

**Joint humanitarian impact
evaluation: options paper**

**To be presented at the 25th ALNAP
Meeting, London, 18th November
2009**

Tony Beck

**Commissioned by
Evaluation and Studies Section
United Nations
Office for the Coordination of
Humanitarian Affairs**

Summary

Background and consultation process

This paper, commissioned by OCHA, sets out a process for facilitated consultation over six to nine months on options for joint humanitarian impact evaluation. Consultation is proposed in-country with the following stakeholders: the affected population; local NGOs; the government; the UN Country Team; clusters; donors; and at the regional level. Consultation is also proposed at the international level with: ALNAP members and observers; evaluation bodies; and donors at headquarters. It is recommended that OCHA coordinate the consultation process by forming and chairing a working group on joint impact evaluation. The working group would guide the consultation process, synthesize the consultation results, and make recommendations on the future of joint impact evaluation. A short background brief on each option in this paper, a set of key discussion points, and a method for capturing and writing up the findings from each consultation process, will need to be completed.

Options

The following options are recommended for discussion during the consultation process:

The purpose of joint impact evaluation needs to be clearly defined before proceeding, so the intended purpose, use and users, and related evaluation questions, should be the first area for discussion:

Option 1: Conceptual Framework and Purpose

Should the purpose be mainly summative/judgment oriented; and/or mixed summative and lesson learning about specific programmes; and/or generalized knowledge generation?

Option 2: Focus - Institution-oriented, affected population- oriented, or both?

Should the scale be the entire system, or parts of the system?

Option 3: Method - Emphasis on quantitative, qualitative, or a combination?

If a combination is chosen, how will this work in practice? There is a major debate on this issue so it is important to discuss it in depth.

Option 4: Capacity

Use of national or international evaluators, or both?

Option 5: Piloting impact evaluation in humanitarian action

It has been suggested that the system could proceed with two joint impact evaluation pilots after the consultation period. Where might this be undertaken?

Option 6: Coordination

A common framework approach? How “joint” should the evaluation be, who will participate, and how?

Option 7: Optimal management arrangements for joint impact evaluation

A dual structure involving a Steering Committee that provides overall guidance and a management group that runs the evaluations on a day-to-day basis?

Table of contents

1. Background	1
1.1 Why joint impact evaluation?	
1.2 Joint impact evaluation as part of a monitoring and evaluation system	
1.3 About this paper	
2. Consultation processes	4
3. Options for discussion during the consultation period	5
Option 1: Conceptual framework and purpose	5
Option 2: Focus	8
Option 3: Method	8
Option 4: Capacity	11
Option 5: Piloting	11
Option 6: Coordination	12
Option 7: Management arrangements	13
Table 1: Types of consultation: in two countries	4
Table 2: Types of consultation: international	4
Table 3: Main users and questions related to evaluation purposes	5
Table 4: Timing for a year-long impact evaluation	11
Bibliography and Annexes	
Bibliography	<i>i</i>
Annex 1 Options Paper Terms of Reference	<i>iv</i>
Annex 2 Interviewees	<i>vii</i>
Annex 3 Sample Consultation Questions	<i>viii</i>
Annex 4 Purposes and methods for joint humanitarian impact evaluation	<i>ix</i>

1. Background

Several agencies, including ECB, OCHA and UNICEF, have been discussing the possibility of joint humanitarian impact evaluation for some time. In June 2009 an OCHA commissioned paper on *Evaluability Assessment for Impact Evaluation of the Humanitarian System at the Country Level* was presented at a workshop looking at the future of interagency evaluations. Participants indicated support for further discussion, and requested OCHA commission an options paper on joint humanitarian impact evaluation for discussion at the 25th ALNAP meeting in London in November 2009. See Annex 1 for this paper's terms of reference. (1)

The purpose of the current paper is to set out a process for facilitated consultation, over six to nine months, on options for joint humanitarian impact evaluation. It is proposed that consultation take place from the affected population to agency HQ on the different options presented below.

After discussing proposed consultation processes in Section 2, the paper sets out options related to:

Joint impact evaluation purpose/concept

- The different purposes of impact evaluation and implications for the proposed joint impact evaluation, as well as key evaluation questions linked to different purposes.

Evaluation scale and methodology

- How impact might be approached at various levels (e.g. system-wide, humanitarian reform effort, cluster approach, Consolidated Appeals Process (CAP)).
- Methodologies that could be employed.

Coordination, management and governance arrangements

- Suggested coordination mechanisms and roles and responsibilities of management and governance bodies.

1.1 Why joint impact evaluation?

The humanitarian system is attempting to be more coherent at a number of levels – through clusters, pooled funding arrangements, principles for partnership, with partner country governments, and the Emergency Capacity Building (ECB) project. The reasons for carrying out joint evaluation are similar to those for carrying out other joint activities: allowing for a broader view of operations and potentially drawing conclusions on an entire humanitarian operation; more effective coordination and use of resources; lower transactions costs for partner countries; and a higher quality of performance. Donors are also promoting joint evaluation as part of the Good Humanitarian Donorship and Paris Declaration processes (e.g. DFID 2009; ECHO 2007). There is also evidence to suggest that joint evaluations are of higher quality than single agency evaluations (Beck and Buchanan-Smith 2008). (2)

Joint impact evaluation has the potential to look at areas that cannot easily be evaluated in single agency evaluations, for example the overall results of multiple interventions vis-à-vis the affected population, and the ways in which coordination and management processes have led to these results. Beck and Buchanan-Smith throw in words of caution (2008: 85): “The sector is on a steep learning curve in terms of how to do joint evaluations, including how best to manage and organize them, when they are appropriate, and with whom.” Transaction costs in joint evaluations are high and usually underestimated, and increase with the numbers of actors involved.

1 Annexes and the bibliography are provided in a separate Word file.

2 Reviews of lessons from carrying out joint evaluations can be found in the development and humanitarian sectors can be found in Beck (2009) and Beck and Buchanan-Smith (2008).

Joint impact evaluation is likely to be particularly challenging in humanitarian contexts because of issues of lack of data, access, and security. Joint impact evaluation, like all evaluation in the humanitarian sector, needs to proceed using the principle of do no harm, and ensure that the affected population is not subject to any risks because of the evaluation process.

1.2 Joint impact evaluation as part of a monitoring and evaluation system

Impact evaluation is one piece of a larger framework which needs to be in place to assess results and improve programming. As Bamberger (2009: 8) notes: “A successful IE [impact evaluation] program can only be achieved when it is part of a broader M&E system. It would not make sense, or even be possible, to focus exclusively on IE without building up the monitoring and other data-collection systems on which IE relies. Although IEs are often the highest profile (and most expensive) evaluations, they only provide answers to certain kinds of questions; for many purposes, other kinds of evaluation will be more appropriate.” Joint impact evaluation will be more effective if it is linked to strengthened needs assessments and monitoring, and to other evaluation processes such as formative evaluations and Inter-Agency Real-Time Evaluations (RTEs). This will also help meet the different information needs of stakeholders.

1.3 About this paper

The review of background literature and interviews carried out for the earlier *Evaluability Assessment* were supplemented with further literature review and interviews – see Annex 2 for those interviewed. Interviewees strongly supported the idea of stakeholder consultation concerning joint impact evaluation. As one respondent put it: “It’s well worth having the conversation”. (3)

3 Thanks for comments on the paper to Kimberly Lietz (OCHA, project manager), Scott Green (OCHA), John Mitchell (ALNAP), and Karen Proudlock (ALNAP).

2. Consultation processes

There was broad agreement among respondents about the importance of early and adequate consultation for impact evaluation (see also Jones *et al* 2009; TEC 2006; NONIE 2009). Consultation is required both to determine if joint humanitarian impact evaluation should be taken forward, and on design and implementation issues. Wherever possible, consultation should fit into other coordination processes; and will need to take into account capacity available for facilitation.

Consulting with host governments should take place where feasible. One criticism of the Tsunami Evaluation Commission (TEC) process concerned lack of affected-country involvement. (4) Beck and Buchanan-Smith (2008) found that involvement of government and national institutions in joint evaluations may be no better than for single-agency evaluations, because of: suspicion and lack of trust between international agencies capacity in particular in complex emergencies; lack of evaluation capacity in the governments of affected countries; and the focus of much EHA on international agency performance.

It will be important to consult with donors, as they have been promoting evaluations focusing on results. When donors ask for “impact evaluation” they may not always be clear what it is they are requesting; and donors’ information needs may be best met by a range of monitoring and evaluation reporting, of which impact evaluation is one part.

Despite agreement that it is crucial, consultation with the affected population is one of the weakest areas of humanitarian evaluation; so **it is important to include affected women, men, girls and boys from the earliest stage possible.**

Consultation processes are set out in Tables 1 and 2, and are planned over 6-9 months. Country-level consultation would preferably take place in two countries where there is relevant OCHA capacity to facilitate this. Needless to say consultation will need to be carefully planned and implemented:

- It is recommended that OCHA coordinate the consultation process by forming and chairing a working group on joint impact evaluation. The working group would guide the consultation process, synthesize the consultation results, and make recommendations on the future of joint impact evaluation.
- A short background brief on each option in this paper, a set of key discussion points, and a method for capturing and writing up the findings from each consultation process, will need to be completed. An example of discussion points is given in Annex 3.

4 [http://www.alnap.org/pool/files/tecsurvey\(2\).pdf](http://www.alnap.org/pool/files/tecsurvey(2).pdf)

Table 1: Types of consultation: in two countries

Stakeholder	Type of consultation	Facilitated by	Cost
Affected population	Focus groups and key stakeholder interviews using participatory methodologies in six communities in each country	INGO evaluation offices	Internal
Local NGOs	As part of ongoing meetings or fora between ALNAP members and local partners	INGO evaluation offices	Internal
National and local government	Meetings with planning/finance ministries, humanitarian departments	HCT members	Internal
Clusters	As part of regular cluster meetings	Cluster members	Internal
UNCT/HCT	During UNCT/HCT coordination meetings	OCHA	Internal
Donors	2 donor forums (e.g. as part of donor coordination meetings)	OCHA/agency evaluation offices (e.g. donors in country)	Costs of up to \$8,000 (e.g. for a workshop), or internal
Regional Level	During UN Regional Director meetings	OCHA	Internal

Table 2: Types of consultation: international

Stakeholder	Type of consultation	Facilitated by	Cost
ALNAP members and observers (including UN agencies, IFRC, INGOs, academics and research institutes, independent evaluators)	Moderated online forum; use of social media; virtual seminars with invited contributors such as Tufts, 3ie, and southern evaluation associations	ALNAP – this may be an appropriate forum to discuss some of the conceptual issues around joint humanitarian impact evaluation	Internal resources
Evaluation bodies (UNEG, 3ie, AEA); Paris Declaration evaluation Phase 2	Email and as part of ongoing coordination function	OCHA	Internal resources
Donors	2 donor forums, e.g. tied in with donor consultation mechanisms such as MOPAN	OCHA/ALNAP	Costs of up to \$8,000 (e.g. for a workshop), or internal (e.g. via MOPAN)

3. Options for discussion during the consultation period

Option 1: Conceptual framework and purpose

Should the purpose be mainly summative/judgment oriented; and/or mixed summative and lesson learning about specific programmes; and/or generalized knowledge generation?

The first step in designing joint impact evaluation is to agree on the conceptual framework and purpose of the evaluation. One lesson from the TEC is that without an agreed conceptual framework it becomes more challenging to ensure a common understanding of the evaluation purpose, define key evaluation questions and users, and enable consistency between different evaluation products (TEC 2006a). The first question to ask is: are we really talking about impact evaluation? Many people referring to impact evaluation mean any kind of results, and aren't always sure about the difference between outputs, outcomes and impacts, and the kinds of methodologies needed to assess impact. In order to find out if we really mean impact evaluation, we need to define with evaluation users the evaluation purpose(s).

Patton's (2008) typology of six kinds of evaluation purposes provides a useful framework for discussion. Table 3, based on Patton (2008), illustrates how different evaluation purposes will lead to different uses and different questions being answered, and can be used to facilitate consultation on the advantages and disadvantages of various impact evaluation purposes. Further details on Patton's typology can be found in Annex 4.

Table 3: Main users and questions related to evaluation purposes

Evaluation purpose	Main users	Main questions
Summative, judgment oriented evaluation	Funders, boards, and policy makers	Did the program work? Did it attain its goals? Should the program be continued, ended, or expanded to other sites? Did the program provide good value for money?
Improvement-oriented, formative evaluation	Program staff	What are the program's strengths and weaknesses? What works and what doesn't? To what extent are participants progressing towards the desired outcomes? How can quality be enhanced?
Accountability	Those with executive, legislative and funding authority	Are funds being used for the intended purposes? Are goals and targets being met?
Monitoring	Program managers	Are inputs and processes flowing smoothly? Are outputs being produced as anticipated and scheduled? Where are bottlenecks occurring? What are variations across subgroups or sites?
Knowledge generating evaluation	Program designers, planners, modelers, academics, policymakers	What are the general patterns and principles of effectiveness across programs, projects and sites? What principles can be extracted across results to inform practice?
Developmental evaluation	Social innovators	What's happening at the interface between what the program is accomplishing and what's going on in the larger world around it? What can we control and not control, predict and not predict, measure and not measure, and how do we respond and adapt to what we cannot control, predict or measure?

For joint impact evaluation, the options would appear to be summative/judgment oriented purposes, and/or lesson-learning purposes aimed at improving performance, and/or knowledge generation purposes, or a mix of the three. Further discussion of these options can be found in

Annex 4. Two illustrative examples of evaluation purposes and linked questions, users and methodology, are provided in Boxes 1 and 2.

Box 1: Impact evaluation of humanitarian action in Darfur

Main purpose: Lesson learning

Key questions: What has worked well and not worked well in the Darfur intervention, and why? What are the programs strengths and weaknesses? What are the implications for future programming?

Main users: Programme staff

Focus: On the affected population, to make a casual link between the introduction of interventions and changes in the affected population's welfare and livelihoods. Interventions include those set out in the nine Sudan clusters.

Scale: System-wide, analysing the total humanitarian intervention since 2008.

Timeline: 18 months for the impact evaluation, including a six-month preparatory phase, including hiring of impact evaluators.

Methodology:

- Use of a program theory approach to construct the logic behind interventions, based on document Review and interviews with key stakeholders at HQ and in country; assessment of whether causal linkages worked as anticipated, to determine why interventions have been successful or not.
- Multiple visits to affected population households over the period of a year, to evaluate changes in welfare (especially mortality, morbidity) and livelihoods using participatory evaluation techniques.
- Sample size: 300 households.
- Baseline developed for the pre-2008 period by document review and recall with the affected population.
- Counter-factual established through use of proportional piling techniques with the affected population, to examine changes in welfare and livelihoods and how far these can be attributed to specific interventions.
- Additional qualitative studies on protection, governance, gender equality and the environment.

Capacity: Mix of national and international evaluators and researchers.

Coordination mechanism: Two-tier structure made up of a steering committee and management group.

Advantages and disadvantages: The main disadvantage is that the information on results achieved will be based on recall and mainly from the perspective of the affected population; the main advantage is that there should be robust lessons for future programming about the relative efficacy of different interventions.

Box 2: Impact evaluation in the post-disaster recovery stage in Bangladesh

Main purpose: Summative

Key questions: What results were achieved? Were intended objectives met? Did the intervention provide value for money?

Key users: Donors

Focus: On the affected population, to determine changes in their welfare and livelihoods.

Scale: System-wide, analysing the total humanitarian intervention in the post-disaster phase.

Timeline: 12 months for the impact evaluation, including a three-month preparatory phase including hiring of impact evaluators.

Methodology:

- Document review and interviews with key agency and government stakeholders.
- Eight village/community studies focusing on results achieved from the perspective of the affected population, using qualitative techniques.
- Counterfactual established through quasi-experimental design with comparison groups also using pre/post design. Some baseline data is likely to be available.
- Sample size: 1200 households for treatment group and 400 households for the comparison group.
- Triangulation of qualitative and quantitative findings.
- Special study on efficiency.

Capacity: Mix of national evaluators, and international evaluators and researchers.

Coordination mechanism: Two-tier structure made up of a steering committee and management group.

Advantages and disadvantages: The main disadvantage is that there would be limited focus on lesson learning; the main advantage is that there would be robust quantifiable evidence on results achieved, and evaluation costs would be decreased because of the main focus on judging results.

Definition of joint impact evaluation

Proudlock and Ramalingam (2009: 10) noted: “many respondents pointed to a significant degree of confusion regarding the definition and purpose of impact assessments.” Twelve definitions (5) of impact and impact evaluation were reviewed for this paper, and commonalities are that impact evaluation should focus on:

- the end of the results chain – this could be several years, or it could be a few months, dependent on the kind of intervention.
- attribution and establishing a counterfactual, that is determining the causes that led to specific results.
- changes in the lives of the affected population.
- unintended as well as intended results.
- negative as well as positive results.

5 ALNAP (2009), discussing Roche (2000); Bamberger (2009); DFID (2009); Jones *et al* (2009); NONIE (2009); White (2009); Watson (2008); ECB (2007); ALNAP (2006); Fearon (2006); World Bank website <http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,menuPK:384339~pagePK:162100~piPK:159310~theSitePK:384329,00.html>; Hofmann *et al* (2004).

A proposed working definition of joint impact evaluation which includes these elements and purposes discussed above is:

Joint judgment of the merit of the intended and unintended and negative and positive end results of interventions; attribution of results to particular interventions; and/or program and generalizable lesson learning.

Option 2: Should the focus be institution oriented, affected population oriented, or both? Should the scale be the entire system, or parts of the system?

Impact evaluation focus and scale will depend on the purpose(s) selected. If the purpose is judgment of humanitarian action or knowledge generalisation, the scale may be the whole system. If the purpose is to understand how particular interventions lead to specific results, the focus may be on determining the relative effectiveness of, for example, cash transfers as opposed to food aid.

In terms of focus, NONIE's impact evaluation guidance (2009: 7-8) usefully distinguishes between the institutional and affected population levels:

Interventions that can be labeled as *institutional* primarily aim at changing second-order conditions (i.e., the capacities, willingness, and organizational structures enabling institutions to design, manage, and implement better policies for communities, households, and individuals). Examples are policy dialogues, policy networks, training programs, institutional reforms, and strategic support to institutional actors (i.e., governmental and civil society institutions, private corporations, and hybrids) and public private partnerships. Other types of interventions directly aim at/affect *communities, households, and individuals*, including voters and taxpayers. Examples are fiscal reforms, trade liberalization measures, technical assistance programs, cash transfer programs, construction of schools, etc.

Scale refers to the unit of analysis to be evaluated. It could be defined according to the scope of the crisis, whether geographically, or by affected population groups. A "system-wide" evaluation could include all areas of the humanitarian intervention instigated by humanitarian actors including the UN, donors, IFRC/ICRC and local and international civil society. It could take the CAP as the unit of analysis. Or it could be defined more broadly to include government interventions, and/or remittances and other contextual factors such as migration and indigenous coping strategies. In terms of costs and benefits, clearly the wider the scale, the more challenging the evaluation process will be, but the greater the ability to include a focus on overall impact.

Option 3, method: should the method be mainly quantitative, mainly qualitative, or a combination?

Impact evaluations are currently conceptualised as quasi-research projects that may require academic input. For example DFID's (2009: 14) evaluation policy defines impact evaluation as: "a specialised type of evaluation which uses research methods to give us rigorous evidence on whether a policy, programme or project has actually changed people's lives and whether outcomes are directly attributable to the interventions."

There is an ongoing debate as to the most appropriate methods to be used for impact evaluation which can only be touched on here, but which should be a topic for discussion during the consultation process. Almost all specialists on impact evaluation agree that it needs to be

able to attribute results to specific interventions, and many argue that establishing a counterfactual (what would have happened without the intervention) is the only method for doing this. One recommended approach for impact evaluation (e.g. Bamberger and White 2008; Bamberger *et al* 2006; White 2006) is to combine some form of experimental design (6) with qualitative process evaluation, e.g. program theory. The advantage of this methodological approach is that it is seen to allow both an assessment of what happened and why it happened. Application of such an approach implies a mixed-methods evaluation design.

As Bamberger (2009: 9) points out: “Although many impacts cannot be fully assessed until an intervention has been operating for several years, planners and policymakers cannot wait three or five years before receiving feedback; consequently, many IEs are combined with formative or process evaluations designed to provide preliminary findings on whether a program is on track to achieve its intended outcomes. These designs are strengthened if a program theory model is used to define key milestones and indicators at each stage of the program cycle.” Further details on using program theory and experimental design methods in combination can be found in Annex 4.

Quantitative experimental design is unlikely to be feasible in many humanitarian contexts because of ethical concerns about, and practical difficulties of, establishing control or comparison groups. Jones *et al* (2009: 32) argue that: “quantitative methods are not appropriate in emergency response work...In an emergency, it is agreed that qualitative work, and work which does not necessarily look at longer-term impact, should be carried out. We need to know things like: if people were reached, if the timing was good, if people were satisfied, the quality of goods and services and who was left out.” Much evaluation of humanitarian action uses “informal” experimental design, for example reviewing targeting practices and through discussions with affected people who haven’t received appropriate support. Using quantitative experimental design may be feasible where the operating environment is relatively stable – e.g. long-term refugee camp situations, or in the post-disaster reconstruction phase. (7) An important example is the four impact evaluations undertaken by IRC which use randomized control trials in post-conflict and protracted emergency settings (Proudlock and Ramalingam 2009; Nelson 2008).

Methodologies to determine causality do not necessarily need to rely on quantitative approaches such as experimental design. For example one tested qualitative technique for determining causality is proportional piling. This is a participatory rural appraisal method described in detail in the *Participatory Impact Assessment Guide* (Catley *et al* 2008), and involves the affected population ranking the impact of different interventions on their livelihoods; it can also be used to attribute the relative significance of different factors on their lives and livelihoods – e.g. the changing dynamics of the conflict versus humanitarian action versus their own coping strategies and adaptations. For example informants are asked to distribute one hundred counters amongst the different variables or indicators, with the largest number of counters being assigned to the most important indicator, and the smallest number of counters being assigned to the least important indicator. Among others, Buchanan-Smith and Jaspers (2006) used this method to rank livelihood strategies in Darfur. While lacking the scientific rigor of quasi-experimental design it does involve more sustained interaction with the affected population. Similarly establishing a counterfactual may not require formal experimental design; it

6 White (2007: 6) defines experimental approaches as: “the random selection of two groups – control and treatment, beneficiaries and non-beneficiaries of an intervention such that the only difference between the two groups is the variable of interest, i.e. the impact of the intervention.” Experimental design includes randomized control trials and quasi-experimental design, and usually involves setting up a control or comparison group.

7 3ie is currently working on a paper on impact evaluation during recovery. A draft was reviewed for this options paper. See also Catani *et al* (2009) and van Dijk and van Leersum (2009).

could be done by asking the affected population what would have happened if they hadn't received support from external agents.

Respondents agreed that methodologies for joint impact evaluation should include systematic and sustained affected population engagement, which would necessitate participatory evaluation techniques as one part of the methodology.

One methodological option would involve participatory longitudinal studies to be carried out with the same set of households. The first visit could be used to set a baseline, either by recall or at the time dependent on the timing of the emergency. Three to four household visits over the course of a year would provide more reliable evidence of results than one-off evaluator visits, and will also allow for community profiles. This method was used in the WFP (2002) RTE in Southern Africa. (8)

Jones *et al* (2009) argue that despite the rhetoric around the use of mixed methods in development impact evaluation, this is rarely translated into practice, where the focus is mainly on quantitative methods. At some point fairly early on in the evaluation process a decision will need to be made as to whether there is adequate capacity commitment, resources and time for combining quantitative and qualitative evaluation methods, and what level of combination is feasible. Discussion of this area is planned as one part of stakeholder consultation.

Timing and time-scale

Timing of joint impact evaluation will be dependent on its purpose. For example if the main purpose of the evaluation is to feed into the broader knowledge base, a longer-term perspective can be taken, but if it is to feed into ongoing programming then it needs to be tailored to this. Timing will also be dependent on the kind of intervention being evaluated. If the focus is on immediate life-saving interventions, the time scale may be shorter than longer-term interventions focusing on livelihoods or housing, dependent on how far along the causal pathway impact is to be tracked. For example, food aid can be tracked in terms of its longer-term impact on nutritional levels, gender equality or protection.

Impact evaluation in the development field often has a long preparatory phase. Naudet and Delarue (2008) report on a preparation phase of one and a half years, involving policymakers, researchers, data collectors, and donors meeting and agreeing on the principles and details of the exercise. Humanitarian situations may require a shorter time-frame. Impact evaluation therefore needs to be located within a holistic monitoring and evaluation system which can analyse different levels of results over different time frames, and meet the differing information needs of different stakeholders. One of the challenges facing Inter-Agency Real Time Evaluations (RTEs) was insufficient buy-in at the country level, so adequate time needs to be devoted to consultation with country level actors (see Section 2). Estimated timing for a year-long impact evaluation pilot can be found in Table 4.

8 The evaluation team visited the six countries covered by the southern Africa regional emergency during three missions, from July 2002 to May 2003. The evaluation included a mixed method approach including: document review; interviews with programme staff; focus group discussions with the affected population at Final Distribution Points in each country; and household visits where in-depth semi-structured interviews with the affected population were undertaken. In addition, an ad hoc 'sentinel site' studies were undertaken in Malawi and Zambia where the family was visited three times for an insight into the impact of the operation.

Table 4: Timing for a year-long impact evaluation

Month 1	Evaluation design including sample size and location, discussions with government, UNCT/IASC country team/clusters, defining users, hiring national evaluators and researchers First visit of international evaluation team
Month 2/3	Initial round of data gathering with the affected population and other key stakeholders Review of progress and initial feedback
Month 4/5	Further round of data gathering
Month 6	Initial evaluation report (e.g. to donors/government) Second visit of international evaluation team
Month 7/8	Further round of data gathering
Month 9-11	Data validation (e.g. with the affected population) and analysis Third visit of international evaluation team (9)
Month 12	Final evaluation report

Option 4: Capacity – National or international or both?

Hofmann *et al* (2004: 3) commented: “The humanitarian system currently lacks the skills and capacity to successfully measure or analyse impact. Greater investment needs to be made in human resources and research and evaluation capacity if the desire to focus more on results is to be realised.” The “market” for humanitarian evaluators operates like most markets; those evaluation offices with better networks and incentives tend to corner the market for skilled evaluators available. Most good humanitarian evaluation practice stems from having the right evaluation team in the right place at the right time, and conversely much poor evaluation practice stems from rushed evaluations dependent on busy multi-tasking independent international evaluators or belt-way evaluation companies.

Joint impact evaluations are likely to be a long-term process and could involve a mix of national and international researchers and evaluators. Most of the data gathering could be carried out by national researchers and evaluators, supported on a periodic basis by an international team whose role would be input into evaluation design, data analysis and triangulation, interviews with international agencies, report drafting, and for purposes of international comparison. If the impact evaluation is to take place over a year, 3 - 4 visits of about two weeks per visit would be adequate for the international team. This model is appropriate for joint impact evaluation because it will allow more sustained participation with the affected population, at the same time bringing in international expertise as relevant. It will also help to support national capacity, and be more cost-effective.

Option 5: Piloting impact evaluation in humanitarian action

Dependent on the outcomes of the stakeholder consultation there may be evaluation pilots after this in at least two countries/regions. As well as agreement about the purpose of the evaluation, factors leading to successful pilots (10) are likely to be:

- One agency as lead agency and champion. Without a lead agency to take on primary responsibility a joint evaluation can be frustrating and unsuccessful.

9 Dependent on local capacity four rather than three visits by the international evaluation team may be necessary.

10 Based on Jones *et al* (2009); Beck and Buchanan-Smith (2008); ECB (2007).

- Adequate international and national capacity.
- Country level support from government and the HC and UNCT.
- Availability of baseline and monitoring data.
- Adequate funding. After reviewing a range of joint evaluation budgets, Beck (2009) estimates the cost of a year-long joint impact evaluation at some \$300,000.
- Security. Access to affected populations may be restricted due to security issues. A year-long evaluation process may help overcome this constraint.

During discussions with respondents the idea was floated of carrying out evaluation pilots in two settings, one complex emergency such as Darfur, and one natural disaster such as Bangladesh, which is subject to almost yearly major natural disasters (see Boxes 1 and 2); CRED (2008) notes that Bangladeshis are the most susceptible population in the world to natural disasters. A decision would need to be made as to whether the two pilots would have the same purpose, focus and scope and methodology. It may be useful to consider a third pilot using a different methodology to attempt to determine the most appropriate approach.

Option 6: Coordination - A common framework approach? (11)

There are several decisions that will need to be made about the coordination and management arrangements for joint impact evaluation. OECD-DAC's (2006) typology of joint evaluations will help determine which modality is most appropriate:

- **Classic joint evaluation:** Participation is open to all stakeholder agencies. All partners participate and contribute actively and on equal terms (e.g. Joint Evaluation of Emergency Assistance to Rwanda, General Budget Support)
- **Qualified joint evaluation:** Participation is open only to those who qualify, through membership of a certain grouping (e.g. EU) or through participation in the activity that is being evaluated (e.g. Basic Education evaluation)
- **Hybrid joint evaluation.** Includes a wide range of alternative ways of joint working:
 - a. responsibilities are delegated to one or more agencies while others take a 'silent partnership' role.
 - b. some components of the evaluation are undertaken jointly while others are delivered separately.
 - c. various levels of linkage are established between separate but parallel and interrelated evaluations.
 - d. the joint activity is agreeing a common evaluation framework, and responsibility for implementation of individual evaluations is devolved to different partners.
 - e. research, interviews and team visits are undertaken jointly but each partner prepares a separate report.

11 This Section is based on Beck (2009) and a review of five joint evaluations: Borton *et al* (2005) on the IDP evaluation; Eriksson (2009) and Wood *et al* (2008) on the Paris Declaration evaluations; IDD (2007) on the evaluation of General Budget Support; Netherlands Ministry of Foreign Affairs (2003) and Samoff (nd) on the basic education evaluation; and TEC (2006, 2006 a and b, 2005) on the Tsunami Evaluation Coalition.

Respondents noted that the “hybrid joint evaluation”, option d., might be the most appropriate to test for a joint humanitarian impact evaluation. This was because the system would be best able to manage this level of “jointness”. This option was used in the IDP and Basic Education evaluations among others, and the experience of these two evaluations points to some of the costs and benefits of this approach. Details are reviewed in Beck (2009) and can be used as a basis for consultation in this area.

Option 7: Optimal management arrangements for joint impact evaluation - a dual structure?

Current good practice suggests that:

- joint evaluation will not work well without appropriate leadership;
- lack of an adequate management structure appears to ensure problematic joint evaluations;
- management structures should be kept simple and light;
- it is critical to have a core group of 4 or 5 agencies involved at an early stage to move things forward;
- there should be a clear delineation of roles and responsibilities; and
- joint evaluations require full-time coordination, administration and research support.

OECD DAC (2006: 17) notes: “Lead-agencies need to take early decisions on balancing demand for wide participation with the need to keep the evaluation process streamlined and efficient.”

The most commonly used management structure in joint evaluations is a two-tier system made up of a steering committee which provides overall guidance and oversight, and a smaller management group which manages the evaluation on a regular basis (e.g. the Paris Declaration 2, Basic Education and General Budget Support evaluations). The advantages and disadvantages of different management structures are reviewed in Beck (2009), which can form the basis for planned consultation.

A further element involves host country participation in the management structure. This has been limited to date in joint evaluations, although achieved in the case of the Interagency Health Evaluation initiative (Beck and Buchanan-Smith 2008). In complex emergencies this may be problematic, especially where governments are party to conflict. Phase 2 of the Paris Declaration evaluation, which will take place in 2009 and 2010, is proposing the use of Country Reference Groups, which will consist of key government and non-government development stakeholders in a given partner country, including but not confined to the central government and key donors. Following this evaluation process will likely be useful for any proposed joint humanitarian impact evaluation.

Experience suggests that a full time coordinator is needed for joint evaluation work of significant scale. The quality of joint impact evaluations will be to a large extent dependent on the skills of this person. Funding arrangements and the coordinator’s institutional location would need to be determined.

A trigger mechanism?

The idea of an evaluation “trigger” was floated during discussions of Inter-Agency RTEs (OCHA 2007). One issue that surfaced during the Inter-Agency RTEs was that whether or not an Inter-Agency RTE should take place required negotiation in each case with the RC and UNCT. Having a mandate from the IASC/ERC stating the situations where a joint impact evaluation

takes place would mean that all actors would be aware when this would happen and be able to prepare accordingly. A joint impact evaluation could for example be triggered at the request of an HC, and/or when the total humanitarian budget exceeded a specified amount. This area should be discussed during the planned consultation period.

Bibliography

ADB (2006) *Impact Evaluation Methodological and Operational Issues*. Philippines: ADB.

ALNAP (2006) *Evaluating Humanitarian Action Using the OECD-DAC Criteria*. London: ALNAP.

Baker, J. (2000) *Evaluating the Impact of Development Projects on Poverty*. Washington, DC: World Bank.

Bamberger, M. (2009) *Institutionalizing Impact Evaluation within the Framework of a Monitoring and Evaluation System*. Washington, DC: IEG.

Bamberger, M and White, H. (2008) "Introduction: Impact Evaluation in Official Development Agencies." *IDS Bulletin* (39) 1-11.

Bamberger, M. J. Rugh and L. Mabry (2006) *Real World Evaluation Working Under Budget, Time, Data, and Political Constraints*. Thousand Oaks: Sage Publications.

Beck, T. and Buchanan-Smith, M. (2008) "Joint evaluations coming of age? The quality and future scope of joint evaluations." *Review of Humanitarian Action*. London: ALNAP

Borton, J. et al (2005) *Support to Internally Displaced Persons – Learning from Evaluations. Synthesis Report of a Joint Evaluation Programme*. Stockholm: Sida.

Buchanan-Smith, M. and S. Jaspers (2006) *Conflict, Camps and Coercion: The Continuing Livelihoods Crisis in Darfur. Final Report*. A Report to WFP Sudan: mimeo.

Catani, C., Kohiladevy, M., Ruf, M., Schauer, E., Elbert, T. and Neuner, F. (2009) "Treating children traumatized by war and Tsunami: a comparison between exposure therapy and meditation-relaxation in North-East Sri Lanka." *BMC Psychiatry* (9).

Catley, A. et al. (2008) *Participatory Impact Assessment: A Guide for Practitioners*. Feinstein International Center, Tufts.

CRED (2008) *Annual Disaster Statistical Review. The Numbers and Trends 2007*. Brussels: Centre for Research on the Epidemiology of Disasters.

DANIDA (2009) *Experiences with Conducting Evaluations Jointly with Partner Countries*. Denmark: DANIDA.

DANIDA (2003) *Framework for a Common Approach to Evaluating Assistance to IDPs*. DANIDA, mimeo.

ECB Project (2007) *Impact Measurement and Accountability in Emergencies: The Good Enough Guide*. Oxford: OXFAM.

ECHO (2007) *Towards a European Consensus on Humanitarian Aid*. Brussels: ECHO.

Eriksson, J. (2009) "Approach Paper for Phase 2 Evaluation of the Paris Declaration. " Prepared for the OECD DAC Evaluation network meetings, February, mimeo.

Forss, K. and Bandstein, S. (2008) "Evidence-based Evaluation of Development Cooperation: Possible? Feasible? Desirable?" *IDS Bulletin* (39), 82-9.

Hofmann, C-A. *et al* (2004) "Measuring the Impact of Humanitarian Aid. A Review of Current Practice." London: *HPG Report 17*.

IDD (2007) *Evaluation of General Budget Support. Note on Approach and Methods*. Birmingham: International Development Department, University of Birmingham.

Ito, S. and Wada Y. (2008) "Learning to Evaluate the Impact of Aid". *IDS Bulletin* (39) 71-81.

Jerve, A. and Villanger, E (2008) *The Challenge of Assessing Aid Impact: A Review of Norwegian Evaluation Practice*. Oslo: NORAD.

Jones, N. *et al*. (2009) *Improving Impact Evaluation Production and Use. A Scoping Study commissioned by the DFID Evaluation Department on behalf of the NONIE*. London: ODI working Paper 300.

Leeuw, F and Vaessen, J (2009) *Impact Evaluations and Development*. Washington, DC: NONIE

Naudet, J and Delarue J (2008) "Fostering Impact Evaluations at Agence Française de Développement: A Process of In-house Appropriation and Capacity-Building." *IDS Bulletin* (39) 12-22.

OCHA (2007) "Learning on the Mechanism for Triggering an IA RTE." Mimeo.

OECD DAC (2006) *Guidance for Managing Joint Evaluations*. Paris: OECD-DAC Evaluation Series.

OECD DAC (2005) *Joint Evaluations: Recent Experiences, Lessons Learned and Options for the Future*. Paris: OECD-DAC.

Patton, M. (2008) *Utilisation-Focused Evaluation: the new century text. Third Edition*. Thousand Oaks: Sage.

Proudlock, K. and Ramalingam, B with Sandison, P (2009) "Improving Humanitarian Impact Assessment." *Review of Humanitarian Action 8*, London: ALNAP.

Ramalingam, B. and Mitchell, J. (2009) "Counting What Counts: performance and effectiveness in the humanitarian sector." *Review of Humanitarian Action 8*, London: ALNAP.

Ruprah, I (2008) "'You Can Get It If You Really Want': Impact Evaluation Experience of the Office of Evaluation and Oversight of the Inter-American Development Bank". *IDS Bulletin* Volume (39), 23-35.

Samoff, J. *et al* (nd) "Consensus, Legitimacy, and Critique in Joint Evaluations An Analysis of *Local Solutions to Global Challenges: Towards Effective Partnership in Basic Education*." mimeo.

TEC (2006) "Lessons about multi-agency evaluation." ALNAP: mimeo.

TEC (2006a) "Notes from an 'After Action Review' held at the Tsunami Evaluation Coalition (TEC) Core Management Group (CMG) meeting, Copenhagen." Mimeo.

van Dijk, S. and A. van Leersum (2009) "Measuring the socio-economic impact of post-disaster shelter: experiences from two Red Cross programmes." *Humanitarian Exchange* 44.

Watson, C. (2008) *Literature Review of Impact Measurement in the Humanitarian Assistance Sector*. Paper submitted to the Feinstein International Center, Tufts University for the Bill and

Melinda Gates Foundation Project: Impact Assessment of Innovative Humanitarian Projects in Sub-Saharan Africa.

WFP (2002) *Full Report of the Real Time Evaluation of WFP's Response to the Southern Africa Crisis. EMOP 10200.0*. Rome: WFP.

White, H (2008) "Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation." NONIE Working Paper No. 7.

White, H. (2007) *Evaluating Aid Impact*. Munich Personal RePEc Archive Paper 6716.

Wood, B. *et al* (2008) *Evaluation of the Implementation of the Paris Declaration. Phase 1 Synthesis Report*. Kabell consulting.

Annex 1 Options Paper Terms of Reference

United Nations Office for the Coordination of Humanitarian Affairs (OCHA)

Terms of Reference

Options Paper for Impact Evaluation in Humanitarian Action

Background

The issue of impact is high on the current humanitarian agenda. This is driven by a number of factors including a shift towards a language and practice of managing and being accountable for results; calls for more ‘evidence-based’ policy; methodological advances for assessing causality; increasing awareness of the potential unintended, negative effects of humanitarian aid; and an increasingly urgent need to generate knowledge about ‘what works’.¹²

Within the humanitarian sector, many individual agencies¹³ now include impact in their evaluation guidelines. At the interagency level, there are also a number of initiatives aimed at monitoring and evaluation of impact,¹⁴ as well as increasing collaboration through joint evaluations. Yet despite some progress in this area, humanitarian impact assessments have not become common practice and the sector has yet to establish a common understanding of what ‘impact’ means and how to measure it. The lack of shared understanding has muddled the discussion about the desirability and feasibility of impact assessments in different contexts, and muddled the selection of appropriate indicators and methodologies. Moreover, many disincentives exist within agencies and the wider sector, including aversion to perceived risk of failure.

In June 2009, OCHA commissioned an *Evaluability Assessment for Impact Evaluation of the Humanitarian System at the Country Level* which was presented by its author, Tony Beck, at a workshop coordinated by OCHA looking at the future of interagency evaluations. Participants indicated support for the initiative and requested OCHA commission a follow on paper, expanding upon the initial work, to be presented for discussion at the 25th Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP) Meeting in London on 17 – 18 November 2009. The paper will serve as the foundation for an OCHA-facilitated discussion during the ALNAP meeting while also complementing OCHA efforts to strengthen needs assessment, monitoring and evaluation at the cluster/sector level.

Scope of Work

The objectives and aim of impact evaluation are myriad. The paper will address conceptual questions while also providing insight into practical approaches for implementation. It will clearly articulate different models/approaches for variant levels (e.g. system as a whole, reform effort, cluster approach, CAP), with associated resource requirements, management and governance structures, and propose a path for strengthening stakeholder engagement and consensus. This information will be framed within two situational examples, one complex emergency and one natural disaster, and provide specific key questions of potential interest to the humanitarian community, (eg. *What difference has the roll-out of the cluster approach made to the lives of crisis-affected populations in Darfur?*).

In consultation with OCHA ESS and ALNAP, and based on existing work, interviews and the consultant’s own analysis, the paper should attempt to outline/identify:

- (a) Conceptual and theoretical questions which will need to be considered during any joint consultation process in order to proceed with developing a framework for impact evaluation in humanitarian action.

¹² Proudlock, K. and B. Ramalingam (2009) “Improving Humanitarian Impact Assessment: Bridging Theory and Practice” in *ALNAP 8th Review of Humanitarian Action* London, ODI.

¹³ WFP, ECHO, UNICEF and USAID all include impact in their evaluation guidelines.

¹⁴ These include the Standardized Monitoring and Assessment of Relief and Transitions (SMART), the Health and Nutrition Tracking Service (HNTS), the Tsunami Recovery Impact Assessment and Monitoring System (TRIAMS), the Fritz Institute Humanitarian Impact project, and the Collaborative for Development Action Listening Project.

(b) Within two situational examples, a set of specific, key questions about the impacts/results of humanitarian action that a joint humanitarian impact evaluation could usefully seek to address e.g. *What difference has the roll-out of the cluster approach made to the lives of crisis-affected populations in Darfur?*

Then, for each specific question identified, the paper should present preliminary suggestions regarding the following:

(c) Who the key stakeholders are and potential strategies for identifying/consulting with them. Specifically, the consultant will be required to:

- Identify innovative approaches to stakeholder analysis and consultation, including approaches to coordination and leadership, which will help to make interests explicit and identify common ground.
- Present a map of proposed approaches to ensuring continued and increased stakeholder engagement post-ALNAP meeting in November 2009 to strengthen and expand existing interest and support.

(d) Definitions of ‘impact’, paying specific attention to particular theories of change i.e. the ways in which a particular intervention or aspect of humanitarian action is expected to achieve outcomes/impacts.

(e) A range of feasible methodological approaches and associated strengths and weaknesses of each (e.g. in terms of methodological rigour and ability to address the attribution problem; internal/external validity of results; associated resource requirements etc.). It will be important to ensure that participation of affected populations is prioritised.

(f) Potential management and governance structures and trigger mechanisms

- Suggested chairmanship, composition and roles and responsibilities of management and governance bodies
- Suggested mechanisms, pathways or situations which would serve to ‘trigger’ an impact evaluation

Questions of impact should not be limited to the evaluation process. In the humanitarian sphere, a concern with change in the short term implies a need for impact to be considered in ongoing monitoring processes, and through techniques such as real-time evaluation. In such a manner, the problem of aggregation inherent in looking at the humanitarian system as a whole – comprised of its separate ‘parts’ – is addressed. As such, it is possible that a realistic approach to measuring humanitarian impact is located at the nexus of scientific, analytical and participatory approaches, applied over time as opposed to a once-off evaluation.

Management Arrangements

The external consultant will report to OCHA’s Evaluation and Studies Section (ESS).

ESS will assign a manager to oversee the conduct of the exercise and assure quality control. His/her responsibilities are to: 1) provide guidance and institutional support to the external consultant, especially on issues of methodology; 2) facilitate the consultants access to key stakeholders and specific information or expertise needed to perform the assessment; 3) ensure that all stakeholders are kept informed; 4) recommend the approval of the final product; 5) help to coordinate and design workshop; and 6) facilitate and monitor subsequent follow up.

Duration and Tentative Workplan

The external consultant will be contracted for 15 days from 10 October – 31 December 2009 to undertake the desk review, interviews with key stakeholders, prepare a paper and outline for discussion at the 17 – 18 November 2009 ALNAP meeting whose audience would include key humanitarian actors, donors and academics.

<i>Activity</i>	<i>Number of Days</i>	<i>Timing</i>
Literature review	3	October 2009
Key informant interviews and conference calls	2	October 2009
Report writing	5	October/November 2009
Workshop preparation	1	November 2009
Workshop attendance	2	November 2009
Documentation of workshop results, report revisions and follow up work	2	November/December 2009
<i>Total days</i>	<i>15</i>	

Competency and Expertise Requirements

The assessment will require the services of a consultant with the following experience and knowledge:

- Extensive experience with both theoretical approaches and practical applications of impact evaluations – preferably in humanitarian settings
- Demonstrated ability in the development of innovative new approaches
- Experience in conducting evaluations of humanitarian programmes and the capacity to work collaboratively with multiple stakeholders
- Previous work in or knowledge of the UN, NGO and/or donor institutions
- Workshop facilitation skills
- Strong analytical skills and ability to synthesize and present findings, draw practical conclusions and articulate complex concepts in a clear and concise manner

Reporting Requirements and Deliverables

- Draft report addressing each of the points outlined above under *Scope of Work* no later than 3 November 2009
- Revised draft report incorporating comments no later than 10 November 2009.
- Outline for discussion at 18 November ALNAP meeting in London no later than 10 November 2009.
- Attendance at and support to facilitation of discussion at 17 – 18 November ALNAP meeting in London.
- Follow up consultations and paper revisions based on information elicited during 17 – 18 November ALNAP meeting, resulting in final document no later than 9 December 2009.

Annex 2 Interviewees

Person	Organization
Jock Baker	CARE/ECB
Tijana Bojanic Krishna Belbase Finnbar O'Brien	UNICEF Evaluation
John Borton	Independent
Margie Buchanan-Smith	Independent
Jeff Crisp Maria Rijskaaer	UNHCR Evaluation
Alexis Hoskins Niels Scott	OCHA Assessment and Classification of Emergencies
Janey Lawry-White	UNDP BCPR
Simon Lawry-White	IASC
John Mitchell Karen Proudlock Ben Ramalingam	ALNAP
Robert Smith	OCHA CAP Unit
Caroline Heider Sally Burrows	WFP Evaluation
Ron Bose	3ie

Annex 3 Sample Consultation Questions

1. What do you think should be the purpose of joint impact evaluation? Should it be for judging and attributing the results of the intervention, promoting lesson learning within an ongoing program, or generating knowledge for the system as a whole? Or a combination of these? Do you think it would be feasible to include more than one of these purposes as the main focus of a joint impact evaluation?
2. What do you think should be the scale of a joint impact evaluation? Should it cover the whole humanitarian intervention by the international system, or part of this? Should it include government interventions? What are the advantages and disadvantages of a “system-wide” scale?
3. What are the best methods to use in joint impact evaluation? Do you think that an experimental design using control or comparison groups would be feasible in a humanitarian setting? In a post-disaster setting? Or do you think qualitative methods such as proportional piling and focus groups are more appropriate? Do you think it would be feasible to combine quantitative and qualitative methods in one evaluation? If so, how would you see this working? Is there capacity currently to carry out effective joint impact evaluation? Would you see a mixed national/international team as appropriate?
4. How long do you think a system-wide joint impact assessment should take to carry out from the development of the Terms of Reference to final report?
5. What is the optimal management structure for a joint impact evaluation? Do you think that a two tiered structure involving a committee providing overall guidance, and a management group managing day to day activities, would work for joint humanitarian impact evaluation?

Annex 4 Purposes and methods for joint humanitarian impact evaluation

Patton (2008) defines the following evaluation purposes:

1. Summative, judgment oriented evaluation. Evaluations aimed at determining the overall merit, worth, significance or value of something are judgement oriented. Summative evaluation fits into this category – the aim is to report on results achieved – summative evaluation provides data to support a judgment about the program’s worth so that a decision can be made about the merit of continuing the program. Patton includes impact evaluation under this heading.

2. Improvement-oriented, formative evaluation. The main focus is on using evaluation results to improve a program, as opposed to judging the program. Patton includes lesson learning under this heading.

3. Accountability. Program and financial audits, which are aimed at assuring compliance with intended purposes and mandated procedures. It’s important to draw a distinction between accountability and judgment-oriented evaluation. The latter is what happens in many evaluations, with the assumption that determining the merit of an intervention will lead to greater accountability. It might, but there is often no direct link between summative evaluation and accountability.

4. Monitoring. Ongoing monitoring serves managers, serving those internal to the program with the information they need to know where their managerial attention will do the most good.

5. Knowledge generating evaluation. Knowledge generation changes the unit of analysis as evaluators look across findings from different programs to identify general patterns of effectiveness. This is a kind of meta-synthesis.

6. Developmental evaluation. Summative evaluation makes an overall judgment of merit or worth based on efficient goal attainment, replicability, clarity or causal specificity, and generalizability. None of these traditional criteria are appropriate or even meaningful for highly volatile environments, systems-change-oriented interventions, and emergency social innovations. Staff in these situations aspire to continuous progress, ongoing adaptation, and rapid responsiveness.

The first point to note from Patton’s typology in thinking about the purpose of joint impact evaluation is that the traditional differentiation between accountability and lesson learning purposes of evaluation of humanitarian action (EHA) – as found in hundreds of terms of reference – is too simple. When evaluation managers talk about accountability, they usually mean reporting functions – reporting to an executive board, to parliamentarians, to the funding agency, to the affected population, or to the taxpaying public. Humanitarian actors are rarely accountable in the sense of their being some consequence after negative evaluation findings. Accountability in the public sector which is the source for most humanitarian funding is quite different from accountability in the private sector where there is a less complex relation between cause and effect. In any case, a joint evaluation process is *less* likely to promote accountability, because there will be less focus on one agency and more on joint results achieved.

Patton doesn’t deal with the issue of “downward accountability” to the affected population, which many would argue should be central to all humanitarian and

development activity. This can partly be promoted through evaluation processes (see Option 3 in the main paper).

There is broad consensus that impact evaluation can be used for determining what happened and attributing results to specific interventions, or judgment oriented purposes. Bamberger (2009:9) sums up this perspective (15):

The primary purpose of an IE [impact evaluation] is to estimate the magnitude and distribution of changes in outcome and impact indicators among different segments of the target population and to assess the extent to which these changes can be attributed to the interventions being evaluated.

There is less certainty that impact evaluation can get inside the “black box” and determine why things happened and the implications of this for programming (i.e. learning). Here is a word of caution from Patton (2008: 148) to those who think that judgment and learning can be combined equally in one impact evaluation:

one purpose is likely to become the dominant motif and prevail as the primary purpose informing design decisions and priority uses; or else, different aspects of an evaluation are designed, compartmentalized, and sequenced to address these contrasting purposes. I also find that confusion among these quite different purposes, or failure to prioritize them, is often the source of problems and misunderstandings along the way, and can become disastrous at the end when it turns out that different intended users had different expectations and priorities.

There are opinions on both sides of the debate as to whether judgment and lesson learning can be combined in one impact evaluation. Let's start with those who think this is possible.

NONIE's (2009: xv, xxi; see also Watson 2008) guidance on impact evaluation is clear that the two can be combined:

Overall, for impact evaluations, well-designed quantitative methods are usually preferable for addressing attribution and should be pursued when possible. Qualitative techniques cannot quantify the changes attributable to interventions, but should be used to evaluate important issues for which quantification is not feasible or practical and to develop complementary and in-depth perspectives on processes of change induced by interventions.....The authors of this Guidance document believe that the ultimate reason for promoting impact evaluations is to learn about “what works and what doesn't and why” and thus to contribute to the effectiveness of (future) development interventions.

White (2006) and White and Bamberger (2008) discuss methods for joint judgment and lesson learning purposes. They recommend combining quantitative experimental design (16) to determine results and attribution, with a more qualitative theory-based approach mapping out whether the connections in the causal results chain have been made and implications of this for programming.

15 See also Catley 2009; NONIE 2009, Jones *et al* 2009; Ito *et al* 2008; ADB 2006; Baker 2002.

16 White (2007: 6) defines experimental approaches as: “the random selection of two groups – control and treatment, beneficiaries and non-beneficiaries of an intervention such that the only difference between the two groups is the variable of interest, i.e. the impact of the intervention.” Experimental design includes randomized control trials and quasi-experimental design, and usually involves setting up a control or comparison group.

White (2006: 7) expands on what is meant by a theory-based approach:

A theory-based evaluation design is one in which the analysis is conducted along the length of the causal chain from inputs to impacts. Many impact evaluations concern themselves only with the final link in the chain: final outcomes. But to do this is often to lose the opportunity to learn valuable policy lessons about why an intervention has worked (or not), or which bits have worked better than others.

Applying a theory based approach requires mapping out the channels through which the inputs are expected to achieve the intended outcomes. In many cases this analysis will already be contained in the project log frame. The log frame may also specify indicators at the various levels. Indeed the M&E system may have collected these indicators for project areas, and can be a useful source of analysis of process aspects of the intervention. A theory-based approach examines the links in the causal chain. Were there missing or weak links? There can be missing links if the project design missed some key determinants at the next level it should have sought to influence.

White (2006) also provides a number of examples of impact evaluations of development interventions which have been successful in combining experimental design with qualitative assessment of processes. The determining factors leading to these impact evaluations appeared to be similar to those for most successful evaluations: a high quality evaluation team with relevant skills, a receptive audience, and adequate resources and time.

Others agree with Patton and are more circumspect about the potential for combining as equal partners judgment and lesson-learning. The TEC had among its aims both improving the quality of humanitarian action – or learning, which was its primary aim - and providing accountability to the donor and affected-country populations. Reviews suggest it was challenging to attempt to meet both aims in one joint evaluation.⁽¹⁷⁾ ALNAP's (2009: 50) summary of four single agency humanitarian impact evaluations concluded:

All four initiatives brought together multiple stakeholders, including implementing agencies, donors, academics, consultants, and recipients of aid. It was challenging in all cases to agree a common sense of purpose and specific questions which are valuable and interesting for key stakeholders. A common reason for the difficulty was perceived tension between the separate goals of accountability to donors and learning; in some cases, this caused problems in later stages of the impact assessments.

Forss and Bandstein (2008: 87) conclude:

As a first requirement it is necessary to distinguish evaluations that are done for managerial purposes, and so need information on management and implementation, as well as on relevance, efficiency and sustainability. Given that such issues are not best addressed through experimental studies, the risk is that they 'contaminate' the impact studies. The urgency of decision support and learning weighs heavier in the short run and determines the choice of evaluation

¹⁷ The TEC questionnaire survey in 2007 to 25 respondents found that 44% found the purpose of the TEC 'very clear', 48% 'quite clear', and 8% 'not very clear'. One respondent noted: "At the field level there was a lot of confusion about who TEC was and the purpose of TEC." [http://www.alnap.org/pool/files/tecsurvey\(2\).pdf](http://www.alnap.org/pool/files/tecsurvey(2).pdf)

design. When the aid agencies want impact information they should not confound those evaluation tasks with a number of other issues and questions, but should focus on impact and do nothing else.

Ruprah (2008: 33) concurs reflecting on the IADB experience:

Success should be measured by the degree to which impact evaluations are adopted as the norm in the institution. This has not occurred. The demonstration effect is non-existent. Success could also be judged by the creation of an effective virtuous cycle of institutional learning, whereby independent evaluation leads to the identification and utilisation of lessons by the institution, leading to improved operational work that in turn leads to improvement in lives. This has not entirely materialised. The few examples of success are due to idiosyncratic factors not institutional ones.

Jones *et al* (2009: 9) found the following two hypotheses to hold for impact evaluations:

- The production of experimental IEs is driven largely by upward accountability to donors.
- Experimental IEs tend to be commissioned less frequently to fulfil downward accountability, or operational learning purposes.

Jones *et al* (2009) also found that impact evaluations based on experimental design are rarely used instrumentally. Rather, some donors are driving the demand for impact evaluation for accountability purposes, which has led to impact evaluations as a 'box ticking' exercise. They concluded that while there has been little systematic analysis of how impact evaluations have been used in policy processes, an important orientation of the impact evaluation movement is to focus more on generalisable policy lessons rather than the mechanics of specific programmes - that is Patton's "knowledge generating evaluation".

Each of these evaluation purposes has different advantages and disadvantages, and costs and benefits. Summative evaluation will tell us what happened, but not why, lesson learning evaluation may tell us why things happened, but not give the same overview of results as summative evaluation. Evaluation users are likely to be interested in different types of information. HCs and UNCTs may be more interested in learning about why programmes are working or not. Donors may be more interested in results and reporting to their public. Policy makers and researchers may be more interested in generalizable knowledge. Timing of the evaluation will play a large role, as evaluations conducted after programs have been completed can't feed into ongoing programming. Stakeholder consultation needs to determine which of these interests is primary, and how they can best be met by different kinds of monitoring and evaluation processes. In theory there's no reason why a quantitative survey using an experimental design focusing on summative judgment and attribution (which will tell us if something happened and what caused it) can't be combined with discussions with the affected population and other stakeholders about what went well or not well in terms of achieving results, using a theory based approach (which will tell us why something happened or didn't happen). Whether these two techniques should and can be combined depends on interest, evaluation capacity, resource levels available, and timing.