Jyotsna Puri
Anastasia Aladysheva
Vegard Iversen
Yashodhan Ghorpade
Tilman Brück

What methods may be used in impact evaluations of humanitarian assistance?

December 2014

Working Paper 22

International Initiative for Impact Evaluation

About 3ie

The International Initiative for Impact Evaluation (3ie) is an international grant-making NGO promoting evidence-informed development policies and programmes. We are the global leader in funding and in producing high-quality evidence of what works, how, why and at what cost. We believe that better and policy-relevant evidence will make development more effective and improve people's lives.

3ie working papers

These papers focus on current issues, debates and enduring challenges facing development policymakers and practitioners and the impact evaluation and systematic reviews communities. Policy-relevant papers draw on relevant findings from impact evaluations and systematic reviews funded by 3ie, as well as findings from other credible and rigorous evaluations and reviews, to offer insights, new analyses, findings and recommendations. Papers focusing on methods also draw on similar sources to help advance understanding, design and use of rigorous and appropriate evaluations and reviews.

About this working paper

This working paper examines the extent to which impact evaluation methods can provide evidence to help improve the effectiveness and efficiency in humanitarian action. It is part of background scoping research and consultation undertaken to assess the scope and methods for impact evaluation in the humanitarian sector. The scoping paper, *What evidence is available and what is required, in humanitarian assistance?,* provides an independent analysis of the evidence base of evaluations in humanitarian assistance and identifies key gaps and priorities in need of rigorous evidence. All of the content is the sole responsibility of the authors and does not represent the opinions of 3ie, its donors or its Board of Commissioners. Any errors and omissions are also the sole responsibility of the authors. Any comments or queries should be directed to the corresponding author, Jyotsna Puri, jpuri@3ieimpact.org

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

3ie Working Paper Series executive editor: Howard White
Production managers: Kanika Jha and Omita Goyal
Assistant production manager: Pradeep Singh
Copy editor: Warren Davis
Proofreader: Mathew PJ
Cover design: John F McGill
Printer: VIA Interactive
Cover photo: Pablo Tosco/Oxfam

What methods may be used in impact evaluations of humanitarian assistance?

Jyotsna Puri
International Initiative for Impact Evaluation (3ie)

Anastasia Aladysheva
Stockholm International Peace Research Institute (SIPRI)

Vegard Iversen
University of Manchester

Yashodhan Ghorpade
Institute of Development Studies

Tilman Brück
SIPRI

3ie Working Paper 22
December 2014

International Initiative for Impact Evaluation

Acknowledgements

This paper is part of a larger study supported by UKaid through the Department for International Development and USAID. The scoping paper examines the scope of evidence and need for evidence in humanitarian assistance. The scoping paper has been prepared by Evidence Aid with support from 3ie. 3ie, along with assistance from the SIPRI and humanitarian assistance experts, led the work on the methods paper.

Jyotsna Puri provided overall leadership and management of these papers with support from Deo-Gracias Houndolo and Peter Giesen. Bharat Dhody provided research assistance.

The team that worked on the scoping and methods paper included: Anastasia Aladysheva, Claire Allen, Frank Archer, Tilman Brück, Mike Clarke, Anneli Eriksson, Yashodhan Ghorpade, Peter Giesen, Vegard Iversen, Jyotsna Puri and Diana Wong. The authors are also grateful to Howard White, Jeannie Annan, John Mitchell, Francesca Bonino, Joanna Macrae, Christine Kolbe, Alison Girdwood, Jonathan Patrick and Joanna Macrae.

We are also grateful to members of the steering committee which included: Caroline Andreson, Jeannie Annan, Alison Girdwood, Langdon Greenhalgh, Penny Hawkins, Christine Kolbe, Joanna Macrae, John Mitchell, John Murray, Jennie Richmond and Howard White.

Executive summary

Humanitarian crises are complex situations where the demand for aid has traditionally far exceeded its supply. The humanitarian assistance community has long asked for better evidence on how each dollar should be effectively spent. Impact evaluations of humanitarian assistance can help answer these questions and also respond to the increasing call to estimate the impact of humanitarian assistance and supplement the rich tradition for undertaking real-time and process evaluations in the sector. This working paper gives an overview of the methodological techniques that can be used to address some of the important questions in this area, while simultaneously considering the special circumstances and constraints associated with humanitarian assistance.

Key findings for the scope of future study

This working paper is part of a larger study undertaken to assess the scope and methods for impact evaluation in the humanitarian sector. Findings from the scoping paper show that:

• **Insufficient high-quality evidence:** High-quality evidence that can causally relate changes in the conditions of people and their outcomes to specific programmes and interventions undertaken in humanitarian assistance are clearly scarce. In an investigation of studies conducted since 2005, we found 39 studies that could be described as impact evaluations that used (implicit or explicit) comparison groups to measure attributable change. However, these too were deficient in many ways: 29 had a theory of change but 23 did not show whether the choice of comparison groups was valid (i.e. did not have balance tests); 29 did not discuss the confidence with which their results were measured (i.e. did not undertake power analyses or show sample size calculations) and only five discussed ethical issues.

• **Sectors:** Using gap-maps of evidence, the study finds that most high-quality studies of humanitarian assistance are in the area of health (and particularly mental health), nutrition and peace building.

• **Timing:** Most existing impact evaluation studies examine changes in conditions and resilience once the affected area is in the recovery phase (there are approximately 27 studies that examine the results of peace building and conflict prevention). There are few studies of unanticipated disasters (four) —all of which examine recovery and resilience and few studies (six) of efforts of immediate relief.

• **A needs map:** A needs map drawn from interviews and strategy documents helped us visualise main areas in which practitioners require additional evidence and research. In particular, more than 20 per cent cited accountability, food security, protection, water and sanitation, and health and said that it was important to assess their impact not just on food security but also on nutrition, income and, in the longer term, on recovery and resilience. Education, humanitarian assistance as a whole, nutrition and logistics were said to be important for study by 10–20 per cent and

less than 10 per cent suggested emergency telecommunications and camp management as areas that require additional evidence and research.

Key findings for methods used in impact evaluations of humanitarian assistance

There are many constraints that impact evaluations need to overcome in humanitarian situations, in addition to those that are faced in studies that are undertaken in less complex and challenging situations. The robustness of studies can be especially compromised in the absence of baseline data and inability to plan for and construct counterfactuals. The need for speed of action and low predictability of such situations also means that little advance preparation is possible. Furthermore, most humanitarian situations have a multiplicity of actors and it is usually difficult to de-couple actions and outcomes. High-covariability or the fact that conflict and disasters don't usually have clean boundaries means that it is also difficult to find or establish comparable groups that can serve as counterfactuals in a scientifically robust and ethically sound way. Last but not least, there is a lack of impact evaluation experts in the humanitarian sector and a lack of humanitarian experts in the impact evaluation sector.

However, traditional evaluations that either monitor processes or assess if targets have been achieved are clearly insufficient for robust evidence by themselves. They are unable to examine unintended consequences; or to deal with a variety of biases, such as selection bias (i.e. areas targeted by humanitarian assistance are likely to have attributes that make them more or less likely to recover, compared to the average), non-random attrition (are unlikely to count people who either migrated as a result of the intervention or those who perished as a result of the disaster), and contamination bias (areas targeted by one actor are also likely to have other sources of assistance that may make it difficult to separate the different sources of changes); and are unable to *measure* the change that has occurred as a consequence of their action.

Clearly, new ways must be forged to combine the strengths of methods and traditions that exist in the sector so that these can be used to enhance evidence, while responding to increased demand for accountability in the context of rapid-onset and protracted crises.
Impact evaluations can help answer other questions, such as:

- How much of the change in conditions was a result of the programme? Was the affected population able to recover to their pre-disaster levels? Are we 'building back better'?
- How much of the programme or intervention should be delivered, at what time, and with what frequency?
- What is the best *way* to deliver an assistance package? What difference did it make? Can it be delivered in a more cost-effective manner?
- How much difference did an agency make?
- Were some groups better off as a result of the programme compared to others?

- Do protocols for coordination and planning make a difference? If so, by how much?[1]

Methods for undertaking impact evaluations should therefore address concerns of costs, speed, multiplicity of actors, absence of data, and the ethics of responding speedily to disrupted communities while ensuring that the needs of the most vulnerable are addressed.

We use six case studies in the paper and discuss the possible methods to address these concerns in different humanitarian contexts that range from unanticipated natural disaster-related emergencies to protracted crises. We discuss a variety of situations and present examples of possible methods that may be used to answer important questions. These situations include understanding the effectiveness of ready-to-use supplementary food in a post-conflict situation in Sri Lanka, the effectiveness of trust-building interventions in post-conflict areas in Kyrgyzstan, education for women in earthquake affected areas, the effectiveness of cash-based flood relief on nutrition and food consumption in a complex emergency in Pakistan, the relative effectiveness of food coupons versus cash for internally displaced populations in a protracted emergency in northern DRC, post-typhoon emergency assistance for pregnant women in the Philippines, and an intervention to improve adherence to tuberculosis therapy in South Sudan.

The main findings from these case studies are as follows:

Ethics: In our review and discussions of case studies we use the 'no-harm principle'.[2] Methods used for assessing and measuring changes in outcomes all use approaches that turn the challenge of limited resources and the inability to cover all affected populations at once into an opportunity. The use of phased roll-outs of interventions and rolling out programmes with small changes while keeping the basic package the same (also called a factorial design) are discussed in relation to the effectiveness of nutrition-related interventions and cash transfers. In other cases, where it is not possible to randomly assign packages for an impact evaluation ex-ante, we discuss the use of ex-post quasi-experimental designs (such as propensity score matching and regression discontinuity design) that can be used to assess the effect of short messaging services on TB adherence and programmes to increase the uptake of iron supplements.

Data availability: In all cases, we discuss the use of other available data that may be used creatively in the context of both unanticipated disasters and for protracted crises. So, in the case of examining food distribution in Sri Lanka, training in Kyrgyzstan and post-earthquake education in Pakistan, we discuss the use of the Living Standards Measurement Surveys (LSMS). Additionally, in the case of post-typhoon Philippines, the Demographic and Health Surveys (DHS) may be used to

[1] Humanitarian Accountability Project (HAP) has been on a long quest to demonstrate the 'business case' for beneficiary accountability but was never able to provide evidence that being accountable to beneficiaries using HAP's standards led to improved impact.

[2] The approach to be tested may significantly improve but will not worsen outcomes for emergency relief recipients.

understand minimum detectable effect sizes, calculate appropriate sample sizes and create a baseline. In other cases, such as post-typhoon Philippines, we discuss the use of satellite images and Geographic Information Systems (GIS), and the use of mobile phones for easy and cheap data collection in the case of TB adherence in South Sudan.

Costs and speed: The cost of robust impact evaluations that help to clearly understand and measure changes in outcomes is more than that of no evaluations at all. However, agencies regularly undertake many evaluations, whether these are process evaluations, monitoring or real-time evaluations. Good theory-based impact evaluations contain all of these but go a step further in assessing attributable changes and/or contribution to change. Rapid impact evaluations that use individual assignment and are quick and less costly have been conducted, for example, to study malaria interventions, to assess ready to use supplementary food, and to determine the effectiveness of approaches to increase voluntary medical male circumcision in mostly development contexts. These types of studies can be customised to the needs of the humanitarian community and should be discussed widely. In the six case studies, most examples use individual assignment and lend themselves to rapid evaluations.

Sample sizes: Another concern for most practitioners is the size and cost of data collection. Using the six case studies, we show that it is usually not necessary to collect data on all beneficiaries. Surveys that inform impact evaluations usually need to be conducted only on a subset of the affected population. Using case studies we show that depending on the outcomes, the range of sample sizes required for undertaking good impact studies varies from 300 individuals (for food distribution and its effect on haemoglobin in Sri Lanka), to 690 households in 30 clusters (for comparing cash versus food coupons in DRC), to 2,000 households in 30 clusters (to understand factors affecting trust-building in Kyrgyzstan). Thus, the cost of data collection may be neither prohibitive nor so widespread as to interfere with field operations.

Unintended consequences and vulnerable groups: An important concern for many respondents is understanding the impact on vulnerable groups such as women, children, people living in remote areas and chronically poor populations. Investigating the differential distribution and uptake of humanitarian assistance amongst these groups is clearly important. It is possible to design impact evaluations that can measure this differential impact but also assess other unintended consequence with the help of good theories of change. Another important question raised by respondents is the need for a better understanding and estimating the effect of protocols for increased coordination amongst humanitarian agencies. We present some hypothetical solutions for understanding these in the case of distribution of nutrition packages in Sri Lanka, for example, where a multitude of agencies work. But we advocate for more thought and discussion, especially with respect to what outcomes may be most useful to investigate.

Since 2005, more than US$90 billion has been spent on humanitarian assistance. Fortunately, several agencies around the world are taking the lead and showing that institutional constraints, capacity bottlenecks and concerns about image-risk that are traditionally associated with impact evaluations can be transcended. Yet, very few

impact evaluations are being conducted. This paper represents a first assessment, to our knowledge, of the possible methods that may be used to undertake high-quality impact evaluations of humanitarian assistance.

Three critical steps can help take this effort forward. Firstly, agencies can pilot impact evaluations in areas that are relevant to them and help demonstrate feasibility and practice. Secondly, a dialogue on the priorities and feasibility of these types of studies of humanitarian assistance can help remove some of the myths that surround the use and planning of impact evaluations. Finally, clearly earmarked resources can contribute to building a critical body of evidence that can inform programming and strategy in humanitarian assistance clearly and consistently.

Contents

List of figures and tables

Abbreviations and acronyms

ACTED	Agency for Technical Cooperation and Development
DHS	Demographic and Health Survey
ERRA	Earthquake Reconstruction & Rehabilitation Authority
ICRC	International Committee of the Red Cross
IRC	International Rescue Committee
IFPRI	International Food Policy Research Institute
LMS	longitudinal monitoring surveys
MSDSP KG	Mountain Societies Development Support Programme, Kyrgyzstan
NFI	non-food items
NGO	non-governmental organization
PSLM	Pakistan Social & Living Standards Measurement Survey
PSM	propensity score matching
RCT	randomised controlled trial
RRMP	rapid response mechanism for population movements
RUSF	ready to use supplementary food
RTEs	real-time evaluations
RDD	regression discontinuity design
MDE	standardised minimum detectable effect
TEC	Tsunami Evaluation Coalition
UNHCR	UN High Commissioner for Refugees
UN OCHA	UN Office for the Coordination of Humanitarian Affairs
WFP	World Food Programme

Understanding the impact of humanitarian assistance is an area where much work is needed… Linking impact measurement and accountability better to the funds agencies receive is a key recommendation of this review (Humanitarian Emergency Response Review, UK Government, March 2011).

The evidence base proving which humanitarian responses are most effective is extremely lacking. Investments must be made in the consolidation of evidence about what works in response to different kinds of needs in different contexts (The Use of Evidence in Humanitarian Decision Making, ACAPS Operational Learning Paper, January 2013).

1. Introduction

In 2011, an estimated 62 million people were directly affected by humanitarian crises across the world. The international community responded by raising US$17.1 billion in funding, but more than a third of the needs identified by the United Nations were unmet. [3, 4,5] In a context where lives are in danger and the demand for resources overwhelmingly exceeds supply, effective and efficient delivery of services is key. However, and despite countless ex-post evaluations routinely being conducted in the humanitarian sector, there is a dearth of *theory-based*, *reliable* evidence *causally* linking the interventions with the observed outcomes. The objective of this paper is therefore to examine the extent to which impact evaluation methods can provide evidence to help improve effectiveness and efficiency in humanitarian action.

The use of conventional impact evaluation methodologies in assessing humanitarian action has been meagre so far. Correspondingly, there is a significant gap in the literature on how to conduct impact evaluations in humanitarian emergencies.[6] There are numerous reasons for this. Humanitarian action is usually implemented in emergency situations. Therefore, undertaking impact evaluations can be challenging. Impact evaluations of both rapid-onset and slow-onset protracted crises must deal with a mismatch between resources and needs, disrupted communities, an absence of baseline data, difficulty in collecting information, security concerns, and finding a valid counterfactual. Moreover, there is a lack of impact evaluation experts in the humanitarian sector and a lack of humanitarian experts in the impact evaluation sector.

This paper explores the methodological options and challenges associated with collecting and generating high-quality evidence needed to answer key questions about the performance of humanitarian assistance, including whether assistance is

[3] Global Humanitarian Assistance. *GHA Report 2012*. Rep., 2012. <http://www.globalhumanitarianassistance.org/reports>
[4] Ibid.
[5] "Financial Tracking Service (FTS) Tracking Global Humanitarian Aid Flows." Financial Tracking Service (FTS). UN Office for the Coordination of Humanitarian Affairs. Web. 18 Nov. 2012.
[6] See Bozzoli, C., T. Brück and N. Wald (2013). "Evaluating Programmes in Conflict-affected Areas". In: P. Justino, T. Brück and P. Verwimp, eds. *A Micro-Level Perspective on the Dynamics of Conflict, Violence and Development.* Oxford University Press, Oxford for an exception for the case of discussing impact evaluations in conflict settings.

reaching the right people and at the right time, whether it is bringing about the desired changes in their lives (effectiveness), and whether it is being delivered in the right doses and ways, and with manageable costs (efficiency).

Structure of paper: The rest of this paper is organised as follows: Section 2 defines and discusses humanitarian emergencies and their responses. Section 3 defines and discusses evaluations, focusing on impact evaluations and how best to conduct them both in general and in emergency situations. Section 4 provides a brief conceptual framework for using impact evaluations in humanitarian emergencies, while Section 5 reviews the relevant if small literature. Section 6 discusses how to overcome valid ethical concerns by suitably adopting the research designs. Section 7 reviews a number of case studies, while Section 8 summarises the lessons learnt.

2. Defining and categorising humanitarian emergencies and humanitarian action

Definition of a humanitarian crisis: A humanitarian crisis is a situation in which there is an exceptional and generalised threat to human life, health or subsistence. These crises usually appear within the context of an existing situation of a lack of protection where a series of pre-existent factors (poverty, inequality, lack of access to basic services), exacerbated by a natural disaster or armed conflict, multiply the destructive effects.[7]

There are two differing but linked definitions of humanitarian action that should be noted. The Dunantist interpretation defines humanitarian action as action designed to save lives, alleviate suffering and maintain and protect human dignity during and in the aftermath of emergencies. [8] According to this interpretation, humanitarian action is different from other forms of foreign assistance and development aid because it is governed by principles of humanity, neutrality, impartiality and independence. Furthermore, humanitarian assistance is usually short-term in nature and provides for activities in the immediate aftermath of a disaster. The Wilsonian definition of humanitarian action broadens the scope of humanitarian action to include slow-onset disasters and situations that demand prolonged assistance for human life and health (recent examples include famine in Somalia, the earthquake in Haiti and conflicts in the Democratic Republic of Congo and in Sudan). With this definition, humanitarian action is as much directed at building resilience as it is for providing immediate relief and aiding recovery. This study combines both these definitions and discusses methods for impact evaluations in both these contexts.

Categories of humanitarian crises: A variety of taxonomies have been attempted to help understand humanitarian emergencies and responses to them.

Buttenheim (2009) provides five categories of disasters based on the immediate cause: (i) biological (epidemics, insect infestations, animal attacks), (ii) geophysical

[7] The School for a Culture of Peace, Barcelona <http://escolapau.uab.cat>
[8] "Defining Humanitarian Aid | Global Humanitarian Assistance." Global Humanitarian Assistance. Web. 18 Nov. 2012. <http://www.globalhumanitarianassistance.org/data-guides/defining-humanitarian-aid>

(earthquakes, volcanoes, dry mass movements), (iii) climatological (droughts, extreme temperatures, wildfires), (iv) hydrological (floods, wet mass movements), and (v) meteorological events (storms). We add a sixth category to this list: violent conflict. The addition differs from other events in that, in contrast to the others, which are all categories of natural disasters, conflict is anthropogenic. But as we discuss later in Section VII, conflicts are exacerbated by natural disasters and, in turn, natural disasters and consequent scarcities often create conflict.[9]

Another distinction is between anticipated and unanticipated humanitarian emergencies. Indeed, many humanitarian agencies use this differentiation in planning their actions. It may be argued that few disasters, man-made or otherwise, are truly unanticipated but we differentiate between those that can be predicted with more than even odds. The illustrations below show that although disasters are often unexpected, they are rarely random events. Figures 1 and 2 illustrate historical observations of earthquakes greater than five (> 5) on the Richter scale, and areas of water scarcity around the world that may lead to drought (and in some cases to conflict).

Furthermore, humanitarian crises can occur suddenly (as in the case of an earthquake) or emerge slowly (as in the case of famine). This differentiation also presents a useful way to think about operational and evidence requirements that are different in slow and sudden-onset situations: severe drought conditions do not translate into a famine overnight, while pre-conflict tensions frequently simmer long before an outbreak of armed or other hostilities. Earthquake is, perhaps, the most compelling example of a sudden-onset emergency, while meteorological events usually come with at least a short forewarning.

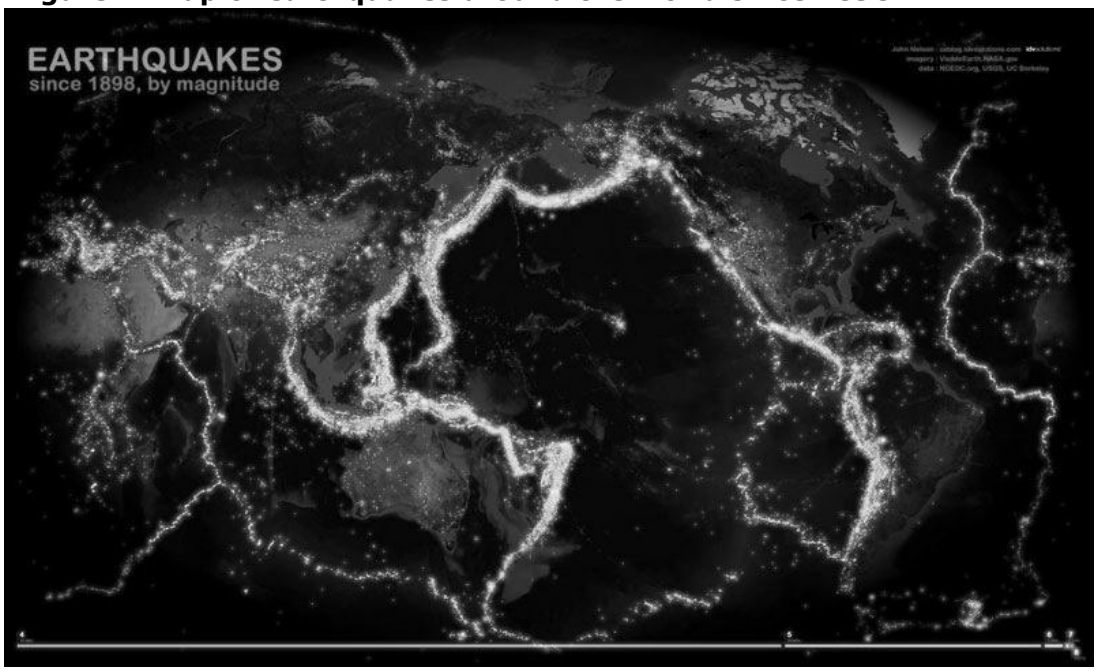
To direct effort and plan for assistance, a number of relief organisations have introduced a policy of tracking crisis hotspots. The World Bank maintains a list of fragile and conflict situations and another of natural disaster hotspots.[10,11] The access to such information can, as we will discuss in more detail later, crucially aid impact evaluation efforts.

[9] Paul Collier. And: Justino, P., T. Brück and P. Verwimp (2013). "Micro-level Dynamics of Conflict, Violence and Development: A New Analytical Framework". In: P. Justino, T. Brück and P. Verwimp, eds. *A Micro-Level Perspective on the Dynamics of Conflict, Violence and Development*. Oxford University Press, Oxford.

[10] "Fragility, Conflict and Violence." *Fragile and Conflict Situations*. World Bank, n.d. Web. <http://web.worldbank.org/WBSITE/EXTERNAL/PROJECTS/STRATEGIES/EXTLICUS/0,,contentMDK:22978911~menuPK:4168000~pagePK:64171540~piPK:64171528~theSitePK:511778,00.html>

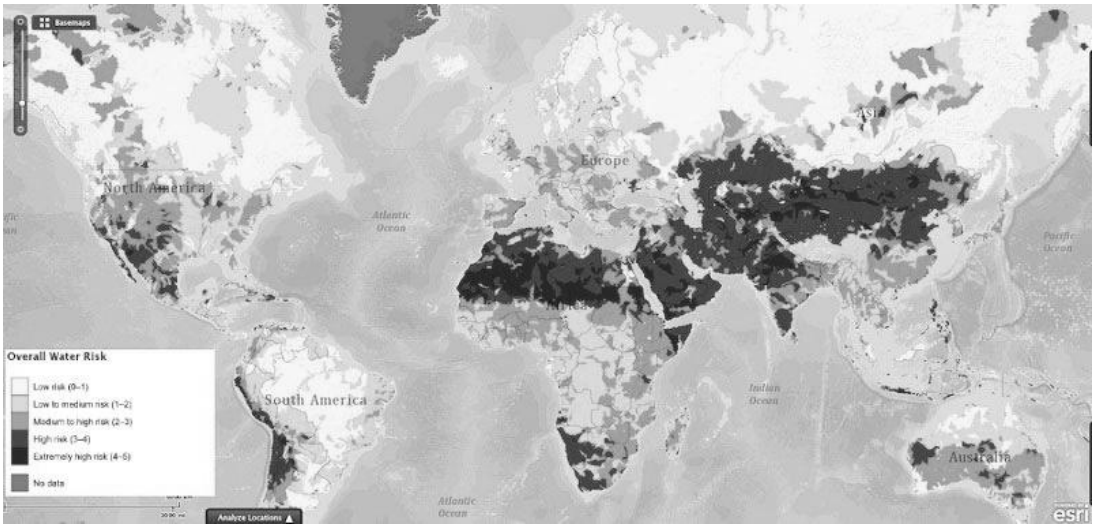
[11] Dilley, Maxx, Robert S. Chen, Uwe Deichmann, Arthur L. Lerner-Lam, Margaret Arnold, Jonathan Agwe, Piet Buys, Oddvar Kjekstad, Bradfield Lyon, and Gregory Yetman. *Natural Disaster Hotspots: A Global Risk Analysis*. Disaster Risk Management Series 5. Publication no. 34423. Washington, D.C.: World Bank, Hazard Management Unit, 2005. Print.

Figure 1: Map of earthquakes around the world since 1898



Source: John Nelson

Figure 2: Overall water risk around the world. Source: UNHCR



Source: WRI

Categorising humanitarian assistance: Another taxonomy distinguishes between preventive action (or action that helps to build long-term resilience and presumably reduce the occurrence of future humanitarian emergencies), and humanitarian assistance that is provided in the immediate aftermath of an emergency irrespective of whether the emergency is rapid-onset or slow onset. [12]

Over time, it is clear that policies and programmes that have improved preparedness and increased resilience have reduced the number of losses and casualties, and have helped reduce the type of emergencies that were traditionally responsible for the

[12] The recent review of the UK's emergency relief placed much emphasis on prevention (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/67579/HERR.pdf), and the question of 'prevention versus cure' features as a case study item below.

largest casualties during the twentieth century.[13] Among the most important insights from Devereux's (2000) analysis of famines in the twentieth century is that while famines during colonial times were directly linked to droughts, after 1980 famines have occurred during conflicts (e.g. those in Uganda, Sudan, Ethiopia, Eritrea, Niger). This is also well illustrated by the casualties recorded over a 200-year period in India. Estimates from the Great Bengal famine in 1769–70 suggest that one-third of the population of the province was wiped out.[14] During the November 1970 cyclone that affected the coastal belt of Bangladesh, approximately 300,000 people perished.[15] Still later, after 1974, Dréze has highlighted the absence of famines in South Asia. [16] In 1997, during a cyclone in Bangladesh, 111 people lost their lives.[17]

These different typologies (types of humanitarian *emergencies* and types of humanitarian *assistance*) have implications for the way impact evaluations may be designed and planned in humanitarian settings. These are discussed in the next section.

3. Defining and discussing high-quality, theory-based impact evaluations

3.1 Various forms of evaluations

Most evaluations in the humanitarian sector use monitoring data and/or are outcome and perception studies or are real-time evaluations (RTEs) (see Box 2). According to the OECD-DAC glossary, monitoring is a "continuing function that uses systematic collection of data on specified indicators to provide management and the main stakeholders of an ongoing development intervention with indications of the extent of progress and achievement of objectives and progress in the use of allocated funds".[18] Monitoring does not include evaluation in terms of addressing efficiency, effectiveness, impact and sustainability, which are the main issues that impact evaluations address.

High-quality outcome and perception studies that use well-collected quantitative data measure the changes in conditions of beneficiaries in the target area before and after a programme. Such outcome and perception studies can thus help inform whether the targets of the programme were achieved in the area targeted. Although these studies are important, they do not tell us whether the change in the status would have been the same without humanitarian assistance, or whether the change was *caused* by the intervention, or due to some other programme.

[13] Devereux 2000
[14] Menon 2013
[15] WB 2010: 34
[16] Lessons from South Asian famine prevention are not, as von Braun et al (2009) rightly point out, directly transferable to areas of Africa that remain famine prone: African droughts remain responsible for the largest number of casualties (WB 2010: 27).
[17] Menon 2013
[18] OECD-DAC "Glossary of Key Terms in Evaluation and Results Based Management", 2010.

Often in humanitarian contexts, outcome studies employ qualitative data that use small and unrepresentative samples of the affected population.[19] These also include surveys that are limited to the humanitarian community, including donors, officials, NGO representatives and volunteers who work in the field after the disaster. While there is a certain value in conducting qualitative surveys, they frequently pose interpretation challenges.

Similarly, other evaluation tools such as key stakeholder interviews (KSIs) are important ingredients of impact evaluations, but *by themselves* are insufficient for producing robust analyses of attributable impact, for the reasons discussed in this paper (confounding factors, attrition, non-random placement and biased response).[20]

Box 1: Requirements of robust impact evaluations

Robust impact evaluations in any sector require minimally a few things:

- a well-defined theory of change;
- good formative research to understand the context and background of the initiative;
- explicit or implicit counterfactuals that help measure what would have happened in the *absence* of the intervention;
- qualitative and quantitative baseline and end line data;
- a well-defined set of beneficiaries and outcome variables;
- identification methods that use these data to quantifiably measure changes in outcomes that may have occurred due to the intervention; and
- the ability to use evidence in other situations and contexts.

Source: White, H, 2011.Conducting theory based impact evaluations

[19] See, for example, two studies in Jordan and Haiti, respectively: Washington K, CARE. 2010. "Material assistance and emergency cash assistance evaluation" and
Mandel, J. and E. Sommerfeldt. 2010. "Closing the loop: responding to people's information needs from crisis response to recovery to development. A case-study of post-earthquake Haiti."

[20] See, for example, Few *et al.* 2014

Box 2: Perception studies vs. Real-time evaluations vs. Impact evaluations

Perception studies	Real-time evaluations	Theory-based Impact evaluations
• Usually qualitative data • Sometimes do not use representative sample • Short-term • Give a good idea of uptake by beneficiary population.	• Use qualitative and quantitative data • Frequently do not use representative samples • Evaluate processes • Focus on development and implementation of the programme and assess if implementation outputs have been achieved.	• Use qualitative and quantitative data • Include RTEs and perception studies to inform theories of change and implementation of programmes • May be used to measure short-, medium-, and long-term effects • Long-term policy implications • Have a theory of change • Are able to measure attribution and contribution of the programme to overall outcomes.

3.2 Impact evaluations in non-emergency settings

Theory-based impact evaluations can measure and understand the change in outcomes, outputs or long-term impact *caused* by a policy, programme or intervention. They help to understand why, what and how these changes occurred and any unintended consequences of programmes. Additionally, theory-based impact evaluations can inform lessons for other situations. Evidence-based theories of change and analyses of outcomes and impacts for sub-groups and contexts enables the application of lessons for other situations.

Box 1 summarises prerequisites of impact evaluations conducted in "standard", non-emergency settings (White2011). The first step in an impact evaluation is to construct a causal pathway that links activities with processes and outputs, and articulates assumptions required in the hypothesised causal pathway. Formative research and familiarity with the context are important ingredients when developing such theories of change. Creating a credible identification strategy that isolates the change in outcomes/impacts as a consequence of activities is the next step. Most identification strategies require building credible implicit or explicit counterfactuals. Appropriate counterfactuals help attribute impact to the intervention, policy or programme.

Dealing with attrition and response: Factors such as non-random attrition (that is, frequently the better off or the most vulnerable are not counted in evaluations because they are the first to migrate or to perish during a natural disaster or a conflict) and non-random response (the most accessible areas are the ones that get relief first but are also, other factors held constant, less vulnerable than others) can all be accounted for by impact evaluations.

Using impact evaluations to reduce biases: There are primarily three types of biases that evaluations traditionally encounter. These are selection bias, information bias and contamination bias. Selection bias occurs when the most privileged are likely to get relief programme first, not accounting for other covariates such as higher education and higher income, which will also affect how effective the relief programme is for this group. Information bias occurs when the information is biased by the perceptions of the respondents. Contamination bias occurs when the programme or the intervention spills over to non-targeted areas and any assessment of impact provides incorrect results. All these concepts are discussed later in this paper.

Heterogeneous effects: Good theory-based evaluations can also uncover heterogeneous effects. Three of the four evaluations undertaken by the Tsunami Evaluation Coalition (TEC) found that aid was disbursed disproportionately to areas that were easily served by transportation, rather than based on need, and that the old and disabled were often excluded from benefits because they were poorly informed about them.[21] But the TEC could not determine how much worse off these vulnerable groups were. Similarly, an evaluation of poverty impacts in China discovered that the poverty package had significant gains for certain groups – poor with better schooling gain more than others. They thus found that the use of community based beneficiary selection reduced overall impact. [22]

3.3 Impact evaluations in emergency settings

Clearly, impact evaluations become much more complicated to implement when done in humanitarian emergencies. The unique methodological and ethical challenges that arise when doing impact evaluations in humanitarian emergency settings depend (a) on the nature of the emergency (armed conflict involves additional ethical challenges when compared to an earthquake), and (b) on the post-emergency phase and outcomes that the evaluation is focusing on.

Factors that make impact evaluations more difficult in humanitarian contexts include:

- Complex contexts: humanitarian crises are often unanticipated and teams find baseline data do not exist. Furthermore, usually there are a multiplicity of actions and interventions occurring all at once and agencies don't always know how long they will stay in a location, despite their response being precipitated by humanitarian crises (this has been the case, for example, in Somalia, Haiti, Pakistan, Rwanda, and the Democratic Republic of Congo)[23]. All these features make planning for an

[21] Cosgrave, J. (2007) Synthesis Report: Expanded Summary. Joint evaluation of the international response to the Indian Ocean tsunami. London: Tsunami Evaluation Coalition.
[22] Chen, Shaohua, Ren Mu, and Martin Ravallion. Are There Lasting Impacts of Aid to Poor Areas? Evidence for Rural China. Working paper no. 4084. Washington, DC: World Bank, 2008. Print. Policy Research Working Paper Series.
[23] An emergency declaration protocol declares an emergency. Most humanitarian agencies dealing with an emergency stay in the affected area for the first three or six months. For

impact evaluation challenging because there are a multitude of interventions, changing activities and outcomes, and unclear timelines.

- Need for speed: speed and coverage of interventions are critically important in humanitarian assistance. Usually there is no time to train teams and plan an evaluation. Arguably, this constraint is much more critical in the case of unanticipated rapid-onset emergencies than slow protracted crises.

- Multiplicity of actors: after a disaster, many international agencies, donor countries, foreign nationals, domestic and foreign non-governmental organisations (NGOs), national and local governments, their militaries, and others, including relatives and friends of the victims and business communities unaffected by the disaster, may provide financial and technical assistance.[24] The World Bank (2006) reports that the number of international and domestic actors that respond to a disaster has been growing during recent years, but that their roles are not fixed and have blurred over time.[25] For example, the Indian Ocean Tsunami had 42 international agencies and this number does not include national and local agencies. It is, therefore, hard to *ex ante* plan impact of a set of cohesive actions for an impact evaluation.

- Attribution: on a related topic, the array of actors not only makes their coordination a challenge, but creates difficulties for attribution of the impact to a particular programme.[26] Attribution refers to (i) ensuring that a causal pathway runs from the intervention to the outcome, and (ii) accurately isolating and estimating the particular *contribution* of the intervention.[27]

- High co-variability: large areas are often affected during a humanitarian emergency. It is difficult to identify counterfactuals because it is not easy to find locations or beneficiaries that look similar to the affected population (but were not affected) or populations that were affected but not targeted for reasons unrelated to their conditions.

- Evaluations of preventive action: for humanitarian assistance-related activities that are directed at prevention rather than post-emergency assistance, it can be difficult, for ethical and technical reasons, to construct an explicit counterfactual for an intervention that seeks to prevent a severe drought from developing into a famine. Similarly, an escalation of tension from developing into a full-blown conflict.

---

example, the International Rescue Committee (IRC) is constitutionally required to open a country desk if it stays for longer than six months in a country.

[24] Williams 2009; Paul 2006

[25] For example, during the 2004 Indian Ocean tsunami the number of international aid agencies was 42 but this number does not include other actors who provided assistance like the private sector.

[26] See, for example, the critique in Williams (2009) and Paul (2006)

[27] Leeuw and Vaessen 2009

In addition, dealing with bias when conducting impact evaluations is at least as important as it is in non-emergency settings.

Selection bias: An important component of impact evaluations is assessing what would have happened to disaster-affected households that received humanitarian or post-disaster assistance, had they not received it. *Ex ante*, households that are affected most by a disaster are more likely to be vulnerable and suffer damage, than those that are not. This likelihood is likely to be affected by a variety of factors such as poverty, location and other socio-economic characteristics. This is the challenge of 'double selection bias': people who are most likely affected by the disaster are those who are most vulnerable, but humanitarian assistance, in turn, is determined frequently by other factors such as access and voice, factors that are often negatively correlated with vulnerability status. All these factors are also likely to affect the probability of the success of a humanitarian assistance programme.

Neglecting to account for these factors means that assessments of impact and causes that have brought about impact may be flawed. This is the problem of bias of programme placement. Thus, naïve evaluations only measure outcomes for households that have remained in an area after a disaster; they ignore households that may have completely perished in the disaster (in the absence of aid) and ignore households that may have moved (before or after aid was delivered). So, measurements of impact of humanitarian assistance are likely to be understated or overstated. In an analysis that uses counterfactuals to mimic what would have happened without assistance, it is important to find households that are similar to the ones that were affected, with the exception that they were not affected by the disaster. This is challenging in the context of emergencies, because identifying these counterfactual or comparison groups may be impossible and ethically inappropriate.

Example: An example of how to deal with selection bias is the use of a natural experiment. One such study is the contribution of assistance packages provided to Rohingya refugees in Bangladesh.[28] Commissioned by the UNHCR and the World Food Programme (WFP), the study aimed to assess the contribution of food aid to refugee-affected populations. The Rohingya refugees had been living in camp settlements in south east Bangladesh for more than 20 years. However, from 1992 the government stopped recognising any Rohingya refugees who immigrated to Bangladesh. Therefore, only 24,000 of the total 200,000 refugees were officially recognised. In the impact evaluation, Nilsen and Jahan (2012) use a **natural experiment** to evaluate the effects of assistance on the two populations and on host communities. Since the only difference between the registered and the unregistered refugees was when they immigrated to Bangladesh, which is unlikely to be correlated with the outcome of the policy, and the WFP was only allowed to disburse aid to registered refugees, the unregistered refugees served as the counterfactual for the evaluation.

[28] Nielsen, Nicolai S., Kate Godden, Gana Pati, Md. Mamun-ur-Rashid, and Omar F. Siddiki. *The Contribution of Food Assistance to Durable Solutions in Protracted Refugee Situations; Its Impact and Role in Bangladesh: A Mixed Method Impact Evaluation*. Rep. N.p.: World Food Programme, 2012. Print.

In the evaluation design, there were 349 registered refugee households, 620 unregistered households, and 100 host community households in nearby villages that were considered for the evaluation. The evaluation focused on examining outcomes such as livelihoods and coping strategies, movements, protection and the protective environment, and food security and nutrition. The results showed that despite assistance, registered refugees were significantly less economically active, and overall earned less income, than unregistered refugees. Frequency of child labour and youth employment was also more frequent for unregistered refugees. In addition, household expenditures were significantly lower for registered refugees compared to unregistered ones. In comparison to the host communities, all Rohingyas had significant protection concerns. However, in terms of nutrition, dietary diversity scores among the unregistered Rohingya households were the lowest.

Information bias: In the absence of baseline information, many assessments of interventions use information recall to assess changes in welfare. However, it is expected that, during a crisis, respondents do not accurately remember details of their housing, schooling or livelihoods prior to the emergency. Recall error is further compounded if beneficiaries are interviewed by officials associated with the recovery effort. Furthermore, error in self-reporting is likely to be correlated with the severity of exposure to a disaster. This indicates that evaluations of interventions need to be planned as early as possible.

One way to alleviate the consequences of information bias is through the use of mixed methods that can help to triangulate results through a variety of information collection techniques. Impact evaluations that use other sources of data and information such as spatially explicit information, census data and other surveys help to alleviate these errors. This is discussed in section 6.

Contamination bias: Contamination bias occurs in humanitarian assistance when people outside of the intended targeted area receive benefits that were originally not intended for them (thus reducing intensity and amount of the dose in a dose-response equation). Contamination bias can also occur if there are other contributors to the effort in the affected area that may affect the implementation and the impact/outcomes achieved by the intervention/programme. This will make it harder to measure the benefits that can be attributed to the programme.[29]

Example: In the case of Pakistan, the Earthquake Reconstruction and Rehabilitation Authority (ERRA) conducted a social impact assessment and concluded that changes in welfare experienced by earthquake-affected households from baseline to post-intervention were all attributable to ERRA. [30] However, this was an erroneous conclusion because household welfare may have been affected by additional factors as well. For example, remittances are likely to have played an important role in improving welfare.

[29] Note that if other contributors/implementing agencies are equally active in the counterfactual and treatment areas, the impact evaluation is unaffected.
[30] Buttenheim (2009) *Impact Evaluation in the Post-disaster Setting: A Conceptual Discussion in the Context of the 2005 Pakistan Earthquake*. Working paper no. 5. N.p.: 3ie, WP5.

3.4 Objectives of impact evaluations

While recognising that there are many constraints to impact evaluations, they can still be very useful in dealing with biases, in supplementing findings of real-time evaluations (RTEs) and outcome and perception studies, and in providing lessons that are useful for programmatic delivery and overall strategy. Questions that impact evaluations can help to understand include (but are not limited to):

- Magnitude: how *much* did people affected by a crisis gain because of the assistance? Was it delivered in the *right amount*? Would the targeted population have gained more had the intervention *intensity* become less or more? How much would the additional benefit have been? Would the additional costs be mitigated by the additional benefits?

- Implementation: did the intervention *increase the resilience* of the affected population or would they have undertaken steps in any case to bring this about? To what extent did social safety networks, remittances, community coordination and migration alleviate the effects of the disaster? To what extent did humanitarian assistance make an *additional difference* to outcomes? Was there a *better way* to bring about this outcome? Are some training methods *more* effective for volunteers and workers that deliver aid compared to others? Do distance/mobile technologies actually work in providing information, in changing behaviour and in communicating disaster warnings? How should these messages be formulated and delivered so that they are most effective? What interventions work to combat gender-based violence in disaster and relief settings?

- Planning: how can programmes be *better planned* and managed to deliver more effectively? Can planning and operations be made more effective and under what circumstances is this possible? What is the *best way* to deliver assistance amongst competing alternatives of delivery? What are the possible roles for displaced persons in camp management and service provision?

- Coordination and contribution: what was the overall impact of a humanitarian appeal effort? What was the impact of the contribution of a specific agency on the overall achieved impact? What technological options are available and effective for improved relief coordination? To what extent are these effective?

- Cost-effectiveness and impacts on marginal groups: were some population sub-groups (e.g. men, groups living near roads) *better off* compared to other groups? *What and how much* were the overall welfare and distributional consequences of recovery efforts? What is the effectiveness and cost-effectiveness of innovative and new initiatives such as child-safe spaces and ready to use supplementary food? What demobilisation strategies work best?

- Other comparisons: are cash transfers more effective than vouchers to attain outcomes related to child nutrition? What are the optimal content and

packaging for instructional material for emergency relief packages to ensure proper use (e.g. for water treatment supplies)?

Impact evaluations of humanitarian assistance can provide many important insights that are relevant for policy and action. For example, in an impact evaluation of the assistance delivered during the 2005 Pakistan earthquake, the evaluation showed that public schools were much more likely to be damaged by the earthquake than private schools. This evaluation could have important implications for schooling decisions. It could also have implications for the way public schools are constructed in the future.[31] In another impact evaluation of post-conflict reconstruction in rural Sierra Leone, while a community loan programme was successful in conveying material benefits to community members, there was no significant impact on collective action and cohesion.[32] Similar evidence has been found for community driven development assistance in post-conflict situations that have contradicted established hypotheses regarding the effectiveness of policies and programmes in these settings.[33]

3.5 Methods for impact evaluations

There are experimental and quasi-experimental identification methods that may be used in impact evaluations and that may be applied to humanitarian settings. Table 1 presents the description of these methods and the pros and cons of using these methods in complex humanitarian situations.

Table 1: Impact evaluation methods

Methods	Description	Pros	Cons
Experimental design			
Randomised controlled trial (RCT)	A sample of eligible subjects is randomly assigned into those who receive the intervention and those who do not. Impact is the difference in outcomes between the two groups.	- Straight forward estimation (difference in means).	- Requires a comparison group; - Requires check of balance (i.e. whether randomisation was successful). If randomisation is not successful, then the results are not valid.

[31] Buttenheim, (2009) *Impact Evaluation in the Post-disaster Setting: A Conceptual Discussion in the Context of the 2005 Pakistan Earthquake*. Working paper no. 5. N.p.: 3ie, WP5.
[32] Casey, Glennerster and Miguel (2011) "Reshaping institutions: Evidence on external aid and local collective action" National Bureau of Economic Research, Working paper No. 17012.
[33] See, for example, Casey, K., Glennerster, R. and Miguel, E., 2012. *The GoBifo project evaluation report: Assessing the impacts of community-driven development in Sierra Leone.* 3ie Impact Evaluation Report 3. New Delhi: International Initiative for Impact Evaluation (3ie).

Quasi-experimental design			
Difference-in-difference	Outcomes of programme beneficiaries and non-beneficiaries are compared before and after the intervention. The relative change in outcomes is the impact of the programme.	- This approach deals with the problem of unobservable differences between treatment and comparison groups.	- Requires baseline data; - Requires comparison group; - Responsibility of ensuring balance in levels and trends is on the research team and usually requires a lot of data to ensure.
Regression discontinuity	A cut-off determines who is eligible to participate. Outcomes of beneficiaries and non-beneficiaries close to the cut-off line are compared.	- Does not require baseline data although it's desirable to have it.	- Requires comparison group; - Requires establishing that the comparison group at the cut-off is similar to the treatment group.
Matching	Programme beneficiaries are compared to a group of non-beneficiaries that is constructed by finding people whose observable characteristics are similar to those of the people in the treatment group.	- Does not require baseline data, except "matching variables" (that can be obtained from secondary data sources such as RLMS, DHS, etc.).	- Requires a comparison group; - Requires data on "matching variables"; - Assumes there are no differences in unobservables.
Instrumental variables	Participation in a programme can be predicted by an incidental factor, or "instrumental" variable, that is uncorrelated with the outcome (other than by predicting participation).	- Does not require baseline data; - The counterfactual is determined by the programme.	- Requires strong assumption that the instrument affects the outcome only through one specific channel that affects selection but does not directly affect the outcome.

Case-control (from medical studies)	An observational study is one in which subjects are not randomised to the exposed or unexposed groups; rather, the subjects are *observed* in order to determine both their exposure and their outcome status and the exposure status is thus not determined by the researcher.	- Can be less costly than RCTs: no randomisation involved and fewer subjects observed; - Similar to matching: do not require baseline data, except data on "matching variables".	- Requires comparison group; - May require more waves of follow-up data (tracking mechanism); - The burden of proof for showing that the subjects are comparable and that no other reason may have brought about the observed change in outcome.

Source: Hempel and Fiala (2012), Glennerster and Takavarasha (2013) Notes: Subjects may mean individuals, households, districts and other administrative divisions, schools, health centres or other units of observation.

a) Randomised controlled trials

The main advantage of randomised controlled trials is their simple design, which could be applied in many settings, and straightforward impact estimation. If randomisation is successfully implemented, then impact is estimated by comparing the means of the outcome variable between the treatment and the control groups. In humanitarian contexts, however, randomised controlled trials are more difficult to implement than in development contexts. First, random assignment into treatment and control groups should be conducted before the implementation of the programme. This could lead to delay in humanitarian assistance when it is urgently needed. Second, simple randomised control trials assume that control groups do not receive any treatment. In humanitarian contexts, and during the relief, especially, this could be considered unethical.[34] However, variations of the randomised control trial such as factorial design, the pipeline approach and pair-matched randomisation can be used in contexts where such concerns arise. These variations allow researchers to take advantage of programmatic realities (programmes are hardly ever rolled out in one-shot because of constraints in field operations; programme staff are frequently unsure about what mechanism works most effectively, such as cash or in-kind transfers).

The robustness of randomised control trials is that if the randomisation is correctly done, it is correct to assume that all other confounding factors that may also affect the outcome of interest are controlled for.

[34] Hempel and Fiala (2012) argue that random assignment may be even more ethical than any other method for two reasons: (i) uncertainty of programme impact (withholding intervention from one group would be better if a programme has, for example, unintended negative side effects); (ii) limited resources (in reality it is usually difficult to benefit everyone and the vulnerable population may be excluded for certain reasons. Random assignment makes the assignment process 'fair').

b) Difference-in-difference

Among quasi-experimental designs, double difference (or difference-in-difference) is used frequently. However, not only does this require baseline data, but its two main assumptions are quite strong: first, it presumes that the observed characteristics are a good proxy for any unobserved attributes of the treated and comparison sample. Second, it assumes that the differences between the intervention and comparison groups are constant over time. This assumption needs two rounds of baseline so that the time trend assumption can be tested (otherwise, our results will be invalid if we find after the follow-up data collection that the differences are non-constant over time). Taking into account the difficulties associated with data collection right after the disaster, this may increase the cost of the evaluation substantially.

The main advantage of the difference-in-difference method is that it 'differences out' the individual effects or, in other words, the fixed characteristics that are inherent to treatment and control groups. Double difference technically eliminates this problem for observed and unobserved 'fixed effects'. Many researchers combine both randomisation and difference-in-difference methods in order to increase statistical power.

c) Regression discontinuity design

As with difference-in-difference, regression discontinuity design (RDD) may be used if randomisation is not possible or if the evaluation study starts after the programme begins. The application of RDD is quite contextual; that is, it can be used if the selection into the programme has been made based on continuous ranking cut-off (e.g. household monthly income per capita, or test scores of pupils). The advantage of this is that RDD does not require any change in the programme design.

One concern is that the RDD restricts the evaluation to the marginal recipients and non-recipients of emergency relief, thus excluding those further away from the threshold who are likely to have been more seriously exposed and who are expected to benefit disproportionately from the relief effort.

RDD is also less statistically powerful than randomisation and requires larger sample sizes (Bloom 2012). On the other hand, unlike other methods it does not require baseline data collection (although, again, baseline data collection is desirable so that a match may be validated).

d) Matching

An important advantage of matching is that, provided we are able to identify a valid comparison group, the evaluation study using matching can be planned even after a programme has begun. This could be particularly useful considering the difficulties in coordination with implementing agency that may arise after the disaster.

If we already have information on pre-programme characteristics of households, individuals, health facilities, schools or any other unit of observation we are interested in, then matching can be undertaken. Matching characteristics may be obtained, for example, from the census or from longitudinal monitoring surveys

(LMS) that may be present in a country prior to a disaster. The problem with using this approach after disasters is that if the disasters lead to high migration (or high mortality that is unregistered), we would have difficulties finding the households or individuals from the census or LMS. Therefore, it is more likely that we would need to collect a large sample survey before the intervention in order to be able to draw a valid comparison group, which could increase the cost of the evaluation.

Another disadvantage of this method is that matching could only be done using the observed characteristics of the individuals, while the characteristics such as personality or motivation that are intangible and difficult to measure will likely bias the results.

e) Instrumental variables

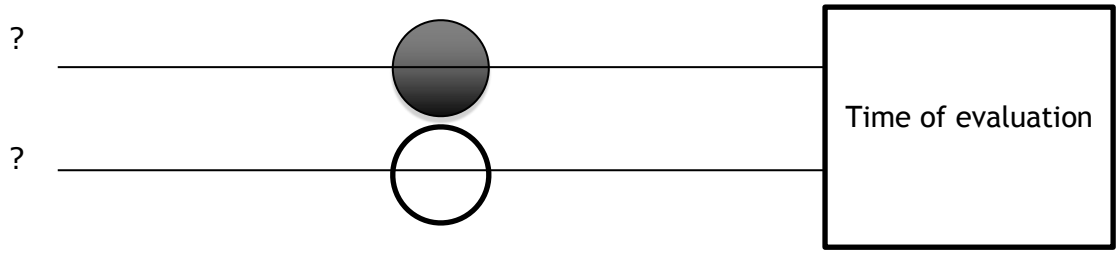
As with other quasi-experimental methods like RDD and matching, instrumental variables is a method for identifying and measuring causal change in outcomes, after the intervention begins. This can be extremely useful in a disaster context.

The main challenge associated with this method is that it requires finding a valid instrumental variable. The instrumental variable is one that is uncorrelated with the main outcome of interest, and correlated with participation in the programme. This then helps us to understand participation or selection into a programme. This can be challenging in most cases. This method does not require baseline data collection and, despite being technically challenging, it could be associated with lower costs than the other methods.

f) Case-control studies

Case-control studies, which are borrowed from medical literature, are observational studies in which two existing groups, different in outcome, are identified and compared in order to identify factors that contributed to this outcome (e.g. disease). Schematically, this method can be illustrated as in Figure 3:

Figure 3: Case-control studies



Case-control studies are similar to the matching method, in that the treatment and control groups are matched based on observed characteristics, and therefore require collecting data on 'matched variables'. The burden of proof to show that no unobserved characteristics are influencing the outcome and behaviour of the control group is upon the researcher, and this can be quite onerous. However, the advantage of the case-control studies, as with other quasi-experimental methods, is that they do not require any change in the programme design.

4. A conceptual framework for using impact evaluations in humanitarian emergencies

In this section we suggest a conceptual framework to capture the types of situations in which impact evaluations may be used. Buttenheim (2009) focuses on sudden onset emergencies and uses the 2005 Pakistan earthquake as a case study. This provides a starting point. Buttenheim divides a sudden onset emergency into the following main phases[35]:

> **(t_{-1}) Baseline:** this is the pre-disaster phase. Most agencies assume that no data exists for this phase.[36]
>
> **(t_0) Emergency:** this is the point (or period) in time when the disaster or conflict occurs. Poverty, social inequality, poor governance and fragility of populations and institutions affect and amplify disaster impacts.
>
> **(t_1) Relief phase:** emergency relief is provided in the immediate aftermath of the disaster and typically as soon as access is restored. This phase usually lasts for three-six months unless the crisis is protracted.
>
> **(t_2) Recovery phase:** longer-term assistance is provided to aid recovery to the pre-disaster 'condition' and to strengthen resilience. This phase starts usually six months after the emergency.

We expand Buttenheim's (2009) framework to include other emergency events and assistance including famines and preventive action, as follows:

> **(t_{-2}):** is the period much before an adverse shock during which the escalation of tension may trigger and eventually result in an emergency situation (conflict).[37] Some crises follow a slow onset trajectory – for example, the gestation period between a severe drought and famine conditions could be weeks, months or even years (Devereux 2000). For some emergencies, therefore, household welfare (e.g. health) deterioration and asset depletion may set in long before the point at which the shock occurs, and then gradually escalate towards the catastrophe. This presumes an emergency discontinuity or threshold beyond which negative health and other gradients deepen and risks to human life, health and other losses dramatically accelerate.
>
> **(t_{-1}):** is the period immediately before the emergency. We make this distinction because a **(t_{-1})** baseline is straightforward for an earthquake, but

[35] Buttenheim, (2009) *Impact Evaluation in the Post-disaster Setting: A Conceptual Discussion in the Context of the 2005 Pakistan Earthquake*. Working paper no. 5. N.p.: 3ie, WP5.

[36] However, pre-disaster administrative or survey data are often present and can be extremely useful. The Pakistan case cited above had LSMS data from 2000 that can be employed usefully.

[37] **(t-2)** is absent from Buttenheim since earthquakes are sudden onset emergencies

not for a famine or conflict. Since a gradual welfare deterioration may begin long before the emergency itself, a **(t_{-1})** based estimate of 'normal' household welfare and asset holdings will be biased *downwards,* thus introducing an *upward* bias in a **(t_{-1})** based assessment of restorations to 'normalcy' during the recovery phase.

(t_0): is the point of time during the emergency (or immediately after it). This raises the question of whether there is a 'correct' time to measure welfare or asset holdings. As noted, emergency environments are often chaotic, with breakdown in public and other service delivery (e.g. water, sanitation, health care). In such environments, indicators or measures of health or human welfare may rapidly deteriorate. The main objective of humanitarian assistance is to quickly arrest such slides.

(t_3): is a point in time much after the disaster during which activities are directed at reducing the likelihood of future disasters/conflicts and to reducing the likelihood of large damages if there is recurrence of the disaster/conflict, by reducing the vulnerability of populations.

In the diagram below, we summarise the different phases of observation and phases of intervention and outcomes/impact that can help us think through different situations during which impact evaluations may be usefully employed.

Using this diagram, impact evaluations may be used to measure changes in a variety of outcome indicators in the following ways ('X' denotes a variable that measures welfare or health or any other indicator of household or individual well-being):

$Xt_{-1} - Xt_0$ = household welfare loss induced by the emergency, which includes losses to health/life and/or asset loss/destruction

$Xt_{-2} - Xt_{-1}$ = household welfare loss induced during pre-emergency drought/escalation of tension, which includes gradual deterioration of health/depletion of assets and deterioration of quality of life

$Xt_{-2} - Xt_0$ = total household welfare loss from adverse shock, which can be decomposed into **$(Xt_{-2} - Xt_{-1}) + (Xt_{-1} - Xt_0)$**

$Xt_1 - Xt_{-1}$ = extent of household recovery from the emergency

$Xt_1 - Xt_{-2}$ = extent of household recovery from the adverse shock

$Xt_2 - Xt_{-1}$ = sustained restoration of households to baseline during sudden onset emergencies (where t-2=t-1)

$Xt_2 - Xt_{-2}$ = sustained restoration of households to baseline during slow onset emergencies, e.g. conflicts and famines (where t-2 > t-1).

Using this framework, we discuss the importance of determining **the right time for collecting data** on outcomes (see Box 3 for a technical exposition). Outcome

measures are likely to be highly sensitive to the time after the emergency and there is a high likelihood that measurements are biased unless care is taken to specify the point of time at which outcomes are measured.

The importance of specifying the point of time for outcome measurement is illustrated by the following example: it is well known that high incidence of diarrhoea and dehydration does not occur immediately after an emergency event, but that it worsens rapidly soon after. Therefore, if an evaluation examines the changes of a programme that aimed to change children's health outcomes and compares the diarrhoea levels much further along the timeline with diarrhoea levels two months after the disaster (for example) when diarrhoea is at its worst, we are likely to overestimate the positive impact that the programme had.

On the other hand, if we compared diarrhoea levels two months after an emergency aid programme had started with diarrhoea levels just before the disaster, this is likely to underestimate the impacts of the humanitarian assistance programme. In this instance, the measurement issue can also create perverse incentives for humanitarian assistance. So for example, if assistance arrives late, and assistance is measured as the change in outcome from a baseline where the baseline is measured much after the disaster has occurred and health indicators have worsened dramatically, then naïve estimates of the contribution of assistance will erroneously attribute a large change to emergency assistance, as compared to assistance that was provided in a much more timely manner.

Box 3: Recognising the importance of timing when collecting data on baselines and outcomes— a technical exposition

In humanitarian contexts, it is especially important to recognise that contexts change rapidly. This, in turn, implies that data values change much faster than in stable developmental contexts. Let us take the example of health variables. Health related measures deteriorate rapidly, but at different rates, after an emergency event. Let h_{max} and h_{min} represent the max value of a health indicator h once the emergency sets in, while h_{min} is the lowest value h can attain if permitted to deteriorate. Such deteriorations may be highly non-linear and involve critical thresholds. Let us consider the case of dehydration and diarrhoea for children. Suppose that assistance arrives quickly after an emergency event and that a slide from h_{max} to h_q occurs in a 'treatment' group with a corresponding slide from h_{max} to h_{qc} in a control group. Let $h_q > h_{qc}$ i.e. let's assume that the treatment does affect the health indicator, so that a positive average treatment effect is observed.

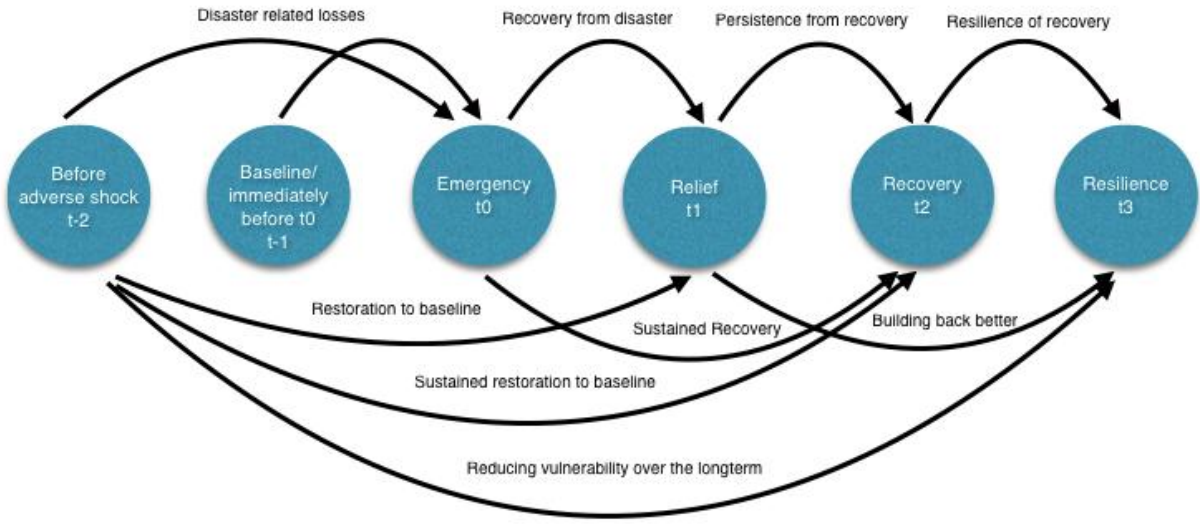
Challenge 1: measurement error in impact estimates

If the health deterioration process is a continuous process, the point in time where h is measured matters. Suppose, first, that we measure h at the 'wrong' point in time, so that $h_{qm} > h_q$ and $h_{qcm} > h_{qc}$. While the correct impact estimate of the impact of a treatment is (h_q-h_{qc}), the obtained estimate is $(h_{qm}-h_{qcm})$. Whether this result is biased or not will depend on the shape of the relevant 'health deterioration function'. To anchor this in a real-life situation, let us assume that the evaluation team arrives an hour after the emergency and starts its survey work on child health indicators right away. Three problems will now be encountered: (a) while, e.g., dehydration and diarrhoea incidence may rapidly accelerate, this acceleration is not instantaneous; (b) stretching out the survey in time will introduce measurement errors when comparing, e.g., children surveyed early and those surveyed later on; and (c) by starting too early, the survey will mis-measure the full impact of the emergency (and underestimate the impact of the emergency assistance on diarrhoea and dehydration).

Challenge 2: perverse incentives and reversal of rankings

A more serious problem occurs if assistance arrives late, so that $h_l < h_q$ (i.e. health indicators deteriorate more because of the late arrival, which is realistic if we think of, e.g., diarrhoea and dehydration) and ditto for $h_{lc} < h_{qc}$: the impact estimate now becomes $(h_l - h_{lc})$ and we observe a 'perverse incentive' problem if $(h_l-h_{lc}) > (h_q-h_{qc})$. In this case, humanitarian assistance can be expected to generate larger positive impacts, if conditions are allowed to worsen. This involves the risk of ranking a highly effective and early arrival intervention as inferior to a less effective late arrival intervention, since the impact size of the latter, because of the slide, may be larger.

Figure 4: Stages of emergency



Using this framework, we examine some of the literature in this area in the next section.

5. Impact evaluations of humanitarian assistance: a review of the literature

Overview

The Inter-Agency Standing Committee classifies types of activities undertaken in delivering humanitarian assistance into clusters:[38]

[38] Source webpage: https://clusters.humanitarianresponse.info/about-clusters/what-is-the-cluster-approach.

Box 4: Clusters of activities in humanitarian assistance used by the United Nations

	Camp Coordination and Camp Management		Logistics
	Early recovery		Nutrition
	Education		Protection
	Emergency Telecommunications		Shelter
	Food security		Water, sanitation and hygiene
	Health		

Although approximately US$90 billion has been spent on humanitarian assistance, few rigorous impact evaluations have been implemented.[39] In one assessment of several databases of evaluations of humanitarian interventions, we found that of more than 900 evaluations, only 38 were impact evaluations.[40] In the rest of this section, we review the impact evaluation literature in this area. The review organises impact evaluation studies by phase of emergency action.

5.1 Emergency relief

Very few studies address the impacts of emergency relief. Such studies evaluate the impact of material relief and are often conducted in camps after a conflict or disaster. Table 9 in the Appendix lists six studies, with the type of intervention, main outcomes, identification methods and main findings. These studies include impact evaluations of food programmes implemented jointly by World Food Programme (WFP) and United Nations Higher Commissioner for Refugees (UNHCR) in Bangladesh;[41] food vs. cash programmes in Colombia by WFP & International Food Policy Research Institute (IFPRI); water cleaning in the refugee camps in Malawi (Roberts *et al.* 2001) and Liberia (Doocy, S. and G. Burnham 2006); food distribution programmes in Chad (Huybregts *et al*. 2012); and reconciliation and

[39] Global Humanitarian Assistance. *GHA Report 2012*. Rep., 2012. <http://www.globalhumanitarianassistance.org/reports>

[40] This number does not look at impact evaluations that examine the impact of humanitarian assistance that may be aimed at increasing resilience or efforts aimed at preventing famines. See Puri and Khosla (2013) for more information.

[41] The food programme was also jointly implemented by WFP & UNHCR in refugee camps in Chad, Rwanda and Ethiopia. These studies are named "impact evaluations", although no rigorous method was used in evaluation.

psychological programmes in Uganda (Bolton *et al.* 2010).[42] These studies employ RCTs (stratified, sometimes with several variations in treatment) with random assignment to control, factorial design or a natural control group.

Finally, there exists literature that does not use rigorous impact evaluation methods of humanitarian assistance and so could not be attributed to the list of studies mentioned above. This literature uses secondary sources of data and no formal statistical techniques in analysing disaster relief programmes. However, these studies also deserve attention in the context of disasters as they include rich qualitative analysis of the relief programmes. Some of the examples include works by Dréze (1991) on the famine in India, and Kunreuther (2006) on the hurricane Katrina.

5.2 Recovery and resilience

5.2.1 Impact evaluations of anticipated emergencies

Most impact evaluations of humanitarian assistance measure programme effects during recovery or resilience stages, which may be several years after the disaster or conflict. Most impact evaluations examining this phase of humanitarian assistance focus primarily on the impact of community-driven development programmes of peace building and stabilisation projects in fragile states.

Samii, Brown and Kulma (2012) review 25 most recently completed or ongoing impact evaluations of *stabilisation* interventions in post-conflict countries. Building on the review by Samii *et al*.,[43] Gaarder and Annan (2013) explore (i) evaluation design issues in conflict-affected situations; (ii) evaluations as interventions, and the implications for the risks and reliability of results; (iii) the importance and value-added of impact evaluations in post-conflict situations; and (iv) ethical concerns about impact evaluations in conflict prevention and peace-building. Fourteen of these studies employ randomised control trials (RCT) and 10 studies use quasi-experimental methods.[44] Counterfactual are chosen using one of the five types of methods: (i) random assignment; ii) delayed random assignment or the pipeline approach; (iii) factorial model; (iv) matching; and (v) natural control group. Table 10 shows a list of these studies, the methods they used, the interventions examined and the main conclusions of these studies.

This list can be updated with one recently completed impact evaluation study on female participation in local governance programme in Afghanistan by Beath, Christia and Enikolopov.[45] Beath *et al*. (2013) randomly assign 500 Afghan villages

[42] A similar study was conducted by Staub, E. *et al.* (2005) "Healing, reconciliation, forgiving and the prevention of violence after genocide or mass killing: an intervention and its experimental evaluation in Rwanda"; however, with no random assignment to treatment, so it is not listed as a rigorous impact evaluation.
[43] Gaarder and Annan list 24 out of 25 studies from Samii *et al*. Their list does not include the baseline report by Humphreys, M. 2008. "Community-driven reconstruction in the Democratic Republic of Congo". Columbia University and International Rescue Committee.
[44] See Annex A pp. 26-27 in Gaarder and Annan
[45] Beath *et al*. 2013 "Empowering women: Evidence from a field experiment in Afghanistan" World Bank Policy Research Working Paper 6269.

while clustering proximate villages to control for spillovers to test the impact of the National Solidarity Programme on women's empowerment and gender equality. Comparison villages are chosen using matching techniques and quantitative baseline, and end line data collection was collected for 13,000 individuals. The study found that the development programme increasedfemale mobility and involvement in income generation, but does not change female roles in family decision-making or attitudes toward the general role of women in society in rural Afghanistan.

A few other studies are undertaken by IRC, and focus on ongoing development programmes to understand their impact on (i) women's empowerment in Burundi and Côte d'Ivoire; (ii) poverty reduction and a parenting intervention in Burundi; (iii) access to education; and (iv) a mental health intervention and savings intervention in DRC. These studies examine variously the impact of these programmes on recovery and resilience stages of humanitarian assistance. These studies also use RCTs with baseline and end line data collection and employ mixed method approach. The counterfactual is chosen through one of the following methods: (i) simple random assignment; (ii) delayed treatment control group; (iii) factorial design, or a combination of delayed treatment control group and factorial design.

To summarise, Table 10 in the Appendix presents all reviewed studies of peace-building and conflict prevention interventions, methodology used to identify and measure impacts, and the main results.

5.2.2 Impact evaluations of unanticipated emergencies

There are very few studies that focus on unanticipated natural disasters. Two studies investigate the impact on outcomes related to recovery and resilience.

The first study by De Mel, McKenzie and Woodruff (2010) investigates the recovery of private firms in Sri Lanka following the 2004 Indian Ocean tsunami.[46] Using dataset of 209 enterprises, authors employ four-arm RCT with delayed treatment control group. They randomly assign four types of treatments to firms: two values of monetary grant are distributed either as cash or in-kind. By comparing treated firms with comparable firms, they found positive effect of grant programme on profits, representing a 9.9 per cent real monthly return on the treatment. Further, direct aid is more important in the recovery of enterprises operating in the retail sector than for those operating in the manufacturing and service sectors. By comparing different treatments, they also found that the use of cash grants is more helpful than the use of in-kind, but only in limited cases.[47]

Using evidence from floods in Bangladesh in 2004, Shoji (2010)[48] found that a newly introduced policy, which allowed rescheduling savings and instalments, acted as a

[46] De Mel *et al.* (2010) "Enterprise recovery following natural disasters" The World Bank Policy Research Working Paper 5269.
[47] International Federation of Red Cross and Red Crescent Societies Mid Term Shelter Tsunami http://www.ifrc.org/Global/tots-mid-term-review.pdf: This study uses random sampling (not random assignment) and studies the effects of Transitional Shelter Programme in Indonesia.
[48] Shoji, M (2010) "Does contingent repayment in microfinance help the poor during natural disasters?" Journal of Development Studies, vol. 46, no. 2, pp. 191-210.

safety net during natural disaster by decreasing the probability that people skip meals during negative shocks by 5.1 per cent, with a higher effect on females and the landless. However, the authors did not estimate the effects on nutritional outcomes, and no conclusions could be made about whether these households are better off nutritionally. The author employed random sampling and, by using the instrumental variable method, compared the same 326 households after and before the introduction of the rescheduling policy using recall data.[49]

Table 11 in the Appendix provides the summaries of the studies, with methodology used and main findings.

5.3 General discussion on methods used by studies

For this part of the study, we directly reviewed 38 studies, of which 30 studies are impact evaluations of peace-building and conflict prevention interventions, only two studies are impact evaluations of unanticipated disasters covering recovery and resilience periods, and six are impact evaluation studies of humanitarian relief. Of the 38 studies:

- **16** studies had a formal test of balance between intervention and control groups, i.e. a test whether a control group and an intervention group are similar in observed characteristics, proving that the randomisation was successful;
- **10** studies did power analysis when selecting an optimal sample size for the evaluation;
- **29** studies had a narrative of underlying economic theory;
- **Only five** studies mention ethical approval or discuss ethical concerns during evaluation. These are the studies of humanitarian relief programmes; and
- **most (27)** studies used randomised control trial (RCT) as an identification method to select subjects for intervention and for control, with the remaining **12** studies using quasi-experimental methods.

Ethics: To select comparisons while still maintaining ethical standards, most studies use two types of methods: factorial design and delayed treatment (or a combination of both). When using factorial designs, researchers use a range of two–four treatment arms. For example, in the study by IRC in Burundi, the households were first randomly selected to participate in the village savings and loans association intervention. Of those selected, half of the households were randomly selected and assigned either to a waitlist control group (phasing in) that receive treatment in the future (first arm), or a treatment group that immediately receive the treatment (second arm). Of the households in the treatment group, half were selected to also

[49] Buttenheim (2010) describes two impact evaluations that were conducted after the 2004 Pakistan earthquake: the ERRA study and the World Bank and South Asia region study. The ERRA study, which will be discussed later, used before- and after- comparison of the same group of households. This method has its limitations as the estimated effect may capture the effects of other factors or interventions. The World Bank and South Asia region study used quasi-experimental methods to assess the impact of housing and livelihood grant programmes. The publications of evaluation studies are not available.

participate in the family-based discussion group (third arm). [50] However, note that few studies discuss explicitly the ethical validity of imposing such treatment alternatives (or of withholding them).

6. Using appropriate methods to overcome ethical concerns

Concerns about the use of impact evaluations in humanitarian settings are often anchored in ethical concerns. Given that experimental approaches distinguish between treatment and control groups, the non-treatment of a control group in its conventional meaning is, many argue, simply not acceptable in an emergency situation. We accept this general sentiment.

We also assert that experimental approaches are very versatile and can help put such ethical concerns to rest. A good research design can account for ethical concerns while delivering important learning.

The do no-harm principle: In impact evaluations, we propose the use of a 'do no-harm' principle that is used in medicine; we re-phrase it as 'the approach to be used may significantly improve but will not worsen outcomes for emergency relief recipients'. A few examples can help deal with these ethical concerns.

Example 1: An important question in humanitarian assistance is whether transfers should be given cash or in-kind, during the emergency stage, and whether reverting to cash transfers is preferable once normality is restored.[51] Assessing this choice can be done easily, without violating the rights of people affected by humanitarian crises. This case is frequently encountered, especially since local markets disappear temporarily during emergencies and price gouging is frequently encountered.

To assess the effectiveness of alternative strategies such as cash transfers versus in-kind transfers, it is possible to create two treatment groups through a lottery so that households randomly receive either a cash transfer or an in-kind transfer. The outcomes from these two groups can then be compared with each other, without violating any ethical concerns, to assess the relative effectiveness of one programme against the other. Figure 5 shows this method schematically.

Another important question in emergency relief is whether targeting emergency relief to women is more effective than a general targeting of households. Testing this intervention can also use the same method where some households are randomly selected to either get relief where anyone in the household can come and receive the relief package, or, the household is required to send the woman to receive the relief package. Again, since the evidence is sparse about what works best, this random selection into one of two groups can be done without violating any ethical concerns.

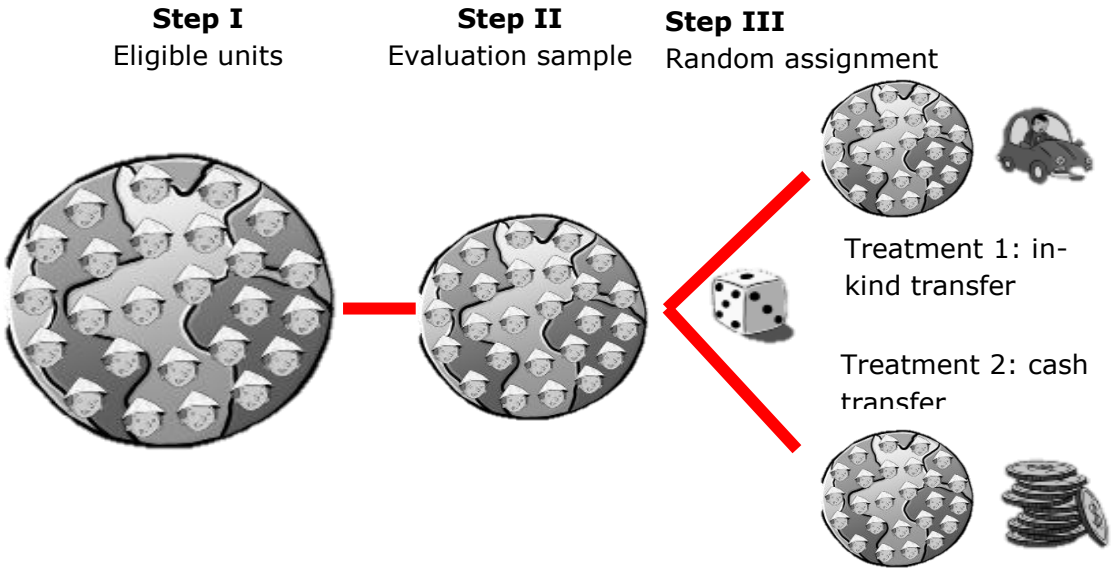
---

[50] IRC (2011) "Urwaruka Rushasha: A Randomized Impact Evaluation of Village Savings and Loans Associations and Family-Based Interventions in Burundi".

A key problem during the emergency phase of a disaster is that local markets often destabilise e.g. price gouging occurred in the aftermath of Hurricane Charley in 2004.

[51] Increases in food prices very quickly erode the purchasing power of any cash.

Figure 5: Illustrative figure showing possible randomisation design for testing the effectiveness of a cash transfer against an in-kind transfer in a humanitarian context



Source: Credits for schematic are with Gertler *et al.* The figure has been adapted to illustrate this example.

Example 2: We consider a logistics example here. A crucial and controversial question in the literature on famines in Africa is whether famine deaths occur because of starvation, increased susceptibility to infections from nutritional deprivation, or from exposure to the disease environments of refugee or other camps.[52] Irrespective of the merit of each of these explanations, logistics affect health hazard exposure particularly for impoverished, nutritionally deprived or otherwise vulnerable population groups. The organisation of an emergency operation needs to build in the risks that alternative logistical set-ups expose relief recipients to, while recognising that a standardised best practice may work well in one setting but be very harmful in another.

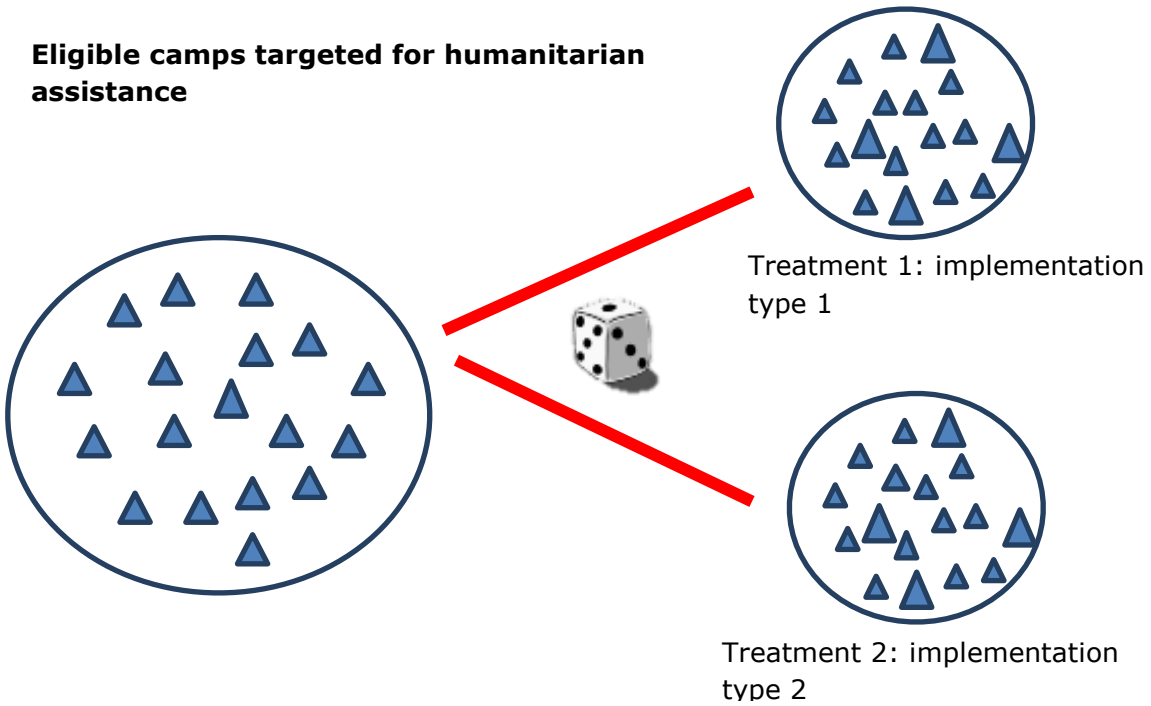
Ethical concerns can once more be mitigated by introducing tweaks to established 'best' practice that satisfy the no harm principle.[53] In Figure 6 we show this schematically. The two different arms of the impact evaluation provide the same programme targeting the same outcomes, but are implemented in two alternative ways. Since the impact evaluations are testing two different forms of delivering an intervention (or testing, for example, training workers in two different ways), these types of interventions are tested at the camp level.

[52] de Waal 1989 as laid out in Devereux 2000

[53] Apart from paying attention to how camps should be organised and the health and other gains from alternatives, there are also potentially valuable lessons from systematic reviews of hygiene, water and sanitation interventions by 3ie that document the returns to hand-washing and other improvements in personal hygiene practices. Hugh Waddington and Birte Snilstveit *(2009)* "Effectiveness and sustainability of water, sanitation and hygiene interventions in combating diarrhoea." *Journal of Development Effectiveness Volume 1, Issue 3, 295-335.*

(We submit, that the critical constraint in this context is not an ethical one. Instead, for robust evaluation designs that use camps as the unit of assignment, and not individuals, the big constraint is likely to be having the requisite number of camps to get robust estimates. These obviously need to be examined on a case by case basis and depend on the type of delivery or implementation methods being tested, but we advise a minimum of 30 camps in each group.)

Figure 6: Illustrative figure showing possible evaluation design for comparing implementation methods seeking the same outcome, across camps in a humanitarian context



Example 3: A third example of dealing with possible ethical concerns is illustrated by providing trauma counselling in the short, medium and long term. Norris (2005) reviews a number of mostly small-sample studies that highlight the presence of post-disaster psychological disorders.[54]

Very few studies have followed larger samples of affected people over several years. A study of Aceh tsunami survivors' post-traumatic stress reaction scores showed that such effects persisted for more than a year, but subsequently declined across the board irrespective of whether people received treatment/counselling or not.[55] While these results may have been specific to Aceh, either culturally or because of the types of treatments that were available there, it is clearly very important to know whether counselling or other psychological treatment and the combination of the two can (a) help mend psychological health and (b) speed up the economic and other recovery of individuals and households through improvements in mental health.

[54] WB: 2010: 47
[55] Frankenberg *et al.* 2009

Large n trials that treat people with individual or combination treatments over long periods of time can help determine what is most effective in mental health treatment. However, it is probably not ethically acceptable to conduct randomised studies that withhold treatment from traumatised persons in the control group, especially if these control group members are also interviewed about the trauma suffered. Again, factorial designs that provide everyone the standard treatment, but then randomly allocate additional/new treatment, can help to understand what types of treatments are most effective in helping with post-traumatic stress while avoiding the ethically unsatisfactory case of withholding treatment to control group members.

Example 4: Pair-wise matching is another technique that is often used. In an evaluation of the impacts of a radio programme in Rwanda, researchers wanted to find out whether a soap opera (called *New Dawn*) implemented in 2004 to promote inter-ethnic reconciliation after genocide and war was successful in changing the individuals' own beliefs about the other ethnic group, whether it had changed perceptions of norms related to prejudicial behaviour, and whether it was leading to greater cooperative behaviour in practice.[56] One hundred and twenty communities were matched first into pairs, and within pairs communities were randomly assigned to either being a control community or a treatment community. In treated communities, listening groups were organised to listen to the radio programme. In comparison communities, listening groups were organised to listen to an alternative (health) programme at the same time as *New Dawn* was being aired. That way, all communities received some benefits. This example provides an example of an innovative way in which an experiment may be created. In the end, the evaluation found that the radio programme was influential in changing people's norms about what constituted 'acceptable behaviour', but was not influential in changing their own behaviour.

Methodological innovation: The four examples above underscore an important attribute of emergencies: humanitarian agencies usually design assistance packages that contain multiple interventions across a variety of sectors. This represents an opportunity for methodological innovation. Just as in medical interventions, a basic care package is provided to both the control and treatment group it is possible to provide a basic care package to all affected. The possibility is then to employ the '**factorial method**' to assess the causes and pathways of impact. The effectiveness and efficiency of innovative or untested interventions can thus be tested if one group of beneficiaries is provided with these interventions, but not the other (*everyone* receives the basic care package). This method helps to measure the incremental contribution attributable to a humanitarian assistance package. These designs are extremely flexible and can be scaled up easily to compare different types of interventions.

[56] Staub, Ervin, Laurie Anne Pearlman, Alexandra Gubin, and Athanase Hagengimana. "Healing, Reconciliation, Forgiving and the Prevention of Violence after Genocide or Mass Killing: An Intervention and Its Experimental Evaluation in Rwanda." *Journal of Social and Clinical Psychology* 24.3 (2005): 297-334. Print.

Example: An example of this is an impact evaluation conducted by Action Contre la Faim. [57] Action Contre la Faim wanted to examine the impact of ready to use supplementary food (RUSF) on wasting in children aged 6 to 36 months. Huybregts *et al.* (2012) examined the question using a two-arm cluster randomised trial. Households in both arms got the general food distribution package.[58] However, in addition to the food package, the intervention arm receives supplementary food. Households in this arm were given 46 grams of RUSF daily for four months. The results of the evaluation showed that adding RUSF to existing food programmes did not reduce the cumulative incidence of wasting. However, targeted children in the treatment group did record higher levels of haemoglobin concentration. The supplement also resulted in significantly lower levels of self-reported diarrhoea.

Such methods can also be used to examine and compare the relative effectiveness of cash, non-food and in-kind transfers during an emergency, and in the relief phases of assistance to examine if there are differences in outcomes and welfare, and if these change with the phase during which they are implemented.

Using other sources of data to alleviate concerns about ethics: We now discuss using other data sources to alleviate ethical concerns about impact evaluations. The main requirement in impact evaluations is to introduce or exploit a variation that helps to either naturally or artificially create comparison groups and intervention groups that allow us to understand what would have happened in the absence of the intervention. This variation needs to be exogenous to the intervention being examined, i.e. not affected by it nor affecting it. RCTs create this exogenous variation by random selection of who gets the treatment and who does not. However, many other sources of variation can be exploited. One such opportunity is served by spatially disaggregated or GIS data that allows us to use this variation.

The use of GIS data: The application of geographic information databases has been relatively unexplored, but has great potential. These are spatially explicit databases that have data for every layer (variable) for each pixel (data point). GIS can contain physiographic data on, for example, weather, elevation, slope, location and distance. In an impact evaluation of protected areas in Thailand, it was found that the use of protection might be overstated, after one accounts for the fact that areas that are usually protected are those that have low agricultural productivity, and that the likelihood of them being cleared for cultivation is lower than otherwise imagined.[59] The study used physiographic attributes such as elevation, slope and location attributes and found that including socio-economic factors such as population density and travel time weighted distance to the market, which are usually used to explain the opportunity cost of clearing land, did not affect the estimates. Using the exogenous variation in physiographic variables that are relatively easily available provided the opportunity to use instrumental variables.

---

[57] Huybregts L, Houngbe F, Salpeteur C, Brown R, Roberfroid D, *et al.* (2012) The Effect of Adding Ready-to-Use Supplementary Food to a General Food Distribution on Child Nutritional Status and Morbidity: A Cluster-Randomized Controlled Trial. PLoS Med 9(9): e1001313. doi:10.1371/journal.pmed.1001313

[58] 2009

[59] Cropper, Maureen, Jyotsna Puri, Charles Griffiths. *How the Location of Roads and Protected Areas Affects Deforestation in North Thailand*. 2001.

Currently, much data are being collected using either satellites or mobile phones, both of which represent cheap and quick methods for collecting rich, spatially disaggregated data that can be used for undertaking impact evaluations without violating ethical concerns.

7. Case studies

In this section, we discuss several *hypothetical* case studies. The purpose of this section is to highlight the scope for applying impact evaluation methods to disaster relief and recovery interventions. Each case study will describe: (i) the intervention and the challenge; (ii) important implementation questions for the evaluation; (iii) the method of identification (RCT/quasi-experimental); (iv) the unit of assignment and unit of observation; (v) the indicators that are important to track and measure during the short run and the long run; (vi) the sample size requirements; and (vii) the lessons that may be learnt.

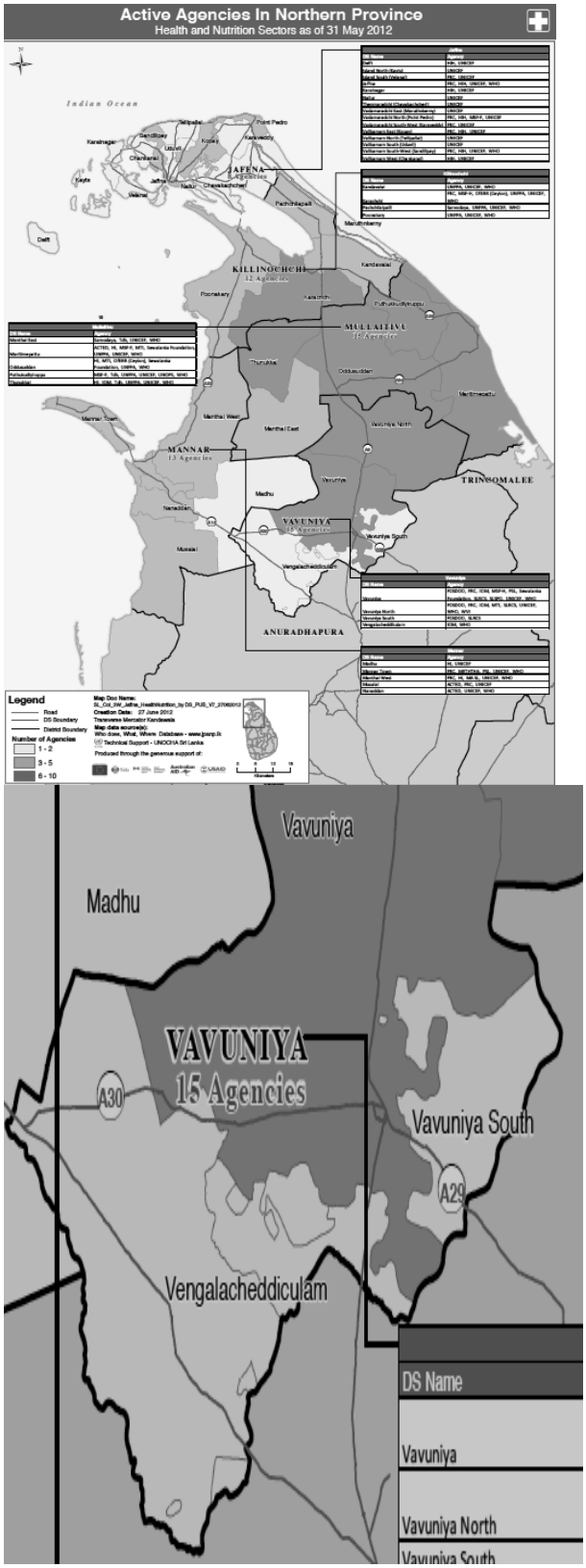
Case study 1: Multiple interventions or a multi-agency intervention

Challenge: with multiple actors and multiple interventions, it is difficult to isolate the impact of the intervention of one actor from that of another, especially if both interventions target an improvement of the same outcome for the same group of people. In this case study, we provide an example of the multiple interventions after the emergency and discuss difficulties and possible solutions for conducting impact evaluations that aim to measure the impact of a single agency. We illustrate this by examining two cases: the 2004 Indian Ocean tsunami, and the 2010 conflict in the south of Kyrgyzstan.

Case study 1a: The 2004 Indian Ocean tsunami – the case of Sri Lanka

Background: in 2004, the Indian Ocean tsunami affected multiple countries. India, Indonesia, Sri Lanka and Thailand were all hit hard. About 275,000 people were killed, tens of thousands were injured and 10 million left homeless and displaced. Sri Lanka was one of the most tsunami-affected countries, with approximately 35,400 people killed, 23,100 injured, and 500,000 displaced. Approximately 114,000 homes were destroyed or damaged by the disaster. As a response, 637 camps and welfare centres provided temporary shelter to displaced persons. In January 2005, more than 180 agencies and NGOs were operating in Sri Lanka. In 2012, UN Office for the Coordination of Humanitarian Affairs's (OCHA) "3W" (Who, What, Where) survey published the distribution of agencies located in the Northern Province of Sri Lanka, operating in health and nutrition sectors (see left sub-figure of Figure 7). The number of agencies varied from eight to 15 per district.

Figure 7: Distribution of humanitarian agencies in health and nutrition sector in Northern Province of Sri Lanka



Source: UN OCHA

Intervention description: suppose that after the disaster, an agency such as Agency for Technical Cooperation and Development (ACTED) intervenes with a general food distribution programme, covering all villages in Vavuniya South district of Northern Province, Sri Lanka (see right sub-figure of Figure 7). Existing evidence suggests that nutritional supplements can prevent wasting and reduce anaemia in populations at risk of periodic food shortages. Simultaneously, another Sri Lanka NGO decides to intervene with an additional food programme in the same region. The national NGO aims to distribute ready to use supplementary food (RUSF) to households with children aged 6–36 months. In order to evaluate the effectiveness of additional food programmes, and to be able to separate the effects from the ACTED general food distribution programme, the NGO intervenes in a neighbouring district, Vengalacheddiculam, in which ACTED does not operate. The NGO decides to intervene only in bordering Vavuniya villages. As is frequent in such cases, let us assume that the NGO can operate only in a few villages.

Evaluation design: a study that uses a factorial design to investigate the relative effectiveness of general food distribution through RUSF is laid out in Table 2. ACTED is working in all households in Vavuniya South to distribute food. Simultaneously, the Sri Lankan NGO is working to distribute RUSF to supplement this action. If the NGO wants to understand its incremental and additional contribution to nutrition and health, and since it is likely constrained by resources, it distributes RUSF to a subset of households in affected villages. Since all villages and households are affected by the tsunami, ACTED and the national NGO randomly assign households in two areas to one of three treatments: 1) households in Vavuniya South that are also receiving general food packages from ACTED also receive RUSF from the national NGO; 2) households in neighbouring Vengalacheddiculam that are outside of ACTED's implementation area are randomly chosen to receive RUSF by the national NGO; 3) households in Vavuniya South that are outside of the NGO's control receive only general food packages from ACTED. Since allocation of RUSF is random amongst households in affected villages, this scheme allows comparing general food distribution vs. RUSF vs. general food distribution and RUSF. Schematically, the treatment arms can be illustrated in the first three columns of Table 2:

Table 2: Identification design for case study 1

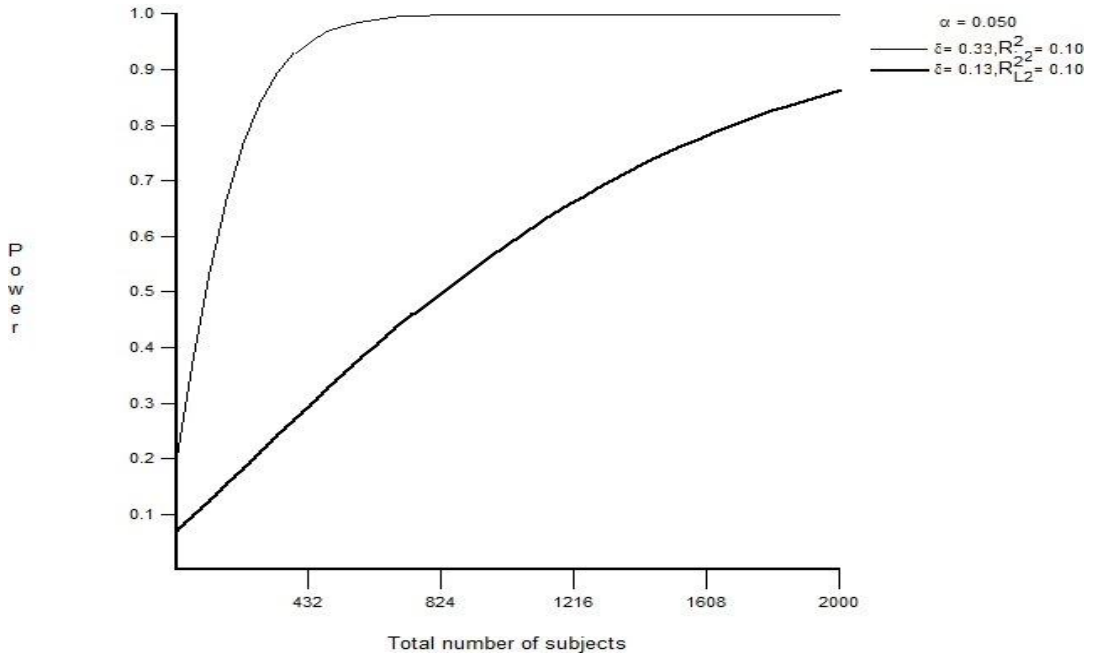
Treatment type	Geographic area	Agency on the ground	Sample size (hhs)	
			2% min. detectable effect	5% min. detectable effect
Comparison group: general food distribution	Parts of Vavuniya South where the Sri Lankan NGO is not present	ACTED	570	100
Treatment group 1: RUSF	Vengalacheddiculam	Sri Lankan NGO	570	100
Treatment group 2: general food distribution + RUSF	Parts of Vavuniya South where the Sri Lankan NGO is present	ACTED and Sri Lankan NGO	570	100

Data and outcomes: quantitative data for impact evaluation are collected before RUSF distribution and after one month following RUSF distribution, to measure the short-term impact from the programme. Units of observation are households and individuals. The key indicators for this hypothetical impact evaluation are captured by a health survey of children aged 0–5 years. The main outcome of interest is the reduction of anaemia among children, which may be measured by the increase in haemoglobin levels of children 6–36 months old.

Power analysis: a simple random assignment for households in the Vengalacheddiculam district enables this impact evaluation. We use the following assumptions for sample size calculations. Significance level: 0.10; power: 80 per cent. Using the health section of the Sri Lanka Income and Expenditure Survey 2006/2007, we assume that the mean value of haemoglobin levels among children aged 6–36 months is 100 g/l (cut-off value to diagnose anaemia is 110 g/l). Standard deviation is 15.

Sample size: suppose that we want to be able to capture a 2% and 5% increase in haemoglobin levels in children. Standardised Minimum Detectable Effect (MDE) for 2% is equal to 2/15=0.13, and for 5% is equal to 5/15=0.33 (δ in Figure 8 below). Therefore, the optimal sample sizes required to capture a 2% and 5% increase in the outcome variable are 300 and 1,700 households respectively (see Figure 8). The total sample sizes are equally split between the comparison group and treatments 1 and 2 groups (see the last two columns in Table 2).

Figure 8: Power analysis for the hypothetical evaluation, case study 1, Sri Lanka



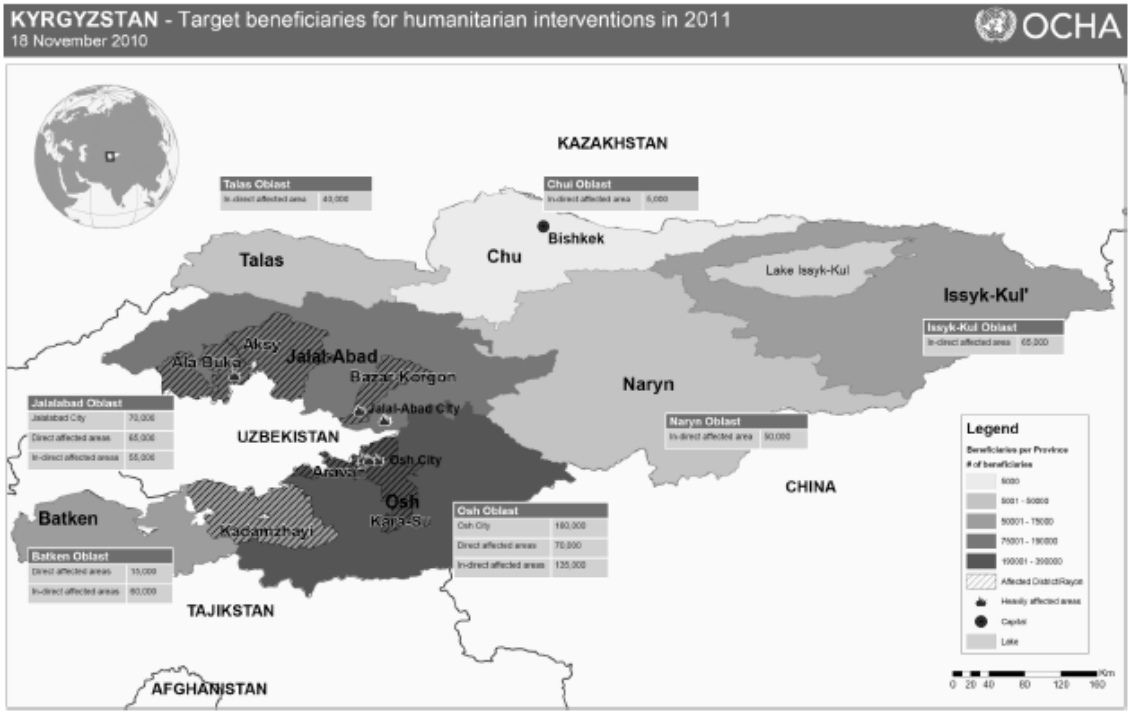
Case study 1b: The 2010 conflict in Kyrgyzstan

Background: in June 2010, political and social tensions climaxed in violent inter-ethnic clashes in Osh City and surrounding areas in the south of Kyrgyzstan, which forms a part of the Fergana Valley and where many Uzbek communities, the country's largest ethnic minority, live. As a consequence of the violent clashes, 400 people were killed according to official statistics. Over 2,500 were injured and, at the peak, 400,000 were displaced with approximately 100,000 crossing as refugees into Uzbekistan. Large-scale destruction of public and private property, especially housing, occurred, notably in the urban centres of Osh and Jalalabad. Unofficial estimates put the number of dead at 2,000.[60] Reports say that the civil conflict led to a poverty level increase of about 2 per cent.

As a consequence, since 2010 many international organisations and international and local NGOs (UN agencies, IRC, ICRC, Save the Children, Mercy Corps, ACTED, Interbilim, etc.) started implementing numerous interventions in Kyrgyzstan. The interventions included humanitarian relief and reconstruction. Figure 9 below presents a map of the most conflict-affected areas and target beneficiaries for humanitarian interventions put together by UN OCHA. The most conflict-affected oblasts include Osh and Jalalabad (marked as red in Figure 9).

[60] Melvin, 2011

Figure 9: Kyrgyzstan. Conflict affected oblasts Source: UN OCHA



More than three years since the conflict (during the stage from recovery to resilience), many agencies are implementing numerous peace-building educational and community development programmes. One of the community development programmes is currently implemented by MSDSP KG (Mountain Societies Development Support Programme, Kyrgyzstan), the implementing arm of the Aga Khan Foundation Kyrgyzstan.

Intervention: the programme is funded jointly by Aga Khan Foundation USA and the World Bank, and is planned for the years 2014–2017. The programme is aimed at promoting social cohesion and building social capital in fragile or post-conflict environments in mono-ethnic and multi-ethnic oblasts in Kyrgyzstan. In particular, this project will identify specific approaches to enhance social cohesion that could be effective in the Kyrgyz context and have potential for integration into a community-driven development approach. The project will then pilot these approaches through community engagement and mobilisation in the delivery of targeted community-driven micro-projects.

Suppose that the programme budget is restricted to intervene only in Osh oblast, one of the most directly conflict-affected areas in the south of Kyrgyzstan. Osh oblast comprises both mono-ethic and multi-ethnic communities. The project is targeted at both types of communities, in order to compare the effects from the intervention in mono-ethnic and multi-ethnic areas. The level of intervention is *aiyl aimak* (administrative unit in Kyrgyzstan), which comprises several villages. Further, the budget allows implementing micro-projects in 15 *aiyl aimaks*.

Data: since the beginning, the impact evaluation team and implementing agency have worked in close collaboration, and the impact evaluation is planned and

designed before the start of the programme. Taking into account the multiplicity of actors and interventions in the southern regions, the impact evaluation team decides to collect as much information about existing interventions as possible. The implementing agency provides a full list of *aiyl aimaks* in the oblast with basic characteristics, such as population size, ethnic composition, and previous and ongoing interventions implemented by MSDSP. UN OCHA provides a repository of valuable information on the list of appeal projects by cluster, location of implementation, name of the implementing agency, budget and the number of target beneficiaries per district.

The impact evaluation team finds that MSDSP has ongoing interventions in several *aiyl aimaks* in Osh oblast. These *aiyl aimaks* were excluded from the programme. Further, according to UN OCHA, three districts – Osh City, Kara-Suu and Aravan – are targeted by ACTED for community restoration. Those districts were also excluded from the programme. After exclusion, the impact evaluation team ended up with 46 *aiyl aimaks,* and all were equally mono- and multi-ethnically representative.

Identification design: impact evaluation design employs cluster-RCTs, with two-arm treatment. The randomisation is conducted in two steps: first, mono- and multi-ethnic *aiyl aimaks* are randomly selected in comparison and intervention groups; second, households are selected randomly in these comparison and intervention *aiyl aimaks*. Two types of projects are implemented in both mono-ethnic and multi-ethnic communities: 'soft' projects (training and exchange of experiences, institutional strengthening) and 'hard' projects (infrastructural projects). This leads to four different types of intervention (we would like to be able to compare the effects of different treatments in mono- and multi-ethnic *aiyl aimaks*):

- Treatment 1: 'Hard' projects in mono-ethnic communities
- Treatment 2: 'Hard' projects in multi-ethnic communities
- Treatment 3: 'Soft' projects in mono-ethnic communities
- Treatment 4: 'Soft' projects in multi-ethnic communities

Each pilot *aiyl aimak* receives one treatment.

Outcomes and measurement: data collection for the impact evaluation is conducted at the *aiyl aimak* and household (beneficiary) levels. Both quantitative and qualitative methods are used. Baseline and follow-up data are collected in order to rigorously evaluate impact from the interventions.

The main indicator that the intervention is targeting is social cohesion at the individual level. The team uses a composite index of several sub-indicators that combine economic (access to land and credit), social (trust and status in society, etc.) and political factors (political exclusion of women and ethnic minorities). The surveys are designed to ask corresponding questions on economic, social and political factors in an ethical and respectful manner.

Power analysis: in order to choose an optimal sample size of the households to survey, the impact evaluation team conducts power analysis. Power analysis may

help answer the questions: How many households per community to survey? With this sample size, what minimum effect are we able to capture?

The number of clusters (i.e. *aiyl aimaks)* is 30 (15 for intervention, 15 for comparison). Assume that the impact evaluation budget may also cover no more than a total of 2,000 households.

For the power analysis, the team used an existing dataset in Kyrgyzstan, Life in Kyrgyzstan survey (2012), in order to choose an outcome variable. The team decides to proxy the social cohesion index with *trust in local institutions,* as it is the closest to the outcome of interest variable in the existing survey. *Trust* index varies from 1 – no trust at all, to 4 – absolute trust. Derived from the Life in Kyrgyzstan survey: mean value of *trust*: 2.71; standard deviation: 0.88.

The following additional assumptions are made:

Intra-class correlation: 0.05

Significance level: 0.10

Power: 80%

Given the constraints of budget and programme intervention, the following allocation is determined for treatment and comparison households:

- Treatment 1: 300 households
- Treatment 2: 300 households
- Treatment 3: 300 households
- Treatment 4: 300 households
- Comparison: 800 households

Table 3: Power analysis for clustered RCT design, case study 1b

Alternative	Minimum Detectable Effect	Percentage of increase
1	0.26	9.6%
2	0.30	11%
3	0.34	13%
4	0.37	14%

In Table 3 we show the MDEs for the number of treatments considered, one to four. Row four of the table illustrates that with four types of interventions and the given sample size, we are able to pick up the effect as small as 0.37 points increase in the level of trust, equivalent to a 14 per cent increase in the average value of trust.

Case study 2: Unanticipated emergencies

This case study builds on the work by Buttenheim (2010) and illustrates how impact evaluations of humanitarian assistance may be conducted after unanticipated disasters. It uses the 2005 Pakistan earthquake as an example.

Background: An earthquake, that was 7.9 on Richter scale, hit Pakistan in October 2005 and killed approximately 73,000 people. The affected area constituted 30,000 sq. km. of Azad Jammu Kashmir (four districts) and North West Frontier Province (five districts).[61]The earthquake destroyed or damaged 570,000 homes, leaving 2.8 million people without shelter. More than 1 million people lost their jobs and thousands of women lost their husbands, who had provided the family income. As of 2006, the relief and estimated total needs for long-term recovery overall amounted to US$1,092 million and US$5.2 billion, respectively (The World Bank).

Intervention: shortly after the disaster, the World Bank made available US$85 million for livelihood support, and the programme was launched in April 2006 in all nine affected districts. The programme gave priority to the most vulnerable groups, including female-headed households, children and orphans, and the poor. Eligible recipients were set to receive a monthly cash grant of Rs. 3,000 (about US$50) for six months (The World Bank).[62] Buttenheim (2010) describes an impact evaluation study of the Livelihood Support Cash Grant conducted after the earthquake in Pakistan. The World Bank, jointly with South Asia region study group,[63] evaluated the impact from the programme to measure its effects on the early recovery (health, assets and education) of the affected population.

Outcomes: the research team decided to evaluate outcomes related to health, assets and education. For simplicity and for this hypothetical case study, let us focus here on education outcomes. We may want to evaluate the effects of the cash grant on short-term and long-term outcomes. An example of short-term outcomes in education could be literacy rates of women or children, and knowledge in general. An example of long-term outcomes could be changes in attitudes and behaviour.

The assumed theory of change of this intervention is that in female-headed families with children, a single mother will choose to allocate the cash grant optimally for her and her household. It is likely, therefore, that she chooses to invest in her own and her children's education, to be able to support her family in the future.

Units of observation: the units of analysis for the impact evaluation are likely to be dictated by the programme design and the groups or subjects that the programme is targeting. The main units of analysis that the World Bank and South Asia region study use are households (to measure assets), individuals within a household (to

[61] Source: Earthquake reconstruction and rehabilitation authority, The World Bank.

[62]http://web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,contentMDK:21082733~pagePK:34370~piPK:34424~theSitePK:4607,00.html

[63] This study was implemented in partnership with Lahore University of Management Sciences (Pakistan) and Pomona College (USA) (Buttenheim 2010).

measure changes in education and health outcomes), and school facilities (to measure education-related outcomes).

Evaluation methodology: to estimate the impact of the Livelihood Support Cash Grant on health, assets and education of the affected population, the researchers used regression discontinuity design (RDD). The benefit of this approach is that baseline data is not required. However, the data on the indicator that determines the cut-off point for participation is required. In an RDD, the households slightly above and slightly below the cut-off point are compared, assuming that they are similar in most characteristics. One of the criteria used to determine eligible households for this grant was the number of dependents. Therefore, households with five or more dependents were given the grant, and those with fewer dependents were not. In the RDD, researchers took advantage of this cut-off and compared the households with four dependents (that did not get the cash grant) with those that had five dependents (that did get the cash grant).

Data: the evaluation sample consisted of 128 randomly chosen villages, drawn from the 1998 population census list of villages in four earthquake-affected regions (Buttenheim 2010). The World Bank undertook two waves of household and facility surveys. They included the first wave of 28,000 households in sampled villages in spring 2009 and 2,500 randomly selected households in autumn 2009. Household surveys included data on variables such as employment, consumption, nutrition, education of children, mental health, and asset recovery. School facility surveys included information on enrolment, and child test outcomes.

The main findings of the World Bank impact evaluation of the Livelihood Support Cash Grant are not yet published, and collected survey data are not publicly available (neither is the reasoning behind the choice of the sample size). Therefore, below, we discuss various other data sources that could be useful for impact evaluations, conducted in the aftermath of the Pakistan earthquake, and we use one of them in particular to demonstrate the possible sample sizes and possibilities for an impact evaluation.

More data: the Pakistan earthquake led to the successful experiment of the international and local research community to create a data management system that contained useful information for humanitarian actors *Risepak*.[64,65] The database was an "online portal, established to collect, collate, and display information on damage, access, and relief at the village level so as to facilitate earthquake relief coordination", and was created within 10 days following the earthquake.[66]

The Pakistan pre-disaster *census of 1998* was published on this database. The census contained data on village-level characteristics such as population, roads, availability of electricity and water, etc. After the earthquake, village-level data was updated with the distance from epicentre, and was constantly updated with village-level data on damage and needs assessment, assistance indicators, organisations

[64] To our knowledge, the Risepak database is no longer active online.
[65] Amin 2005
[66] Amin 2005

that provided assistance, relief aid distributed, etc. The Risepak researchers and volunteers also made field visits to the villages and sent updates to the database. Within two months of the disaster, Risepak contained information on 950 villages. In addition, the teams surveyed 3,840 households in 18 villages, which was "the largest independent survey of households since the earthquake".[67]

Risepak provides an illustrative example of how timely and effectively data collection activities may be implemented in post-disaster settings. It also informed impact evaluation teams about who implemented what assistance, when and where. Buttenheim (2010) mentions several small-scale studies that used the Risepak data for impact evaluations.

Another example of pre-existing data that researchers could use in these circumstances is the *Pakistan Social and Living Standards Measurement Survey* (PSLM). The project was initiated in 2004 by the Government of Pakistan (Bureau of Statistics) and will continue until 2015. As stated on the Government of Pakistan website, "an important objective of the PSLM Survey is to try to establish the distributional impact of development programmes; whether the poor have benefited from the programme or whether increased government expenditure on the social sectors has been captured by the better off". The survey is conducted at two levels: the district, and the provincial. It covers 80,000 households at the district level and 17,000 households at the provincial level. This survey was used to evaluate the impact of ERRA by monitoring and evaluation. If extended beyond 2015, this survey could be very useful for the impact evaluations of humanitarian assistance in the case of future hazards. Another useful survey is the *Demographic and Health Survey (DHS)*, which was conducted in Pakistan in 2006–07 and again in 2012–13.

Power analysis: to determine the statistical power for the regression discontinuity evaluation design, we use *DHS 2006-07,* as this survey is publicly available and is closer in time to the earthquake than, for example, the most recent survey of 2012–13. The statistical power of the regression discontinuity design (RDD) is lower than that of the RCTs.[68] Here we use a 'rule of thumb' for RD design without clustering: it states that required sample size for an RD design without clustering is 2.75 larger than that for a randomised control trial for the same power.[69]

In this power analysis, we estimate the effect of the Livelihood Support Cash Grant on the education of women, and we use a proxy variable such as the literacy rate for women from the DHS survey. From the DHS survey, we select women (with children) who are the heads of the household. There are 510 women in the sample with such characteristics. In the dataset, literacy variable is equal to 1 for a woman who is able to read at least a part of a sentence, and 0 for a woman who cannot read at all. Mean value of the variable is equal to 0.30 (i.e. 30 per cent of women

[67] Amin 2005

[68] See Bloom (2012) "Modern Regression Discontinuity Analysis" Journal of Research on Educational Effectiveness, 5:1, pp.43-82.

[69] Here we assume that the indicator variable is normally distributed (Bloom 2012). Optimal sample size calculations in clustered RD designs are more complex and not presented here for simplicity (Schochet 2008).

are able to read at least a part of a sentence), and standard deviation is equal to 0.45. So the assumptions about the variables required for the optimal sample size are as follows:

Statistical power: 0.80

Significance level: 0.05

R-squared (from the regression of literacy rates on the cut-off variable and covariates): 0.2

For each minimum detectable effect size (MDES)[70], we calculate an optimal sample size for the RCT and then use a 'rule of thumb' to detect an optimal sample size for the RD design. We assume that MDES for the outcome are 5, 10 and 15 per cent. Table 4 provides the corresponding optimal sample size numbers.

Table 4: Power analysis for RD design

Minimum Detectable Effect Size (MDES) (in percentage points)	Optimal sample size for RCT	Optimal sample size for RD design
5%	2,050	5,638
10%	525	1,444
15%	245	674

Notes: Optimal sample size for RCT is calculated using Optimal Design Software.

According to the power analysis, although the RD design is less powerful than RCT, with the RD design we are able to capture the effect of as little as a 15 per cent increase in literacy rates, with an optimal sample size of 674 women. The split between the treatment and control groups of women is optimal at 0.5, suggesting that the sample of a treatment group is only 337 women.

Case study 3: A complex emergency involving flooding and conflict

Floods can entail irreversible losses to human health and nutrition, as well as damage to crops, and household and public assets. In recent years, the world has witnessed instances of severe flooding that have led to massive losses of lives and property, such as across Pakistan in 2010, in parts of Thailand in 2011 and 2013, and in the Indian state of Uttarakhand in 2013. The onset of flooding can be sudden, for instance, when cloudburst caused flooding in the Ladakh plateau (India) in 2010, or more gradual as was the case in many downstream areas of the River Indus in Pakistan in 2010. Flood-preparedness and response time, and therefore flood damage, can thus vary depending on the cause of flooding.

[70] This is the standardised smallest treatment effect that a research design has an acceptable chance of detecting, if it exists.

Background: Post-flood aid comprises largely two aspects: relief in the immediate aftermath of the floods, and reconstruction, which refers to longer-term measures to enable households to rebuild their lives. In specific instances, governments and donor agencies may choose from an array of options to provide relief and reconstruction. While relief could include evacuation services, and emergency supplies of food/cash, longer-term reconstruction can include cash transfers, assistance with rebuilding assets and/or renegotiating loans. In the case of the 2010 floods, the Government of Pakistan rolled out a two-phase cash transfer for identified beneficiaries; the first tranche of the Citizens Damage Compensation Programme comprised a one-time unconditional cash grant of Rs. 20,000 to each household in a flood-affected village; the second tranche provided identified beneficiaries with a much larger cash grant (between Rs. 60,000 and 80,000) aimed at enabling reconstruction.

Intervention: here we illustrate a hypothetical sampling strategy for evaluating a cash-based flood relief programme. For this illustration, we propose to employ a randomised controlled trial-based field experiment for evaluation. In a post-flood scenario, such a set-up is most likely when administrational capacity constraints imply that all flood-affected areas cannot be covered at the same time. This would necessitate a phased rollout, which can then be used to carefully design a rigorous evaluation. Essentially, we propose to randomise the selection of villages for the cash-based flood relief in the first phase (rather than adopting other approaches such as first covering villages that are closer to administrative headquarters, for instance). This will ensure that the villages included in phases 1 and 2 (or in subsequent phases) are different only with regards to the timing of the rollout of the cash relief programme. Phase 2 villages therefore serve as comparisons for phase 1 villages. We assume that the cash transfer is provided to all households in a flood-affected village, defined on the basis of empirical measures of flood damage (as was the case in Pakistan), or flood exposure measures such as surplus rainfall, inundation, etc. In other words, the treatment is at the village, rather than household, level.

Outcomes: a flood relief programme can seek to protect several dimensions of human development that are vulnerable to flood exposure. These may include child malnutrition, household food consumption and household dietary diversity; and can serve as meaningful outcome variables for the evaluation exercise.

Power analysis and sample size: in order to detect any effects of the cash transfer on these outcome variables, we need to ensure that the sample has sufficient power. Table 5 illustrates calculations of optimal sample sizes for alternate specifications of effects, cluster sizes and levels of power. Other assumptions are:

Level of significance: 0.05

Intra-class correlation assumed (ratio of variability *within* clusters to total variability): 0.15

Power: 0.8 or 0.9

Table 5: Power calculations for sample selection

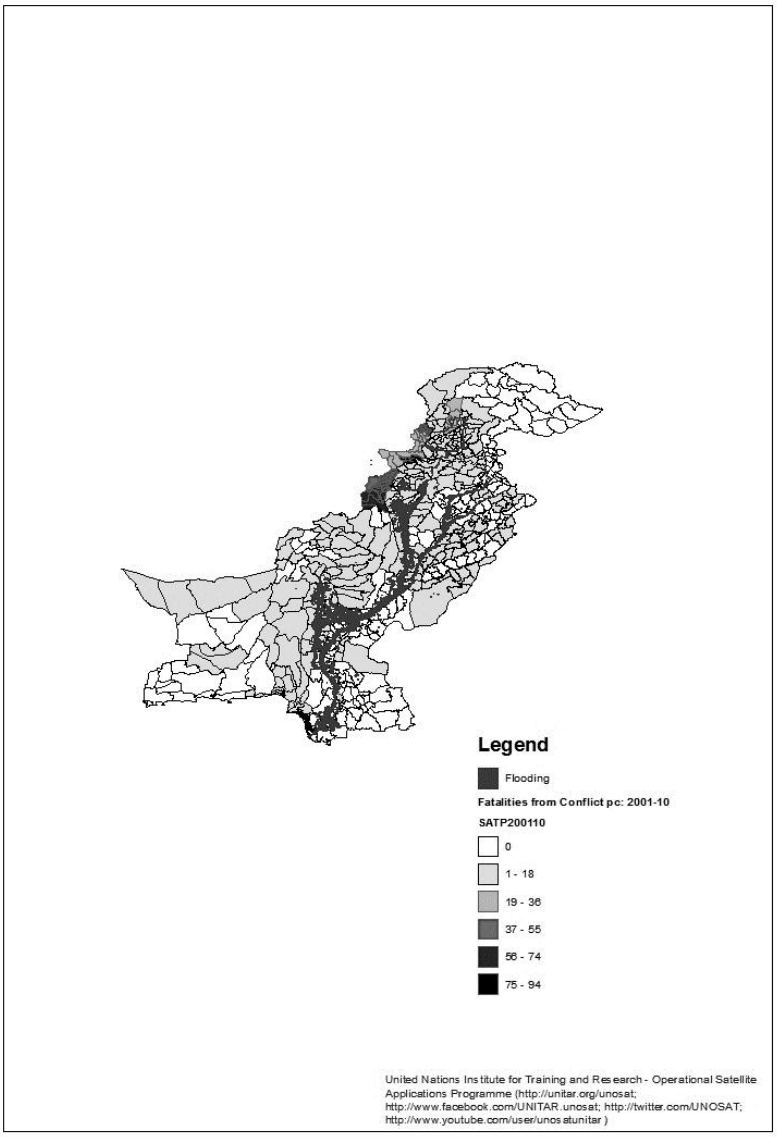
Minimum effect size (% points)	Number of households per cluster (or village)	No. of clusters (villages)		Total number of sampled households	
		Power = 0.9	Power = 0.8	Power = 0.9	Power = 0.8
10	20	806	609	16,120	12,180
20	20	208	154	4,160	3,080
30	20	92	70	1,840	1,400
40	20	54	41	1,080	820
10	30	745	560	22,350	16,800
20	30	190	142	5,700	4,260
30	30	86	66	2,580	1,980
40	30	49	39	1,470	1,170

Notes: calculated using OptimalDesign Software.

In this case, therefore, to detect changes in malnutrition, household food diversity or nutrition, and a 30 percentage point increase in these variables, the evaluation will require 66 villages (power =0.8) and will need to sample 1,980 households within these (30 households per village) to be able to measure the change with a power of 0.8. If greater power is required, for the same percentage point change to be detected, 86 villages will need to be sampled.

In the case of flooding in Pakistan in 2010 it is noteworthy that several areas affected by flooding were also affected by high levels of armed conflict over the past decade, making it a complex emergency with the overlap of multiple sources of vulnerability. The overlap of conflict intensity and flooding exposure is depicted in Figure 10.

Figure 10. Overlap of Conflict Incidence (at District level – 2001 - 2010) and Flooding in 2010 in Pakistan



Source: Author's calculations using South Asia Terrorism Portal and UNITAR Data

The expected programme impact in conflict-affected areas can be different due to a number of factors pertaining to both programme implementation (where conflict affects implementation capacity, security, infrastructure for delivery), and households' requirements and priorities (which may be markedly different from peaceful areas). In order to evaluate programme impact in such a complex emergency context, comparisons must not only be across treatment and control groups, but additionally between these groups, in and out of conflict-affected areas.

In our case, if we were to additionally require impact estimates over sub-populations of conflict-affected and peaceful areas[71] conflict-affected areas. Effectively, this calls

[71] We take these two categories for simplicity of illustration here; more complex situations may require a wider range of conflict settings— depending on the level of violence as Low/Medium/High, or according to the motivation of the violence—including international, local militant or insurgent conflict.

for doubling [72] the sample size, as the sample calculated above needs to be representative and suitably powered not only at the aggregate level, but also over the sub-groups of conflict-affected and peaceful areas.

Undertaking such a heterogeneity analysis also requires reliable and measurable data on the incidence of violence across the sample universe. This is usually available through international and local conflict monitoring agencies that maintain geo-coded and time-marked records of political violence. In the case of Pakistan, leading resources include the South Asia Terrorism Portal, the Pakistan Institute of Peace Studies, the Uppsala Conflict Data Programme and the BFRS Dataset on Political Violence. Care should be taken to analyse a longer time period to identify exposure to violence as greater exposure to violence over time, rather than a recent but one-time incident of violence, is more likely to reflect the inherently different character of a conflict-affected area, which is the aim of our heterogeneity analysis.

[72] Or multiplying by n where n is the number of sub-populations for which we seek separate estimates of programme impact.

Box 4: Challenges of conducting evaluation in conflict-affected settings

Conflict-affected areas are often marked by a distinct set of political actors and processes that may be very different from those in more peaceful areas. The political economy of aid delivery and targeting, including disaster aid, can be particularly complex in conflict-affected areas, with direct implications for data collection for evaluation.

1. The most obvious and direct challenge is the security risk associated with entering conflict-affected areas. Data collectors may be seen with suspicion, by communities and armed groups alike, who may fear that covert intelligence-gathering activities may be taking place under the guise of a survey. Armed groups can threaten violence and in some cases also carry out abduction or physical assault.
2. Aid has the potential to become, or be seen as, an instrument through which the state/donors and NGOs can potentially win hearts and minds of people living in conflict-affected areas. For this reason, aid may be opposed by local armed groups. Consequently, local authorities may be diffident in providing a valuable modicum of support, guidance and logistical help that is otherwise extended to field surveyors when they enter a new community.
3. Aid is contentious in conflict-affected areas and is often opposed/resented by locally powerful armed groups. This can affect respondents' attitudes. They may not cooperate with enumerators, or withhold/misrepresent information on aid receipts, their use, and usefulness for fear of violent backlash.
4. The higher costs of surveys can be an additional factor in conflict-affected areas. This is because such areas may often be more remote/inaccessible, and also because of the higher insurance costs for enumerators entering very high-conflict areas (linked, again, to a higher security risk).
5. During a follow-up round of an evaluation survey, i.e. after the rollout of the aid, any control group communities that lie in a conflict-affected area may feel that they have systematically been denied aid as punishment for supporting rebel groups. Alternatively, armed groups may portray such an image to garner sympathy and fuel resentment against the state. This can further increase the security threat for enumerators and the non-cooperation of respondents.

While the extent to which these challenges apply will vary across settings and over time, these are important points to consider when evaluating interventions in complex emergencies. These challenges can often be overcome if pre-empted and addressed carefully. For instance, an open/covert dialogue between state and armed groups can allow the peaceful rollout and evaluation of disaster aid, as was done between the Sri Lankan government and the LTTE in the aftermath of the 2004 Indian Ocean Tsunami.

To the extent possible, enumerators for conflict-affected areas should be recruited locally. Local recruits are more mindful of the conflict dynamics in the area, are less conspicuous among the communities and can employ local knowledge and networks for making on-the-spot security assessments.

Enumerators must be provided support by engaging in dialogue with individuals and groups that may have a channel of communication with armed groups either at a high level, or locally. This will ensure that the survey teams have a better understanding of the security risk they face, to make informed decisions. In other cases, they may require armed escorts, provided either by the state or privately, to ensure their safety.

Enumerator teams should constantly be in touch with the survey organization, who should monitor any rapidly developing security threats and advise teams. Survey teams should have an evacuation/exit plan, with arrangements of support from local authorities, the police, or any other helpful contacts.

Case study 4: A protracted emergency— internally displaced peoples in DRC

Background: this example presents an emergency that has been ongoing for a long time. In November 2013, European Community Humanitarian Office (ECHO) announced that it would increase its funding to UNICEF to increase assistance to families and children who are affected by armed conflict, natural disaster and cholera in the eastern Democratic Republic of Congo. This programme is called the Rapid Response Mechanism for Population Movements (RRMP). Support is provided in the form of emergency support in health, essential household items, emergency water and sanitation and emergency education. Actions target children and their families in North Kivu, South Kivu, Province Orientale and Katanga. Children and families have suffered from conflict in their home villages. Due to armed conflict, an estimated 2.7 million people, more than half of them children, are internally displaced. More than 96 per cent of the displaced are living with host families. Instability in Katanga led to the displacement of 375,000 people. [73]

The objective of RRMP is not to target *all* internally displaced people – it does not have the requisite resources. But RRMP prioritises its interventions in the following way:

- Areas that have the most displaced people with the most acute vulnerabilities
- Areas characterised by complexity of (physical) access
- Areas characterised by lack of other actors

These are all included in a characterisation matrix.

The overall objective of RRMP is to maintain rapid deployment capability (which means it does not intervene everywhere) and to intervene in areas where displaced and returning populations benefit most. For this purpose, it constructs 'intervention thresholds'. The purpose of these thresholds is to formally trigger RRMP interventions, to ensure that the proposed interventions are within the mandate of RRMP and also provide information to other partners in the cluster. For each sector, therefore, there is a list of activities and indicators over which data is collected. These data are sorted according to whether an area is a displacement area, number of long-term returnees, number of temporary returnees, and a mixture of these. Interventions are divided into non-food items (NFIs), education, and water and sanitation interventions.

It is important to note that the analysis of the threshold and the decision to trigger an RRMP response or not is taken during RRMP committees, using quantitative data collected on indicators and qualitative and context analyses. RRMP committees consist of OCHA, UNICEF, cluster leads and other partners. As part of this exercise,

[73] Please see http://news.sciencemag.org/2010/01/how-many-have-died-due-congos-fighting-scientists-battle-over-how-estimate-war-related for further inspiration, for questions to ask (humanitarian action is defined in the introduction as life-saving activities) and for methodological issues.

RRMP undertakes a vulnerability analysis. This includes the vulnerabilities of communities and the vulnerability of individuals.

The following are the activities that were undertaken January–October 2013:

- 450,699 people were provided essential household non-food items (NFI) and shelter materials, 70 per cent of them through voucher fairs;
- 295,144 people were given access to water, sanitation and hygiene in emergency;
- 99,421 children aged 5–11 years old had access to quality education and recreational and psychosocial activities in a protective environment;
- 40,481 people received free emergency healthcare through RRMP operated mobile clinics and supported health facilities; and
- RRMP gave critical support to the emergency vaccination of 284,143 children against measles.

In addition to these activities, NFI fairs are held. The Norwegian Refugee Council, for example, provides funds for NFI fairs (these are described below).

Outcomes: possible questions that can be asked in this impact evaluation:

- Are voucher fairs more effective compared to direct transfers?
- To what extent do people who had access to quality education actually use it (enrolment is not sufficient)?
- What is the best way to deliver and ensure the use of clean water in camps?
- What is the best way to ensure that people uptake hand washing and sanitary methods for defecation in Internally Displaced Person (IDP) camps?
- How can providers of psychosocial activities be trained to most effectively provide counselling to IDPs?

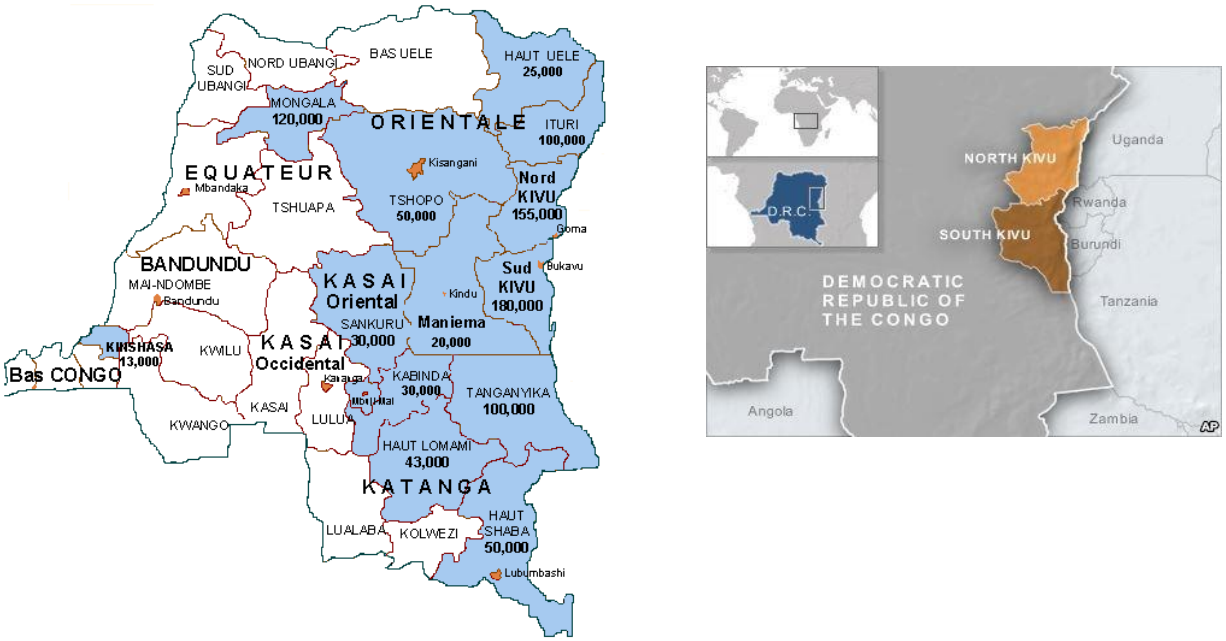
For the purposes of this hypothetical case, our primary question is: what works best for internally displaced people? Direct food transfers, cash and food vouchers can all contribute to reducing malnutrition, especially in the areas where there are high rates of acute malnutrition and difficulties in distribution and access. Do displaced households and returnee households benefit in similar ways from food transfers, cash and food vouchers?

Data: let us focus on North Kivu. International Organization for Migration reports that, as of 2013, there were more than 900,000 people living in 31 North Kivu IDP camps in 2012.[74,75] On average, there are 30,000 people living in each camp. UNICEF, RRMP and other agencies collect data to understand the thresholds, but also to know the births, deaths, new arrivals and departures for camp-based IDP populations. The Data Centre for IDPs in North Kivu collects data on the following[76]:

[74]http://reliefweb.int/sites/reliefweb.int/files/resources/DRC%20Factsheet%20Population%20Movement%20_english_2%20eme%20trimestre%202013.pdf
[75]http://www.irinnews.org/report/95836/briefing-crisis-in-north-kivu
[76]http://www.fmreview.org/DRCongo/church.htm

- undertaking individual registration of camp-based IDP populations, including new arrivals, departures, births, deaths, etc.;
- maintaining an up to date and real-time database that allows for population tracking and the production of disaggregated data on IDP populations;
- managing population movements from, to and between IDP camps by ensuring individual documentation, such as Voluntary Return Attestations, etc.;
- producing accurate beneficiary lists for assistance purposes, taking into account family size, special needs, and vulnerability criteria such as defined by the humanitarian community in DRC;
- helping develop a strong humanitarian data analysis capacity within the framework of the Congolese government's stabilisation plan for eastern DRC;
- ensuring individual registration of Congolese refugee returnees in order to facilitate verification in the countries of asylum and to assist UNHCR North Kivu in planning for protection and assistance activities; and
- maintaining a database for protection and returnee monitoring reports.[77]

Figure 10: Democratic Republic of Congo, North and South Kivu camps

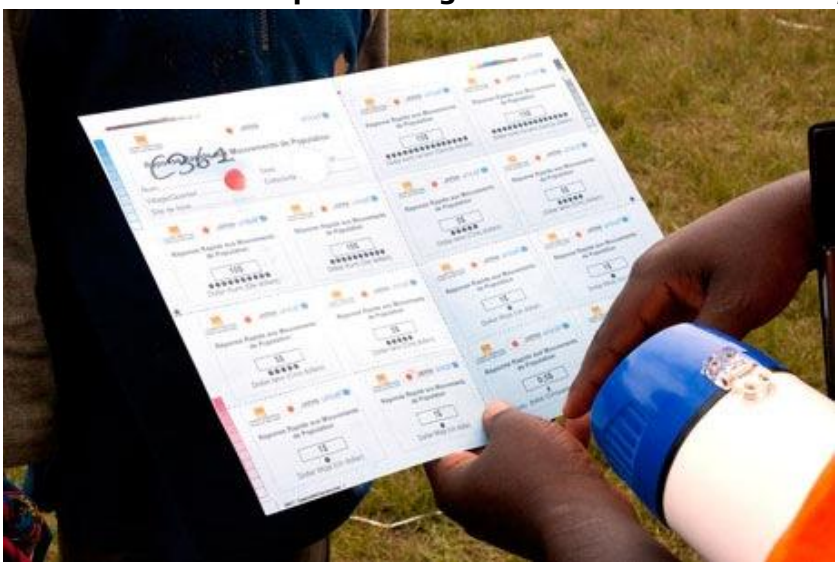


[77]See more at: http://www.fmreview.org/DRCongo/church.htm#sthash.gyEtXjqK.dpuf

Picture 1: An IDP camp in Goma, North Kivu

Photo credits: MSF

Picture 2: Food coupons being distributed in North Kivu, DRC



Picture 3: Bulengo refugees' camp, west of Goma, North Kivu, DRC, 2008

Photo credit: B Smets

Picture 4: A view of the IDP camp on the outskirts of Nyanzale in North Kivu

Photo credit: MSF

The intervention, the target population: there is a general consensus that although healthcare, education and livelihoods are provided by UNICEF in this case, it is not clear what is effective and what is not. The cluster system allows for coordination, and one agency works to provide one set of interventions to people living in IDP camps. The cluster system can also be useful in facilitating a rapid impact evaluation because it ensures that different agencies can coordinate their programmes.

It is believed that food coupons inject cash into the local economy and require fewer logistical expenses compared to injections of cash for donors because vendors are responsible for transporting goods. Coupons can vary in value. Many agencies discuss the effectiveness of coupons against direct cash transfers, and also question their effectiveness compared with cash transfers.[78]

Identification design: clearly an important question in this context is to assess whether cash transfers are more effective than food coupons. There are two possible designs here. The first is clustered randomisation, where all eligible camps are divided into two groups. Camps can randomly belong to the first or the second group. In the first group of camps, all individuals get food coupons. In the second group of camps, households receive cash. The second possible design is randomising households or individuals within the *same* camp into two groups, with one group getting food coupons and the other receiving cash. Randomisation at the individual level in humanitarian assistance has its advantages and disadvantages. Box 5 below discusses some pros and cons of clustered randomisation over simple randomisation where the treatment group receives cash.

[78] For example, the NRC provided each family with non-food item coupons worth US$75 for urgently needed household items. These are especially useful for returning families, who find that in their homes all their belongings are missing. Items that can be bought with these coupons range from tin panels for roofs, to feminine hygiene kits to insecticide treated mosquito nets.

Box 5: Comparisons of cluster randomisation and individual level randomisation

Pros Cluster level randomisation is easier to implement because the same intervention needs to be implemented in the entire camp.

Deals easily with individual level spill-overs or exchanges across households within the same camp compared to individual level randomisation, where it is difficult to determine if people indeed used the coupon or traded it in.

Cons Requires many more clusters or camps compared to individual randomisation to account for within-camp correlated behaviour and outcomes.

Therefore, it is much more costly to implement as an evaluation design because many camps are required. (Managing access, quality and fidelity to the design across camps imposes additional burdens, over and above the cost of reaching and accessing them.)

Power calculations and sample size: given Box 5 above, we use cluster randomisation to test the effectiveness of cash and compare it with food coupons.[79] In this case, we discuss distributing coupons versus direct food transfers and examine the impact on only one outcome – nutrition – although clearly this can be extended to several outcomes across sectors. We determine statistical power with the corresponding sample size for different effect sizes.

Assumptions

An important indicator of nutrition is haemoglobin levels of children aged from 6–59 months in DRC (data from DHS 2007)

Mean: 102 g/l; Standard deviation: 17.4

Number of clusters: 15–66

Significance level: 0.05

Intra-cluster correlation: 0.15

Power: 80%

[79] Note that we could use this design to also compare cash with non-food item coupons. The difference then would be that we'd have to change the outcome variable, and the power analysis and sample size calculations would then change.

Table 6: Power analysis for case study 4 – internally displaced peoples in DRC

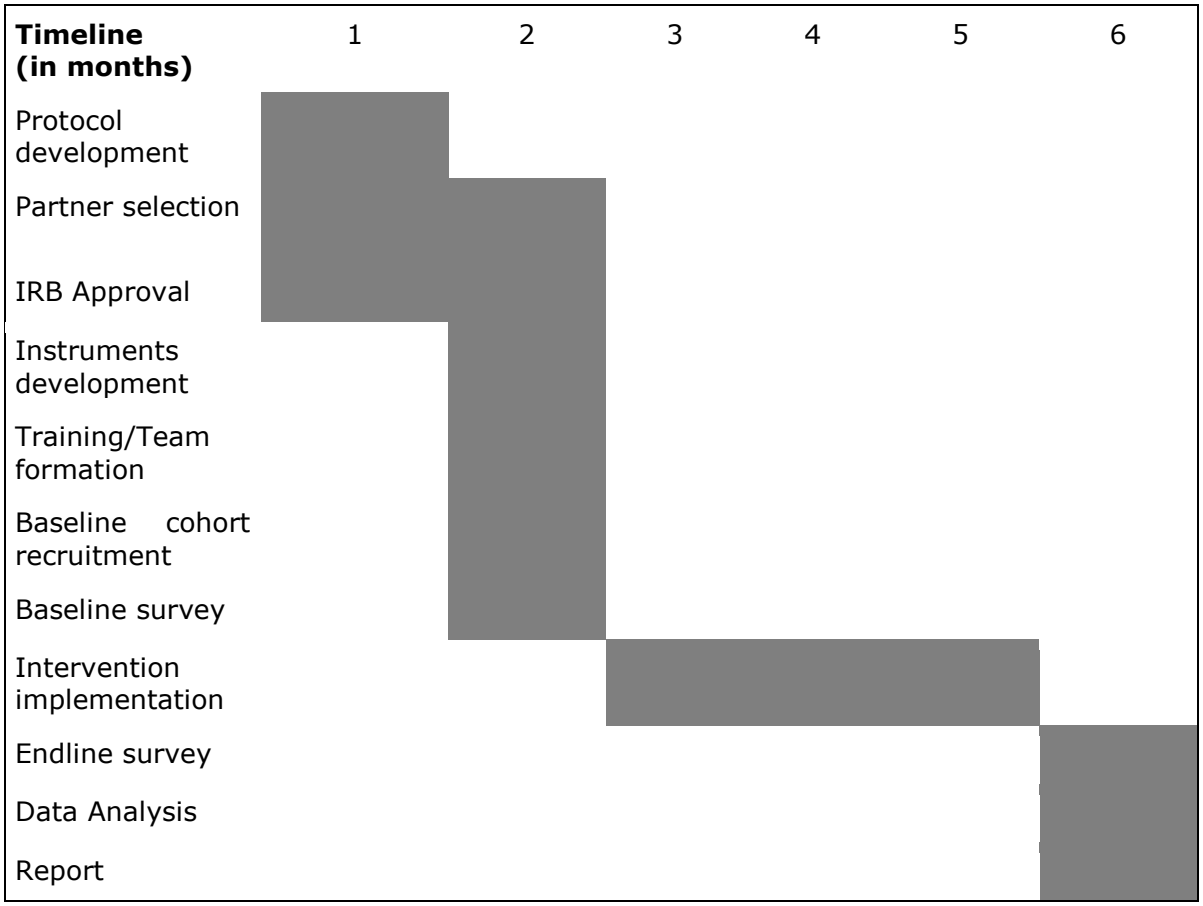
Minimum effect size (percentage point)	Increase in haemoglobin levels in g/l	Number of subjects per cluster (n)	Number of clusters (J)	Total sample size (N=n*J)	Power
5%	5.1 g/l	30	66	1,980	80%
5%	5.1 g/l	60	61	3,660	80%
5%	5.1 g/l	100	59	5,900	80%
10%	10.2 g/l	30	17	510	80%
10%	10.2 g/l	60	16	960	80%
10%	10.2 g/l	100	15	1,500	80%

Note: calculated using Stata module.

Table 6 shows that a sample size of 510 in 17 clusters will suffice to detect a relatively small effect size of a 10 per cent increase in the outcome (haemoglobin) variable due to food coupons. The comparison group is assumed to get cash transfers.

Timeline: the possible total timeline for this study, assuming an intervention of three months, will be eight months, assuming that the effects of better nutrition can display themselves in haemoglobin levels after four months.

Figure 11: Timeline for a rapid impact evaluation in DRC

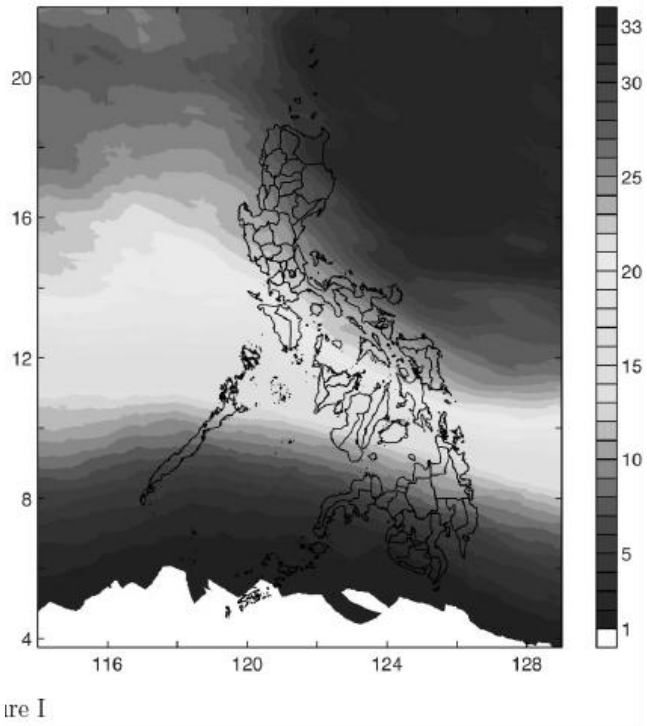


Case study 5: Using impact evaluations to estimate the effect of assistance after typhoons in the Philippines

Background: Typhoon Haiyan hit the Philippines on 8 November 2013, with unusual and brute force and with some of the highest wind speeds ever recorded. Typhoons which are tropical cyclones in the West Pacific affect the Philippines regularly (about 10 times a year), and the death toll and devastation after Haiyan was unprecedented: about four million people were left homeless and electricity and water supply infrastructure was extensively damaged.[80] Health services were severely disrupted.

What was achieved by the rapid humanitarian relief operation? And is it possible to estimate the short- and intermediate-term impacts of such a sizeable relief operation? What can be said about the impacts of smaller-scale, more specialised and targeted interventions that are part of such large-scale initiatives? This case study discusses these questions and the challenges associated with implementing mixed-methods, theory-based quasi-experimental impact evaluation designs in humanitarian emergency settings. Our discussion draws extensively on Hughes and Hsiang's empirical analysis of the impacts of typhoons in the Philippines;[81] Buttenheim and White's presentation[82];

Figure 12: Variation in typhoon exposure across the Philippines



ıre I

[80] Hughes and Hsiang 2012
[81] 2012
[82] 2009: Are Disasters Any Different: Challenges and Opportunities for Post-Disaster Impact Evaluations'

and selective insights from the World Bank's efforts to evaluate relief in the aftermath of the 2005 Pakistan earthquake.[83]

Hughes and Hsiang (2012) exploit the variation in typhoon exposure across the Philippines to estimate household welfare and other typhoon-induced losses. Their map, which is reproduced in Figure 11, highlights this dramatic variation. Our example differs from their study because, unlike them, we are not estimating the impact of the typhoon, but rather the impact of the emergency relief (a) in Haiyan's immediate aftermath, and (b) during the subsequent recovery period.

Hughes and Hsiang show that in contrast to the immediate health deterioration and fatalities resulting from direct exposure to strong winds and the subsequent flooding that were responsible for the large number of casualties in, for example, the 1991 Bangladesh cyclone, the most severe adverse effects on life and health from typhoon exposure in the Philippines occur after a significant time lag. To illustrate, the indirect and lagged infant deaths from typhoon exposure outnumber immediate deaths from direct exposure by about 15 to 1. Another startling finding is that this lagged impact is distinctly gendered, with girl infants systematically endangered by post-typhoon adjustment disinvestments in health and human capital within Filipino households. [84]

If needs determine relief efforts, areas hit hardest by a disaster are also more likely to receive relief.[85] This is a variant of the selection bias touched upon above but cannot be taken for granted: a vast literature shows that local and other politics may distort emergency relief allocations.[86]

Our approach follows Hughes and Hsiang (2012) and a recent and rapidly growing literature using geological, seismic or—as in this case—meteorological physical storm data and variation in typhoon exposure at the province[87] level to estimate disaster impacts. In a difference-in-difference design, Hughes and Hsiang (2012) estimate impacts on household physical assets, consumption, and income and health. These estimates of household losses are net of private and public transfers (where the latter would include emergency relief).

Intervention design: let us suppose that aid organisations allocate relief based on disaster intensity. According to Hughes and Hsiang (2012), spatially weighted maximum windspeed ('windspeed' from now) reliably measures typhoon intensity and predicts household losses accurately. This resembles release clauses for weather

[83] Above, we noted that emergency relief could simply crowd out assistance and relief provided through informal financial and other kinship-based support to affected households. As also noted, such informal insurance is much less effective in mitigating the adverse impacts of the covariate shocks typhoons are an example of.

[84] (Doocy *et al.* 2013)

[85] Note that 'hit hardest' may not be equivalent to the strongest typhoon exposure because of heterogeneity in the resilience of e.g. residential buildings to withstand typhoon level winds.

[86] (e.g. Sen 1971; Chang and Zilberman 2013)

[87] With data from 82 provinces.

insurance in agriculture where the levels of measured rainfall trigger insurance payments. Schemes that link insurance payments to windspeed exposure are presently being discussed for the Philippines.[88]

Vulnerability to high windspeed is likely to be correlated with socio-economic status: the housing structures of the poor may be less robust; their water and sanitation and general health infrastructure more patchy, and so forth. For a given windspeed, therefore, deprived areas are expected to be more severely afflicted.

Following on from the above, one option is to locate a windspeed (or a damage) cut-off or discontinuity for distributing emergency relief that, for example, facilitates comparisons of a narrow band of households just below or just above the threshold that triggers action. Both in terms of the losses incurred from the typhoon and in socio-economic characteristics, these households can be expected to be very similar.[89] For now and in order to identify the impact of overall emergency relief after Haiyan, we need to make comparisons of 'similar' households from areas just above and just below the windspeed threshold.

Data: a first concern is that data on windspeed exposure are not available at the household level. Hughes and Hsiang (2012) rely on windspeed variation using province-level data (there are 82 provinces). Like Hughes and Hsiang (2012), we use the most recent DHS data to locate households from provinces just above and just below the threshold. We need a large enough number of observations and a balanced sample of such 'baseline' households for each of the two selected provinces. Such baseline balancing cannot be assumed and, for the identification design to work well, this needs to be verified using data to make actual comparisons.

Identification design: we use a regression discontinuity design to measure the impact of humanitarian assistance. Let us assume that this intervention takes the form of distribution of iron tablets to improve the anaemia count for people affected by the typhoon.

Outcome: iron tablets or syrup intake for at least 90 days during the last pregnancy.

Pregnant women should take iron tablets or syrup for at least 90 days during pregnancy to prevent anaemia and other complications. Only 34% of women took iron tablets or syrup for 90 days or more during their last pregnancy according to DHS 2008 for Philippines.

For the power analysis, we use a 'rule of thumb' and assume the optimal sample size for RDD is 2.75 larger than that for simple RCT (see Pakistan case).
Assumptions for the power analysis:

[88] http://reliefweb.int/report/philippines/typhoon-haiyan-losses-trigger-major-new-proposal-catastrophe-insurance)
[89] e.g. Khandker *et al*. 2007

Mean proportion of pregnant women who took iron tablets for 90 days or more during their last pregnancy (from DHS 2008): 0.34

Standard deviation: 0.47

Power: 0.80

Significance level: 0.05

R-squared (from the regression of iron intake on the cut off variable and other covariates: 0.2

Table 7: Sample sizes for Regression Discontinuity Design, case study 5

Minimum Detectable Effect Size (MDES), in percentages	Optimal sample size for RCT	Optimal sample size for RD design
5%	2,252	6,193
10%	563	1,548
15%	250	688

Notes: calculations by hand. See Bloom (1995, 2012) for the formulas.

This case study shows that in order to detect a 15 percentage point change in the proportion of women taking iron tablets as a consequence of the intervention, i.e. a change from 34 per cent to 49 per cent, the evaluation team would have to survey nearly 700 women at baseline and endline, in order to be confident about the result.

The exercise above yields the minimum sample size required to evaluate a cluster-randomised post-flood cash transfer programme at an aggregate level. The sample calculations arrived at, however, are insufficient for comparing programme impact at lower levels (such as provinces/districts) or to detect any other underlying heterogeneity of programme impact. This is particularly relevant to the case of Pakistan where the overlap of violent conflict and floods results in a complex emergency situation (see map on p.46).

Case study 6: Using impact evaluations to estimate the effect of assistance in the recovery phase in the absence of *ex ante* planning

When randomising individuals into treatment and comparison groups is difficult or impossible, and when evaluation starts during implementation of the programme, it is possible to use propensity score matching methods (PSM). We provide an illustrative hypothetical example of the use of PSM in the evaluation after a disaster. This example also illustrates the use of mobile technology that may be used to provide reminders for anti-disease pills amongst the affected population. Many factors contribute to low adherence rates of taking anti-disease home treatment. However, the recent studies find that forgetting is the most common form of non-adherence (Costa *et al*. 2011). The use of reminders could help in overcoming this problem.

Intervention: part of the humanitarian assistance package involves sending SMS reminders to use anti-TB pills to a group of individuals with tuberculosis (TB). The incidence rates of TB may increase after a disaster, especially in countries with already high disease prevalence, as in Sub-Saharan Africa. Suppose that the implementing agency arrives to deliver anti-TB treatment (pills) after the disaster in South Sudan.[90] The implementing agency also budgets SMS reminders to take anti-TB pills on a frequent basis, but is unclear whether these work.

Impact evaluation: aims to estimate the effect of SMS reminders on the adherence rates of taking anti-TB pills.

If the selection of individuals who were going to receive SMS reminders was *not* random. In this case, the evaluation team decides to use PSM. The main advantage of using PSM is that it reduces the selection bias, which occurs when treatment individuals (those who receive the reminders) are systematically different from non-treatment individuals (those who do not receive the reminders).

PSM consists of four stages:

1. Estimating the probability of participation: the propensity score for each unit in the sample (which includes those who received the SMS reminders and those who didn't) is calculated using baseline data. Probabilities of being selected try to mimic the selection criteria that was used by the implementing agency, and create an estimated probability using probit or logit regression on an exhaustive set of observable characteristics. In our hypothetical impact evaluation, these characteristics may include the presence of TB, socio-demographic characteristics, and lack of access. Probability or propensity scores are calculated for the entire sample (including those who received the SMS-related intervention) from baseline data.
2. Nearest-neighbour matching: the beneficiaries are matched with non-beneficiaries with the closest propensity scores. Non-beneficiaries with scores dissimilar to those of the beneficiaries are dropped from the analysis – and vice versa. (This dropping of beneficiaries and non-beneficiaries is called 'lack of common support'.)
3. Balance check: this is done to establish whether matching in observed characteristics between beneficiaries and non-beneficiaries was successful.
4. Estimating the programme effect: this is done by averaging the differences in outcomes between each treatment unit and its neighbour in the control group.

[90] According to WHO (2012), prevalence of TB in South Sudan is 257 cases per 100,000 individuals of the population.

Calculating sample sizes for PSM cases

The optimal sample size for a PSM study is larger than that of an experimental study like randomisation, because subjects for treatment are matched with similar subjects for comparison; therefore, there should be a relatively large pool of subjects in the comparison group. In his blog, D. McKenzie argues[91] that a PSM sample of the control group could be 20-200 per cent larger than the treatment sample of an experimental study, provided that we know well the characteristics of the treatment group, and that we can sample accordingly. This is also important because the implicit assumption in PSM is that the observable characteristics (or attributes that can be observed and are measured) are also accounting for unobservable characteristics.

To evaluate the effect of messaging on the adherence rates of anti-TB treatment, we first estimate the optimal sample size of the treatment group, as in a pure experimental study (with no clustering). We then expand the sample sizes by the according rate. For this, we make certain assumptions on the outcome variable (e.g. adherence rate) and test statistics:

Average adherence rate for treatment to TB: 50%

Standard deviation: 0.30[92]

Power: 80%

Significance level: 0.05

Assume further that we want to detect a 2 per cent, 5 per cent and 10 per cent change in adherence rates. The results from the power calculations for the optimal sample size of the treatment group in pure experimental design are presented in column 2 of Table 8:

Table 8: Optimal sample sizes for PSM

MDES	Optimal sample size in pure RCT	Optimal sample size in PSM
2%	7064	8477
5%	1132	1358
10%	284	341

Note: sample sizes for RCT are calculated using Stata

[91] http://blogs.worldbank.org/impactevaluations/power-calculations-for-propensity-score-matching

[92] The assumptions on mean and standard deviation are hypothetical. In the literature, adherence rates of taking anti-TB, anti-HIV, or anti-malaria treatments ranges from 40 to 75% in the developing countries.

Table 8, column 3 shows the optimal sample sizes for the control group in PSM, which are blown up by 20 per cent, which is a lower estimate and a more optimistic scenario when we know well about the treatment group characteristics. It shows that for a 10 percentage point change in the average adherence rate (an increase from 50 to 60 per cent), the optimal sample size for the total sample is 341 individuals.

8. Conclusions

This paper assesses the challenges for impact evaluations of interventions in humanitarian emergencies. Given the complexity of humanitarian contexts, the need for speed, the lack of baseline data, the multitude of actors and requirements of coverage and capacity, and the significant ethical concerns about impact evaluations often expressed, it is usually assumed that theory-based impact evaluation methods cannot be used in such contexts. This explains the scarcity of high-quality studies. At the same time, the need for learning in the context of humanitarian assistance is enormous, with scarce resources barely meeting significant needs for assistance. So, delivering effective and efficient assistance is of significant interest to donors and recipients alike.

In this paper, we demonstrate that it is possible to conduct rigorous impact evaluations in humanitarian emergencies. With the help of six case studies and drawing on real-life examples from the small but growing academic literature, we show how impact evaluation methods can be used successfully and in an ethical manner to learn about providing humanitarian assistance effectively and efficiently. This is often achieved by adjusting research designs to programme realities that, in turn, mean that not all delivery can be done in one fell swoop. In fact, very often there are many decisions that field staff need to make that can be well-informed by impact evaluation methodologies.

We also show that impact evaluation methodologies may be used constructively, not only to understand impact, but to assess what design of programme might be best suited to a humanitarian context, and to help understand what method of delivery might be most appropriate to the time of response and context of the humanitarian disaster. There is, hence, a lot of scope to improve practice in the humanitarian sector as a result of learning based on impact evaluations.

We also use case studies to illustrate that the data requirements of impact evaluations are not as onerous as is often suspected. In many cases, researchers may draw upon pre-existing datasets that can help evaluations in providing proof of balance, as well as providing them with insights into context.

The paper draws the conclusion that while ethical concerns about impact evaluations are valid, they can be addressed, making impact evaluations feasible also from a ethical point of view. The do no harm principle can be usefully adjusted to emergency settings, especially when learning about how to deliver assistance in environments where there is little rigorous knowledge about what works best.

One reason why there is a dearth of hard knowledge is that humanitarian emergencies can occur under so many different circumstances, as argued at the start of this paper. Furthermore, responses to emergencies are also extremely heterogeneous. Taking account of the context—for example, through formative research—is, hence, particularly important. At the same time, theory-based impact evaluation can help to generalise lessons, because the analysis will uncover why something did or did not work. Addressing such causal issues contributes to building a general understanding of behaviour, which can help plan the next emergency response.

A key lesson from our report is that it pays to be prepared. Much information is being collected these days about the risks of various emergencies unfolding, be they sudden onset or slow onset emergencies. Hence, national actors and international donors can prepare on three fronts: (i) they can learn about where emergencies may unfold and where assistance may be required; (ii) they can plan ahead and be prepared to intervene for when an emergency unfolds (including strengthening local resilience *ex ante*); and (iii) they can prepare their impact evaluation designs in advance, drawing on the many insights into how to conduct successful impact evaluations offered in this paper and in the emerging literature on this topic.

Being prepared to conduct rigorous impact evaluations also includes preparing the national and local capacity, understanding and support for impact evaluation among donors. Impact evaluations can answer some questions but they do not answer all questions that donors pose. There may still be strong misconceptions about the inequities of randomised controlled studies for example, and fears about the costs and duration of impact evaluations. Some of these concerns are very valid, of course. Impact evaluations help to create knowledge that, ideally, is a public good. Impact evaluations are less useful for fast learning about how to improve an ongoing intervention. Yet given the dearth of rigorous causal evidence of what works and what does not work in the humanitarian sector, there is a high dividend to be earned from conducting more impact evaluations in emergency settings. We therefore expect there to be many more impact evaluations taking place in the humanitarian sector in the years ahead.

Appendix A : Table on impact evaluations of humanitarian relief

Table 9: Impact evaluations of humanitarian relief

No.	Study	Method & Counterfactual	Outcomes	Main findings
1	UNHCR & WFP (2012) "The contribution of food assistance to durable solutions in protracted refugee situations; its impact and role in Bangladesh: a mixed method impact evaluation" **Country:** Bangladesh **Category:** food assistance **Disaster:** conflict	**Method:** Quasi experimental **Counterfactual:** Natural control group **Rigour:** ToC, PA	(i) Food security; (ii) refugee movement; (iii) global acute malnutrition (GAM); (iv) economic activity and earnings; (v) copying strategy index; (vi) household dietary diversity score (HDDS); (vii) protection indicator	(i) Improved dietary diversity and reduced frequency of negative coping strategies; (ii) positive impact on economic activity; (iii) improved self-reliance and security
2	WFP & IFPRI (2012) "Impact evaluation of cash, food vouchers, and food transfers among Colombian refugees and poor Ecuadorians in Carchi and Sucumbíos" **Country:** Colombia **Category:** cash vs. food **Disaster:** conflict and poverty	**Method:** RCT **Counterfactual:** Random Assignment **Rigour:** -	(i) Food consumption and diversity; (ii) social capital (discrimination and participation in groups); (iii) anaemia; (iv) IPV	(i) Value of per capita food consumption increased by 13 per cent, per capita caloric intake increased by 10 per cent, HDDS improved by 5.1 per cent, dietary diversity index (DDI) by 14.4 per cent, and food consumption score (FCS) by 12.6 per cent; (ii) vouchers lead to the largest gains in dietary diversity and food leads to the largest increase in caloric intake; (iii) did not lead to a significant change in haemoglobin levels or anaemia classifications (negative effects on food group); (iv) discrimination decreased by 6 percentage points and participation in groups increased by 6 percentage points, decrease in IPV
3	Huybregts, L. *et al.* (2012) "The Effect of Adding Ready-to-Use Supplementary Food to a General Food Distribution on Child Nutritional Status and	**Method:** RCT **Counterfactual:** Random assignment with factorial model **Rigour:** PA, ToC,	(i) Anthropometric measures; (ii) and morbidity	(i) Reduction in cumulative incidence of wasting (incidence risk ratio: 0.86); (ii) lower gain in height-for-age (+0.03 Z-score/mo); (iii) higher haemoglobin concentration (+3.8 g/l), thereby reducing the odds

No.	Study	Method & Counterfactual	Outcomes	Main findings
	Morbidity: A Cluster-Randomized Controlled Trial" **Country:** Chad **Category:** food assistance **Disaster:** conflict	ToB, E		of anaemia (odds ratio: 0.52); (iv) lower risk of self-reported diarrhoea (229.3%) and fever episodes (222.5%).
4	Doocy, S. and G. Burnham (2006) "Point-of-use water treatment and diarrhea reduction in the emergency context: an effectiveness trial in Liberia" **Country:** Liberia **Category:** water cleaning **Disaster:** conflict	**Method:** RCT **Counterfactual:** Random assignment with factorial model **Rigour:** PA, ToC, ToB, E	Prevalence of diarrhea	(i) improved storage reduced diarrhoea incidence by 90% and prevalence by 83%, when compared with control households with improved water storage alone; (ii) among the intervention group, residual chlorine levels met or exceeded Sphere standards in 85% of observations with a 95% compliance rate
5	Roberts, L. *et al*. (2001) "Keeping clean water clean in a Malawi refugee camp: a randomized intervention trial" **Country:** Malawi **Category:** water cleaning **Disaster:** conflict and poverty	**Method:** RCT **Counterfactual:** Random assignment **Rigour:** ToC, E	(i) Prevalence of diarrhoea; (ii) water contamination	(i) 31% less diarrhoeal disease in children under 5 years of age among the group using the improved bucket; (ii) less water contamination; (iii) proper chlorination is a less expensive and more effective means of water quality protection in comparison with the improved bucket, but was unpopular and rarely utilised by the camp inhabitants
6	Bolton, P. *et al.* (2010) "Interventions for Depression Symptoms Among Adolescent Survivors of War and Displacement in Northern Uganda" **Country:** Uganda **Category:** reconciliation	**Method:** RCT **Counterfactual:** Delayed treatment control group **Rigour:** PA, ToC, ToB	(i) Score on a depression symptom scale; (ii) scores on anxiety, conduct problem symptoms, and function scales (using the Acholi Psychosocial Assessment	(i) Difference in change in adjusted mean score for depression symptoms between group interpersonal psychotherapy and control groups was 9.79 points; (ii) girls receiving group interpersonal psychotherapy showed substantial and significant improvement in depression symptoms compared with controls (12.61 points), and improvement among boys was not statistically

No.	Study	Method & Counterfactual	Outcomes	Main findings
	Disaster: conflict		Instrument)	significant (5.72 points); (iii) creative play showed no effect on depression severity (−2.51 points); (iv) no statistically different improvements in anxiety and conduct problems or function scores in either intervention group

Notes: PA=power analysis, ToC=theory of change; ToB=test of balance, E=ethics

Table 10: Impact evaluation studies of peace-building and conflict prevention interventions (from Samii, Brown and Kulma and Gaarder and Annnan)

No.	Study	Method, Counterfactual & Rigour	Main findings
1 *	Annan, J. and C. Blattman (2011) "Why men don't rebel: experimental results from an ex-combatant reintegration program" **Country:** Liberia **Category:** Ex-Combatant Reintegration	**Method:** RCT **Counterfactual**: Randomised Control Group **Rigour:** -	(i) Increased engagement in agriculture; (ii) participation rates in unlicensed and illicit activities unchanged, but participation levels dropped; (iii) little change in current income and expenditures, but a large rise in durable wealth; (iv) modest improvements in social engagement, citizenship and stability
2	Beath, A. *et al.* (2010) "Randomized impact evaluation of Phase II of Afghanistan's National Solidarity Programme (NSP): estimates of interim program impact from first follow-up survey" link **Country:** Afghanistan **Category:** Peace Dividends	**Method:** RCT **Counterfactual:** Randomised Control Group **Rigour:** PA, B, ToC	(i) Induced changes in village governance through creating village councils and transferring authority to elderly; (ii) improvements in villagers' access to services and perceptions of well-being; (iii) no effect on household income or consumption (objective measures); (iv) increased engagement of women in community life
3 *	Blattman, C. *et al.* (2011a) "Peace Education in rural Liberia" link **Country:** Liberia **Category:** Peace Structures	**Method:** RCT **Counterfactual:** Randomised Control Group **Rigour:** ToC	(i) Little impact on specific measures of civic participation and community cohesion; (ii) modest increases in respect for human rights and equality; (iii) large impacts on conflict and conflict resolution (though not always in expected ways)

No.	Study	Method, Counterfactual & Rigour	Main findings
4 *	Blattman, C. *et al.* (2011b) "Uganda: Enterprises for Ultra-poor Women after War" link **Country:** Uganda **Category:** Victims of War	**Method:** RCT **Counterfactual**: Delayed Treatment Control Group **Rigour:** ToC	A year after the intervention, (i) monthly cash earnings doubled; (ii) cash savings tripled; and (iii) short-term expenditures and durable assets increased 30 to 50% relative to the control group; (iv) most impactful on the people with the lowest initial levels of capital and access to credit; (v) no effect on women's independence, status in the community, or freedom from IPV (no increase in a woman's probability of experiencing IPV); (vi) little effect on psychological or social well-being; (vii) large spillover effects into these small village economies, etc.
5	Blattman, C. *et al.* (2013) "Generating skilled self-employment in developing countries: experimental evidence from Uganda" **Country:** Uganda **Category:** Victims of war (youth)	**Method:** RCT **Counterfactual:** Random assignment **Rigour:** ToC, ToB	(i) Increased business assets by 57%, work hours by 17%, and earnings by 38%; (ii) formalised enterprises and hired labour; (iii) no impact on social cohesion, anti-social behaviour, or protest; (iv) impacts are similar by gender, but are qualitatively different for women because they begin poorer and because women's work and earnings stagnate without the programme but take off with it
6	Casey, K. *et al.* (2011) "Reshaping Institutions: Evidence on External Aid and Local Collective Action" **Country:** Sierra Leone **Category:** Peace Dividends	**Method:** RCT **Counterfactual:** Randomised Control Group **Rigour:** B, ToC	(i) Positive short-run effects on local public goods provision and economic outcomes; (ii) no sustained impacts on collective action, decision-making processes, or the involvement of marginalised groups (like women) in local affairs, indicating that the intervention was ineffective at durably reshaping local institutions
7	Fearon, J. *et al.* (2009) "Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-Conflict Liberia" **Country:** Liberia **Category:** Peace Dividends	**Method:** RCT **Counterfactual:** Randomised group assignment of villages **Rigour:** B, ToC	Villages exposed to a community driven reconstruction programme (CDR) exhibit higher subsequent levels of social cooperation than those in the control group, as measured through a community-wide public goods game

No.	Study	Method, Counterfactual & Rigour	Main findings
8	Fearon, J. *et al.* (2008) "Community-Driven Reconstruction in Lofa County" **Country:** Liberia **Category:** Peace Dividends	**Method:** RCT **Counterfactual:** Randomised Control Group **Rigour:** E, B, ToC	(i) Improvements in communities' ability to act collectively after the programme's completion to improve their own welfare; (ii) reinforced democratic values and practices; (iii) increased social inclusion in beneficiary communities, especially for marginalised groups
9 *	Glennerster, R. and E. Miguel (2010) "The Role Of Information And Radios On Political Knowledge And Participation In Sierra Leone" **Country:** Sierra Leone **Category:** Peace Messaging	**Method:** RCT **Counterfactual:** Randomised Control Group **Rigour:** -	*No results available.* Possible outcomes measured: household and community awareness of politics, current affairs and local councils, attitudes about outsiders, women and local authorities, participation at local village and council meetings and community activities, and differential impacts on women
10	Paluck, E. and D. Green (2009) "Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda" **Country:** Rwanda **Category:** Peace Messaging	**Method:** RCT **Counterfactual:** Clustered random assignment **Rigour:** ToC, E	Although the radio programme had little effect on many kinds of beliefs and attitudes, it had a substantial impact on listeners' willingness to express dissent and the ways they resolved communal problems
11*	Paluck, E. (2009a) "Entertainment, Information, and Discussion: Experimenting with media techniques for civic education and engagement in Southern Sudan" **Country:** Sudan **Category:** Peace Messaging	**Method:** RCT **Counterfactual:** Clustered random assignment with factorial model	*No results available*

No.	Study	Method, Counterfactual & Rigour	Main findings
12	Paluck, E. (2009b) "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda" **Country:** Rwanda **Category:** Peace Messaging	**Method:** RCT **Counterfactual:** Randomised assignment of clusters with matching **Rigour:** B, ToC	The influence of mass media (radio soap opera) changes social behaviour, but not personal beliefs
13	Pugel, J. (2007) "What the Fighters Say: A Survey of Ex-combatants in Liberia" **Country:** Liberia **Category:** Ex-Combatant Reintegration	**Method:** RCT **Counterfactual:** Randomised selection of 20 person clusters **Rigour:** ToC, PA	Significantly, empirical evidence supports the finding that those former combatants who registered with the national DDR programme and completed a course of reintegration training have reintegrated more successfully than those ex-combatants who chose not to participate and reintegrate on their own
14	Paluck, E. (2010) "Is It Better Not to Talk? Group Polarization, Extended Contact, and Perspectives Taking in Eastern Republic of Congo" **Country:** DRC **Category:** Peace Messaging	**Method:** RCT **Counterfactual:** Randomised assignment of clusters with matching **Rigour:** PA, ToC	Compared to individuals exposed to the soap opera only, talk show listeners discussed more but were more intolerant, more mindful of grievances, and less likely to aid disliked community members
15	Barron, P. *et al.* (2009) "Community- Based Reintegration in Aceh: Assessing the Impacts of BRA-KDP" **Country:** Indonesia **Category:** Peace Dividends	**Method:** Quasi Experimental **Counterfactual:** Matched Control Group **Rigour:** ToC	(i) Mixed success in targeting conflict victims as beneficiaries; (ii) welfare gains and improvements in perceptions of wellbeing; (iii) little evidence in improvements in social cohesion or improved awareness of or faith in governmental institutions at the village or at higher levels

No.	Study	Method, Counterfactual & Rigour	Main findings
16	Biton, Y. and G. Solomon (2006) "Peace in the Eyes of Israeli and Palestinian Youths: Effects of Collective Narratives and Peace Education Program" **Country:** Israel **Category:** Consensus & Dialogue	**Method:** Quasi Experimental **Counterfactual:** Matched-pair randomisation of classes in selected schools/natural **Rigour:** -	(i) Peace education can serve as a barrier against the deterioration of perceptions and feelings; (ii) individuals' perceptions of peace are differentially coloured by their group's collective narratives and more immediate experiences of current events, but are significantly altered by participation in a peace education programme
17	Gilligan, M. *et al.* (2010) "Reintegrating Rebels Into Civilian Life: Quasi-experimental Evidence From Burundi" **Country:** Burundi **Category:** Ex-Combatant Reintegration	**Method:** Quasi Experimental **Counterfactual:** Natural control group with matching **Rigour:** PA, B, ToC	(i) Programme resulted in a 20 to 35 percentage point reduction in poverty incidence among ex-combatants and moderate improvement in livelihoods; (ii) no effect on political reintegration: modest increase in propensities to report civilian life as preferable to combatant life, but no evidence that the programme contributed to either more satisfaction with the peace process or a more positive disposition toward current government institutions
18	Humphreys, M. and J. Weinstein (2007) "Demobilization and Reintegration" **Country:** Sierra Leone **Category:** Ex-Combatant Reintegration	**Method:** Quasi Experimental **Counterfactual**: Matched control group **Rigour:** ToC	(i) Wealthier and more educated combatants face greater difficulties in reintegration; (ii) men, and younger fighters are the most likely to retain strong ties to their factions; (iii) little evidence at the micro level that internationally funded programmes facilitate demobilisation and reintegration
19**	Kondylis, F. (2007) "Agricultural Outputs and Conflict Displacement: Evidence from a Policy Intervention in Rwanda" **Country:** Rwanda **Category:** Victims of War	**Method:** Quasi Experimental **Counterfactual:** Natural control group **Rigour:** ToC	(i) Returns on farm labour are higher for returnees (refugees returning to settlement) relative to stayers, although the evidence suggests that the policy contributed little additional effect to this differential; (ii) these differentials suggest that, upon return from conflict-induced exile, returnees are more motivated to increase their economic performance

No.	Study	Method, Counterfactual & Rigour	Main findings
20	Levely, I. (2010) "Reintegration in Post-War Liberia: A Failed Approach or Simply a Failed Program?" **Country:** Liberia **Category:** Ex-Combatant Reintegration	**Method:** Quasi Experimental **Counterfactual:** Matched Control Group	*Unpublished*
21	Malhotra, D. and S. Liyanage (2005) "Long-Term Effects of Peace Workshops in Protracted Conflicts" **Country:** Sri Lanka **Category:** Consensus & Dialogue	**Method:** Quasi Experimental **Counterfactual:** Natural control group **Rigour:** ToC	(i) Compared with two control groups, the participant group showed greater empathy toward members of the "other" ethnicity, even one year after participation in the peace workshops; (ii) consistent with the attitudinal data on empathy, participants donated more money to help poor children of the "other" ethnicity than did nonparticipants
22 **	Mvukiyehe, E. and C. Samii (2009) "Laying a Foundation for Peace? Micro-Effects of Peacekeeping in Cote d'Ivoire" **Country:** Cote d'Ivoire **Category:** Peace Dividends	**Method:** Quasi Experimental **Counterfactual:** Natural control group **Rigour:** ToC, B	(i) Little effect on the security situation; (ii) associated with less severe economic losses and more confidence in forthcoming elections; (iii) no clear association between deployments and the restoration of local authorities
23 **	Mvukiyehe, E. and C. Samii (2011) "Peace from the Bottom Up: A Randomized Trial with UN Peacekeepers" **Country:** Liberia **Category:** Community Security Initiatives	**Method:** Quasi Experimental **Counterfactual:** Matched Clusters (communities)	*Only project description without results is available*

No.	Study	Method, Counterfactual & Rigour	Main findings
24	Mvukiyehe, E. and C. Samii (2010) "Quantitative Impact Evaluation of the United Nations Mission in Liberia: Final" **Country:** Liberia **Category:** Ex-Combatant Reintegration, Peace Dividends	**Method:** Quasi Experimental **Counterfactual:** Cluster matched sampling **Rigour:** B, ToC	*Main conclusion*: humanitarian community can contribute to consolidating the peace in Liberia by (i) supporting the reintegration of newly resettled households; (ii) supporting efforts to foster social and community cohesion, especially among the newly resettled households; and (ii) providing electoral assistance to sustain political interest among ordinary citizens
25	Beath *et al.* (2012) "Empowering women: evidence from a field experiment in Afghanistan" **Country:** Afghanistan **Category:** Victims of War (women)	**Method:** RCT **Counterfactual:** Matched-pair randomisation of clusters **Rigour:** ToC, ToB	(i) Increased female mobility and involvement in income generation; (ii) unchanged female roles in family decision-making or attitudes toward the general role of women in society
26*	IRC & NYU (2011) "Opportunities for Equitable Access to Quality Basic Education (OPEQ)." **Country:** DRC **Category:** Victims of War (children)	**Method:** RCT **Counterfactual:** Delayed treatment control group **Rigour:** ToC	*Only baseline report available*
27	IRC (2012) "Measuring Impact: Survivors' Social, Psychological and Economic Well-Being" **Country:** DRC **Category:** Victims of War	**Method:** RCT **Counterfactual:** Random assignment with factorial model **Rigour:** -	*http://www.nejm.org/doi/full/10.1056/NEJMoa1211853*
28**	IRC (2008) "Getting down to business: Women's economic and social empowerment in Burundi" **Country:** Burundi **Category:** Victims of War (women)	**Method:** RCT **Counterfactual:** Random assignment with factorial model **Rigour:** -	*No baseline report available* (i) Incidence of IPV decreased; (ii) women reported increased decision-making; (iii) use of negotiation skills increased; (iv) acceptance of violence decreased

No.	Study	Method, Counterfactual & Rigour	Main findings
29	IRC (2011) "Urwaruka Rushasha: A Randomized Impact Evaluation of Village Savings and Loans Associations and Family-Based Interventions in Burundi" **Country:** Burundi **Category:** Victims of War	**Method:** RCT **Counterfactual:** Random assignment with factorial model and delayed treatment control group **Rigour:** PA, ToC, ToB	*Mid-term results:* (i) Increased assets and consumption; (ii) decreased harsh physical and verbal discipline in the home, improved communication between children and caregivers, and a decrease in family problems, including violence and intoxication of family members Final report: http://www.rescue.org/sites/default/files/resource-file/New_Generation_Final_Report_05312013.pdf
30	Humphreys, M. *et al.* IRC & CARE. (2011) "Social and Economic Impacts of *Tuungane"* **Country:** DRC **Category:** CDR	**Method:** RCT **Counterfactual:** Random assignment **Rigour:** PA, ToC, ToB	(i) *Tuungane* was successful in implementing a large number of projects in the target areas and the projects were in line with the preferences of the populations; (ii) populations reported very high levels of exposure to project activities and satisfaction with the outcomes of the project; (i) failure to find evidence that the positive experiences with the *Tuungane* intervention led to behavioural changes

Notes: * ongoing studies, ** preliminary results, PA=power analysis; ToC=theory of change; ToB=test of balance; E=ethics

Table 11: Impact evaluations of unanticipated disasters

No.	Study	Method, Counterfactual & Rigour	Main findings
1	De Mel *et al.* (2010) "Enterprise recovery following natural disasters" **Country:** Sri Lanka **Category:** Entrepreneurship **Disaster:** tsunami	**Method:** RCT **Counterfactual:** Delayed treatment control group **Rigour:** ToC, ToB	(i) Positive effect of grant programme on profits, representing a 9.9 per cent real monthly return on the treatment; (ii) direct aid is more important in the recovery of enterprises operating in the retail sector than for those operating in the manufacturing and service sectors; (iii) the use of cash grants is more helpful than the use of in-kind, but only in limited cases
2	Shoji, M. (2010) "Does contingent repayment in microfinance help the poor during natural disasters?" **Country:** Bangladesh **Category:** Microfinance **Disaster:** floods	**Method:** Quasi experimental **Counterfactual:** Before/after comparison **Rigour:** ToC	(i) Decreasing probability that people skip meals during negative shocks by 5.1 per cent, with a higher effect on landless and females; (ii) the authors did not estimate the effects on nutritional outcomes, and no conclusions could be made about whether these households are better off nutritionally

Notes: PA=power analysis; ToC=theory of change; ToB=test of balance; E=ethics

Publications in the 3ie Working Paper Series

The following papers are available from http://www.3ieimpact.org/3ie_working_papers

What methods may be used in impact evaluations of humanitarian assistance?, *3ie Working Paper 22*. Puri, J, Aladysheva, A, Iversen, V, Ghorpade, Y and Brück, T (2014)

Impact evaluation of development programmes: Experiences from Viet Nam, 3ie Working Paper 21. Nguyen Viet Cuong (2014)

Quality education for all children? What works in education in developing countries, 3ie Working Paper 20. Krishnaratne, S, White, H and Carpenter, E (2013)

Promoting commitment to evaluate, 3ie Working Paper 19. Székely, M (2013)

Building on what works: commitment to evaluation (c2e) indicator, 3ie Working Paper 18. Levine, CJ and Chapoy, C (2013)

From impact evaluations to paradigm shift: A case study of the Buenos Aires Ciudadanía Porteña conditional cash transfer programme, *3ie Working Paper 17*. Agosto, G, Nuñez, E, Citarroni, H, Briasco, I and Garcette, N (2013)

Validating one of the world's largest conditional cash transfer programmes: A case study on how an impact evaluation of Brazil's Bolsa Família Programme helped silence its critics and improve policy, *3ie Working Paper 16*. Langou, GD and Forteza, P (2012)

Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework, 3ie Working Paper 15. White, H and Phillips, D (2012)

Behind the scenes: managing and conducting large scale impact evaluations in Colombia, *3ie Working Paper 14*. Briceño, B, Cuesta, L and Attanasio, O (2011)

Can we obtain the required rigour without randomisation? 3ie Working Paper 13. Hughes, K and Hutchings, C (2011)

Sound expectations: from impact evaluations to policy change, 3ie Working Paper 12. Weyrauch, V and Langou, GD (2011)

A can of worms? Implications of rigorous impact evaluations for development agencies, 3ie Working Paper 11. Roetman, E (2011)

Conducting influential impact evaluations in China: the experience of the Rural Education Action Project, *3ie Working Paper 10*. Boswell, M, Rozelle, S, Zhang, L, Liu, C, Luo, R and Shi, Y (2011)

An introduction to the use of randomized control trials to evaluate development interventions, *3ie Working Paper 9*. White, H (2011)

Institutionalisation of government evaluation: balancing trade-offs, *3ie Working Paper 8.* Gaarder, M and Briceño, B (2010)

Impact Evaluation and interventions to address climate change: a scoping study, *3ie Working Paper 7*. Snilstveit, B and Prowse, M (2010)

A checklist for the reporting of randomized control trials of social and economic policy interventions in developing countries, *3ie Working Paper 6.* Bose, R (2010)
Impact evaluation in the post-disaster setting, 3ie Working Paper 5. Buttenheim, A (2009)

Designing impact evaluations: different perspectives, contributions, 3ie Working Paper 4. Chambers, R, Karlan, D, Ravallion, M and Rogers, P (2009) [Also available in Spanish, French and Chinese]

Theory-based impact evaluation, *3ie Working Paper 3*. White, H (2009) [Also available in French and Chinese.]

Better evidence for a better world, *3ie Working Paper 2.* Lipsey, MW (ed.) and Noonan, E (2009)

Some reflections on current debates in impact evaluation, *3ie Working Paper 1*. White, H (2009)

Since 2005, more than US$90 billion has been spent on humanitarian assistance. Humanitarian crises are complex situations where demand for aid very often exceeds supply. The humanitarian assistance community has long asked for better evidence on how aid money can be spent more effectively.

This paper explores the methodological options and challenges associated with collecting and generating high-quality evidence needed to answer important questions on the impact of humanitarian assistance. These questions include whether assistance is reaching the target populations and at the right time, whether it is bringing about desired changes in their lives and whether it is being delivered in effective doses and ways, with manageable costs.

The paper also uses case studies to discuss methods for undertaking impact evaluations to address these concerns in a range of humanitarian contexts, from unanticipated natural disaster-related emergencies to protracted crises.

Working Paper Series

International Initiative for Impact Evaluation
c/o Global Development Network
2nd Floor, West Wing, ISID Complex
Plot No. 4, Vasant Kunj Institutional Area
New Delhi – 110070
India

3ie@3ieimpact.org
Tel: +91 11 4323 9494

www.3ieimpact.org