of 'validating' through argument, rather than 'verifying' through some 'scientific' process…' The subsequent push towards results measurement appears to have obscured this aspect of evaluation.

degree on triangulated or aggregated personal opinions (e.g. 'a majority of respondents felt that …'). The requirement for objectivity in evaluations – as distinct from impartiality and analytical rigour – therefore needs to be qualified. An approach that stresses the facilitative role of evaluators also needs to recognise and take due account of this element of subjectivity.

The validation function described is a challenging one in more than one sense, and it often involves implicit (and sometimes explicit) criticism of individuals' judgements and decision-making. To be directly useful, evaluation that validates should be done at a time when course correction is still possible. For it to be successful, the organisation needs to be open and willing to question narratives around which whole teams – or indeed a whole organisation – may have come together. It needs managers who believe in a learning culture and who positively encourage (even demand) such self-reflection of their teams and of themselves.[28] Ultimately, it is the culture of reflective and responsive practice, not the generation of evidence for its own sake, that is most likely to lead to appropriate, effective and adaptive humanitarian interventions.

4.2 Strengthening the links between monitoring and evaluation[29]

One of the recurrent themes from the interviews conducted for this paper was the continuing disconnect in practice between the monitoring and evaluation functions. This is despite the clear connections between them in terms of intended purpose – learning, accountability, informed decision-making – and despite the fact that responsibility for both is often located in the same organisational teams. Addressing this disconnect, we suggest, is essential to addressing some of the evidential weaknesses highlighted in this paper.

Evaluation and programme monitoring need to be seen as two parts of the same agenda, albeit with their own distinct functions within the programme cycle. Linking these two functions more explicitly, both in terms of concept and process, would emphasise the importance of using external measures of performance and could provide a stronger and more consistent evidence base for decision-makers throughout the programme cycle. One way in which monitoring and evaluation might be more strongly linked is by considering a more basic set of questions that can be used for both in-programme monitoring and for retrospective evaluations. The DAC evaluation criteria, as modified for use in humanitarian settings (Beck, 2006), provide an essential touchstone for evaluators. Yet from a programme perspective they can seem abstract, and using them to structure evaluations can lead to disconnected analyses. We propose a simple complementary framework comprised of four basic evaluative questions, asked from the perspective of the programming agency:

1. 'Have we done the right things?'
2. 'How well have we done them?'
3. 'Have they worked?'
4. And for each of these questions: 'How do we know?'

These basic questions give rise to several sub-questions that are both evaluable in principle and directly related to programme design, implementation and oversight. While compatible with the DAC criteria, these questions introduce the idea of programme *quality*, distinguish *outcomes* clearly from *outputs*, and emphasise *informed* programming.

One important advantage of this approach is that the same basic framework but with the questions shifted into the present tense – e.g. 'Are we doing the right things?' – provides a real-time template for managers concerned with the success of the intervention in question. In other words, these same questions provide a conceptual bridge between the basic concerns of situational and programme monitoring, on the one hand, and of evaluation on the other. Seeing these processes as part of a diagnostic continuum helps to emphasise the essential links between them. In practice, day-to-day management and operational concerns mean that the questions facing managers and staff on a daily basis are likely to be more immediate and practical. But unless organisations and inter-agency groups find ways to ask and answer these basic questions as the intervention proceeds, they are unlikely to be able to adapt programming as the situation demands.

[28] On the necessary elements of an organisational 'enabling culture' for evaluations, see Hallam and Bonino (2013)
[29] On this topic, and the related subject of outcome monitoring, see Dillon (2019); Dillon & Sundberg (2019)

From an evaluation perspective, unless at least proximate outcome data is being collected during the course of the programme, there will be limited information on which to base conclusions about core evaluation questions, particularly effectiveness. Yet these questions – of relevance, quality and effectiveness – need ultimately to be referred back to the organisation concerned. How does it assure its performance in these areas? How is effectiveness monitored?[30] Is there an adequate quality assurance mechanism in place? How is effectiveness monitored? If an organisation has no real way of answering the question 'Is it working?' with reference to external criteria – even if those are proxy indicators – then something is amiss in its monitoring process.

4.3 Improving evaluation comparability and coverage of system-wide performance

Earlier in this paper we noted the apparent importance of organisational factors – the type of commissioning agency and the choice of evaluator – in determining the evidential quality of an evaluation (section 2). This suggests that organisational approaches matter. Different agencies have varying levels of quality assurance, different capacities to conduct and manage evaluations, and different budgets with which to do so. Sharing good practice between agencies might go some way to reducing quality variance; for example, by replicating the UNEG Norms and Standards (2016) –which inform evaluation policies across the UN agencies – beyond the UN system.[31]

Beyond the question of evaluation quality, there are two particular issues that need to be highlighted here, arising in part from analysis of the SOHS evaluation synthesis results. The first is the lack of inter-comparable data and evidence available from evaluations conducted by different agencies in relation to the same context or theme. This issue was noted in, for example, the case of the Syria Coordinated Accountability and Lessons Learning initiative, an attempt to harmonise evaluation approaches in relation to the Syria crisis response that was frustrated in practice. Retrospective efforts to synthesise the results of individual agency evaluations have proved challenging because of the very different ways in which such evaluations have been framed and conducted, making results hard to compare (Darcy, 2016).

The lack of inter-comparable data and evidence from individual evaluations suggests the need for renewed efforts to harmonise evaluation approaches and related use of evaluation synthesis. But it also points to the second issue of concern here: the relative lack of evaluation coverage of system-wide agendas such as the humanitarian–development nexus. As discussed in section 2.2., this is one of the biggest gaps in coverage from recent evaluations, particularly those that have a single-agency or single-project focus. In compiling evidence for the SOHS 2018 report, it became apparent that there is relatively little evidence about the functioning of the humanitarian system *as a system*. Here, single-agency or single-project evaluations offer too narrow a focus.

This issue is surmountable. But doing so will require commissioning agencies to clearly and jointly signal their intent to increase and improve coverage of system-wide performance. Tools such as strategic or thematic evaluation, together with evaluation synthesis, can contribute to the solution. Improved mapping of evaluation evidence can provide a clearer overview of where the system as a whole is strong, and where it is less so.[32] And synthesis tools that make the best of the available information can reduce the need for duplicative studies and direct evaluation resources accordingly.

Likewise, the humanitarian sector must recognise the value of existing inter-agency, peer review and joint evaluation efforts and endeavour to learn from and replicate good practice. The Inter-agency Humanitarian Evaluation model provides at least some basis for collective accountability and learning, although it been too sporadically applied. The operational peer review and Peer-2-Peer mechanisms, for all their merits, are not a substitute: they focus mainly on inter-agency process and coordination issues based on 'internal' system criteria.

[30] It should be stressed here that questions of quality and effectiveness are intrinsically linked. An intervention that is poorly conducted is unlikely to be truly effective, and indeed may have unintended negative effects.
[31] There is a risk of a 'formulaic' approach here, for example that quality assurance is reduced to a list of tightly defined process criteria. We should foster rigour in the use of evidence and analysis, but not over-prescribe the evaluation process.
[32] See ALNAP *Evalmapper,* https://www.alnap.org/evalmapper (accessed 17 December 2020).

5. Conclusion

The review of evaluations conducted for the SOHS 2018 report highlighted a range of different issues in the humanitarian evaluation system, including variable quality, limited scope, inconsistent approaches to assessing effectiveness and poorly developed links between evaluation and monitoring systems. More generally, there appears to be a persistent gap between expectations and reality when it comes to the contribution of evaluation to the humanitarian system. All of this has an impact on evaluation usefulness and use, over which there are enduring concerns that demand serious attention.

The ALNAP Secretariat has worked to tackle some of these issues. In 2019, three publications in the Monitoring of Humanitarian Action series touched upon the quality and conduct of monitoring systems, including measurement of outcomes and linkages between monitoring and evaluation systems in the sector (Dillon, 2019; Dillon and Sundberg, 2019). ALNAP has developed a mapping tool for its HELP Library database of evaluations to map the geographic and thematic focus of humanitarian evaluations and to help identify gaps in humanitarian evaluation.[33]

The growing pressure on aid agencies to demonstrate results, although understandable in accountability terms, has led to a situation in which the success of a humanitarian intervention is often measured and reported predominantly in terms of 'numbers reached' (or similar quantitative indicators). The availability of reliable quantitative data is essential to effective management; but in practice, the figures involved tend to be based on outputs rather than outcomes, are often arbitrary and incomplete, and tend to lack reference to the wider context (e.g. scale of need).

As a basis for accountability, such figures are therefore often of limited value. Moreover, they tell only part of the effectiveness story. They imply rather than demonstrate a causal link to the achievement of intended outcomes; and they fail to reflect the quality of the intervention, or the lived experience of the intended beneficiaries. To establish the extent to which a project or programme was effective, an evaluator has to dig beneath the reported figures to evaluate the basis for claims of positive effect using other points of reference. In practice, however, there are often strict limits to what evaluation in itself can be expected to achieve in this respect.

However, we will only make humanitarian evaluations more useful and used if these discrete efforts are accompanied by greater reflection on the purpose and function of evaluation in the humanitarian system. At present, evaluations often appear to be conducted just because they are required as a matter of policy or contract, rather than with a clear purpose in mind. Evaluation policies, individual terms of reference, contracting and quality assurance processes all tend to be based on an understanding of evaluation as providing accountability and learning, yet those terms are often undefined.

In particular, the *validation* function played by independent evaluation is often under-recognised. Properly understood, it can help organisations reflect critically on prevailing internal narratives and posit alternative accounts of an organisation's role and effect on the external environment. Validation demands more of a facilitative (rather than a technical) approach to evaluation – although technical (and quantitative) evaluations remain an essential means of generating evidence on the effectiveness of specific interventions and approaches.

It follows that the evaluator's role may often be better understood as one of facilitating organisational reflection and dialogue rather than one of measuring results. One implication of this view is that rather than the external evaluator's key attribute being strict objectivity, it is their independence from the organisational processes and their perspective as outsiders, free from the constraints of adhering to internal narratives.

[33] See ALNAP *Evalmapper,* https://www.alnap.org/evalmapper (accessed 17 December 2020).

Finally, when looking across the breadth of humanitarian evaluations currently available, there is a clear lack of evidence on system-wide performance. This gap is particularly keenly felt when it comes to coverage of affected populations, the humanitarian–development–security nexus and inter-agency coordination. Commissioning agencies should make a clear joint signal of intent to make progress in this area: a greater focus on evaluation synthesis, system-wide evaluation mapping, inter-agency, peer review and joint evaluation would all help increase coverage. And doing so would allow the evaluation community to contribute its part to collective accountability, learning and joint reflection across the humanitarian system.

Annex: Interview list

Name	Organisation	Position
Jo Abbotts	UK Department of International Development	Joint head Humanitarian, Security and Migration Division
Sarah Bailey	Independent	Evaluator
Julia Betts	Independent	Evaluator
Steve Darwill	Australian Department of Foreign Affairs and Trade	Director, Humanitarian Reform and Performance Section
Gaby Duffy	World Food Programme	Evaluation Officer
Marie Gaarder	3ie	Head of Evaluation
Josse Gillijns	International Federation of Red Cross and Red Crescent Societies	Head, Planning, Monitoring, Evaluation and Reporting
Nicola Giordano	Action Against Hunger	Director of MEAL Service
Scott Green	United Nations Office for the Coordination of Humanitarian Affairs	Senior Evaluation Officer
Alistair Hallam	Valid Evaluations	Evaluator
David Heath	Global Affairs Canada	Director, International Assistance Evaluation
Hélène Juillard	Independent	Evaluator
Peter Klansoe	Danish Refugee Council	Head of Programme Division
Jane Mwangi	UNICEF	Evaluation Specialist
Antoine Ouellet-Drouin	International Committee of the Red Cross	Head of Sector, Planning, Monitoring & Evaluation
Anke Reiffenstuel	German FFO	Head of Division Humanitarian Assistance and Operations
Thomas Schwandt	University of Illinois	Professor
Marco Segone	United Nations Population Fund	Head of Evaluation
Ritu Shroff	United Nations High Commissioner for Refugees	Head of Evaluation
Maria Thorin	Swedish International Development Cooperation Agency	Humanitarian Desk Officer
Vivien Walden	British Red Cross	Senior PMEAL Adviser

Bibliography

The following publications can also be accessed via the Humanitarian Evaluation Learning and Performance (HELP) Library: https://www.alnap.org/help-library/missing-the-point-biblio

All entries with a * correspond to evaluation policies analysed for this paper.

* Action Against Hunger. (2011) Evaluation policy and guidance: Enhancing organisational practice through an integrated evaluations, learning & accountability framework. Paris: ACF. (www.alnap.org/help-library/evaluation-policy-and-guidance-enhancing-organisational-practice-through-an-integrated).

ALNAP. (2018) The state of the humanitarian system. ALNAP Study. London: ALNAP/ODI. (www.alnap.org/help-library/the-state-of-the-humanitarian-system-2018-full-report).

Bamberger, M. (2012) Introduction to mixed methods in impact evaluation (no.3). Massachusetts: InterAction. (www.alnap.org/help-library/introduction-to-mixed-methods-in-impact-evaluation).

Beck, T. (2006) Evaluating humanitarian action using the OECD-DAC criteria. London: ALNAP/ODI. (www.alnap.org/help-library/evaluating-humanitarian-action-using-the-oecd-dac-criteria).

Betts, J. and Wedgewood, H. (2011) 'Effective institutions and good governance for development: Evidence on progress and the role of aid'. Evaluation Insights, 4. Paris: OECD. (www.alnap.org/help-library/effective-institutions-and-good-governance-for-development-evidence-on-progress-and-the).

Bovens, M., t'Hart, P. and Kuipers, S. (2008) 'The politics of policy evaluation', in Goodin, R. E., Moran, M. and Rein, M. (eds) The Oxford Handbook of Public Policy. Oxford: Oxford University Press. (www.alnap.org/help-library/the-politics-of-policy-evaluation).

Cosgrave, J., Buchanan Smith, M. and Warner, A. (2016) Evaluation of humanitarian action guide. London: ALNAP/ODI. (www.alnap.org/help-library/evaluation-of-humanitarian-action-guide).

Dahler-Larsen, P. (2012) The Evaluation Society. Stanford: Stanford University Press. (www.alnap.org/help-library/the-evaluation-society).

Darcy, J. (2016) Evaluation synthesis and gap analysis. Syria coordinated accountability and lessons learning (CALL) initiative. New York: UN OCHA. (www.alnap.org/help-library/syria-coordinated-accountability-and-lesson-learning-call-evaluation-synthesis-and-gap).

* DFAT. (2017) DFAT aid evaluation policy. Canberra: DFAT. (www.alnap.org/help-library/dfat-aid-evaluation-policy).

DFAT-MFAT. (2018) MFAT-DFAT humanitarian monitoring and evaluation framework for the Pacific. Canberra/Wellington: DFAT/MFAT. (www.alnap.org/help-library/mfat-dfat-humanitarian-monitoring-and-evaluation-framework-for-the-pacific).

* DFID. (2013) International development evaluation policy. London: DFID. (www.alnap.org/help-library/international-development-evaluation-policy).

Dillon, N. (2019) Breaking the mould: Alternative approaches to monitoring and evaluation. London: ALNAP/ODI. (www.alnap.org/help-library/breaking-the-mould-alternative-approaches-to-monitoring-and-evaluation).

Dillon, N. (2020) Learning from what we know: How to improve evaluation synthesis for humanitarian organisations. London: ALNAP/ODI. (www.alnap.org/help-library/learning-from-what-we-know-how-to-improve-evaluation-synthesis-for-humanitarian).

Dillon, N. and Sundberg, A. (2019). Back to the drawing board: How to improve monitoring of outcomes. London: ALNAP/ODI. (www.alnap.org/help-library/back-to-the-drawing-board-how-to-improve-monitoring-of-outcomes).

* DRC. (2015) Evaluation policy. Copenhagen: DRC. (www.alnap.org/help-library/evaluation-policy).
Engel, P., Carlsson, C. and van Zee, A. (2003) Making evaluation results count: Internalising evidence by learning. Maastricht: ECDPM. (www.alnap.org/help-library/making-evaluation-results-count-internalising-evidence-by-learning).

Galport, N. and Azzam, T. (2016) 'Evaluator training needs and competencies: a gap analysis'. American Journal of Evaluation, 38(1): 80–100. (www.alnap.org/help-library/evaluator-training-needs-and-competencies-a-gap-analysis).

Gasper, D. and Apthorpe, R. (1996) 'Introduction: discourse analysis and policy discourse'. European Journal of Development Research, 8(1): 1–15. (www.alnap.org/help-library/introduction-discourse-analysis-and-policy-discourse).

Hallam, A. (2011) Harnessing the power of evaluation in humanitarian action. ALNAP Working Paper. London: ALNAP/ODI. (www.alnap.org/help-library/harnessing-the-power-of-evaluation-in-humanitarian-action-an-initiative-to-improve).

Hallam, A. and Bonino, F. (2013) Using evaluation for a change: Insights from humanitarian practitioners. ALNAP Study. London: ALNAP/ODI. (www.alnap.org/help-library/using-evaluation-for-a-change-insights-from-humanitarian-practitioners).

House, E.R. and Howe, K.R. (1999) Values in Evaluation and Social Research. Thousand Oaks: Sage. (www.alnap.org/help-library/values-in-evaluation-and-social-research).

IAHE – Inter-Agency Humanitarian Evaluations. (2018) Inter-Agency Humanitarian Evaluations: Process guidelines. Geneva: IAHE. (www.alnap.org/help-library/inter-agency-humanitarian-evaluations-process-guidelines).

* ICRC. (2009) Measuring results. Geneva: ICRC. (www.alnap.org/help-library/measuring-results).
* IFRC. (2011a) IFRC framework for evaluations. Geneva: IFRC. (www.alnap.org/help-library/ifrc-framework-of-evaluation).

IFRC. (2011b) Project/programme monitoring and evaluation (M&E) guide. Geneva: IFRC. (www.alnap.org/help-library/projectprogramme-monitoring-and-evaluation-me-guide).

* IRC. (n.d.) Research, evaluation and learning at the International Rescue Committee. New York: IRC. (www.alnap.org/help-library/research-evaluation-and-learning-at-the-international-rescue-committee).

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/help-library/insufficient-evidence-the-quality-and-use-of-evidence-in-humanitarian-action-alnap-0).

Liverani, A. and Lundgren, H. (2007) 'Evaluation systems in development aid agencies: an analysis of DAC peer reviews 1996-2004'. Evaluation, 13(2): 242. (www.alnap.org/help-library/evaluation-systems-in-development-aid-agencies-an-analysis-of-dac-peer-reviews-1996).

Multilateral Organisations Performance Assessment Network. (2018) MOPAN 3.0 methodology manual. Paris: MOPAN. (www.alnap.org/help-library/mopan-30-methodology-manual).

* NRC – Norwegian Refugee Council. (2015) NRC evaluation policy. Oslo: NRC. (www.alnap.org/help-library/nrc-evaluation-policy).

OECD. (1999) Guidance for evaluating humanitarian assistance in complex emergencies. Paris: OECD. (www.alnap.org/help-library/guidance-for-evaluating-humanitarian-assistance-in-complex-emergencies).

* Oxfam. (n.d.) Oxfam GB evaluation guidelines. Oxford: Oxfam. (www.alnap.org/help-library/oxfam-gb-evaluation-guidelines).

Patton, M.Q. (2010) 'Incomplete success'. Canadian Journal of Program Evaluation, 25(3). (www.alnap.org/help-library/incomplete-successes).

Power, M. (1997) The Audit Society: Rituals of Verification. Oxford: Oxford University Press. (www.alnap.org/help-library/the-audit-society-rituals-of-verification).

Prewitt, K., Schwandt, T.A. and Straf, M.L. (2012) Using Science as Evidence in Public Policy. Washington, D.C.: The National Academic Press. (www.alnap.org/help-library/using-science-as-evidence-in-public-policy).

Pritchett, L., Samji, S. and Hammer, J. (2013) It's all about meE: using structure experiential learning ('e') to crawl the design space. Washington, D.C.: Center for Global Development. (www.alnap.org/help-library/it%E2%80%99s-all-about-mee-using-structure-experiential-learning-%E2%80%9Ce%E2%80%9D-to-crawl-the-design-space).

Raimondo, E. (2018) 'The power and dysfunctions of evaluation systems in international organisations'. Evaluation, 24(1). (www.alnap.org/help-library/the-power-and-dysfunctions-of-evaluation-systems-in-international-organisations).

Rogers, P. (2009) 'Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation'. Journal of Development Effectiveness, 1(3): 217-226. (www.alnap.org/help-library/matching-impact-evaluation-design-to-the-nature-of-the-intervention-and-the-purpose-of).

Sandison, P. (2006) 'The utilisation of evaluations', in ALNAP Review of Humanitarian Action. Evaluation Utilisation. London: ALNAP/ODI. (www.alnap.org/help-library/the-utilisation-of-evaluations-alnap-review-of-humanitarian-action-in-2005-evaluation).

* Save the Children. (2012) Evaluation handbook. London: Save the Children. (www.alnap.org/help-library/evaluation-handbook).

Schwandt, T.A. (2015) Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice. Stanford: Stanford University Press. (www.alnap.org/help-library/evaluation-foundations-revisited-cultivating-a-life-of-the-mind-for-practice).

Scriven, M. (1991) Evaluation Thesaurus. 4th ed. Thousand Oaks: Sage. (www.alnap.org/help-library/evaluation-thesaurus).

Segone, M. (ed) (2012) Evaluation for equitable development Results. New York: UNICEF Evaluation Office. (www.alnap.org/help-library/evaluation-for-equitabledevelopment-results).

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B. (2012) 'Broadening the range of designs and methods for impact evaluations'. DFID Working Paper, 38. London: DFID. (www.alnap.org/help-library/broadening-the-range-of-designs-and-methods-for-impact-evaluations-report-of-a-study).

Stephens, A., Lewis, E. D. and Reddy, S. M. (2018) Inclusive systemic evaluation (ISE4GEMs): A new approach for the SDG era. New York: UN Women. (www.alnap.org/help-library/inclusive-systemic-evaluation-ise4gems-a-new-approach-for-the-sdg-era).

Sundberg, A. (2019) Beyond the numbers: How qualitative approaches can improve monitoring of humanitarian action. London: ALNAP/ODI. (www.alnap.org/help-library/beyond-the-numbers-how-qualitative-approaches-can-improve-monitoring-of-humanitarian).

Swedish International Development Agency. (2018) Evaluation at Sida: Annual Report 2018. Stockholm: Sida. (www.alnap.org/help-library/evaluation-at-sida-annual-report-2018).

* Swedish International Development Agency. (2020) Sida's evaluation handbook guidelines and manual for conducting evaluations at Sida. Stockholm: Sida. (www.alnap.org/help-library/sida%E2%80%99s-evaluation-handbook-guidelines-and-manual-for-conducting-evaluations-at-sida).

United Nations Evaluation Group. (2011) UNEG Framework for professional peer reviews of the evaluation function of UN organisations. New York: UNEG. (www.alnap.org/help-library/uneg-framework-for-professional-peer-reviews-of-the-evaluation-function-of-un).

United Nations Evaluation Group. (2012) Update note on peer reviews of evaluation in UN organisations/UNEG-DAC Peer Reviews. New York: UNEG. (www.alnap.org/help-library/update-note-on-peer-reviews-of-evaluation-in-un-organizations).

United Nations Evaluation Group. (2016) UNEG norms and standards for evaluation. New York: UNEG. (www.alnap.org/help-library/uneg-norms-and-standards-for-evaluation).

* UNFPA – United Nations Population Fund. (2013) Revised UNFPA evaluation policy. New York: UNFPA. (www.alnap.org/help-library/revised-unfpa-evaluation-policy).

* UNHCR. (2016) Policy on evaluation. Geneva: UNHCR. (www.alnap.org/help-library/policy-on-evaluation).

UNICEF. (2010) Bridging the gap: The role of monitoring and evaluation in evidence-based policy-making. New York: UNICEF. (www.alnap.org/help-library/bridging-the-gap-the-role-of-monitoring-and-evaluation-in-evidence-based-policy-making).

* UNICEF. (2018) Revised evaluation policy for UNICEF. New York: UNICEF. (www.alnap.org/help-library/revised-evaluation-policy-for-unicef).

* UN OCHA. (2010) Policy instruction: Evaluations. New York: UN OCHA. (www.alnap.org/help-library/policy-instruction-evaluations).

UN OCHA. (2016) Leaving no-one behind: Humanitarian effectiveness in the age of the sustainable development goals. New York: UN OCHA. (www.alnap.org/help-library/leaving-no-one-behind-humanitarian-effectiveness-in-the-age-of-sustainable-development).

Vogel, I. (2012) Review of the use of 'theory of change' in international development. London: DFID. (www.alnap.org/help-library/review-of-the-use-of-theory-of-change-in-international-development).
* WFP. (2015) Evaluation policy (2016-2021). Rome: WFP. (www.alnap.org/help-library/evaluation-policy-2016-2021).

Williams, B. (2016) Systemic evaluation design: A workbook. (www.alnap.org/help-library/systemic-evaluation-design-a-workbook).

World Bank Independent Evaluation Group. (2009) Institutionalising impact evaluation within the framework of a monitoring and evaluation system. Washington, D.C.: World Bank IEG. (www.alnap.org/help-library/institutionalising-impact-evaluation-within-the-framework-of-a-monitoring-and).