RANDOMISED IMPACT EVALUATION IN HUMANITARIAN ACTION

A Learning Agenda for the Humanitarian Initiative

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
EUROPE

Foreign, Commonwealth & Development Office

March 2022

TABLE OF CONTENTS

EXECUTIVE SUMMARY

What are effective mechanisms for delivering needed humanitarian assistance, promoting resilience and remedying harm for civilian populations affected by crisis and conflict? This document provides a guide to how rigorous impact evaluation can provide answers to these questions and to how this research can be deployed in humanitarian settings. While improving accountability (both to funders and to beneficiaries) is an important goal of evaluation, our aim here is to show how randomised evaluations can be useful to the sector in promoting a rich understanding of how humanitarian action can achieve its aims as well as in providing a robust pathway for validating and scaling innovative programming and informing discussions on how to improve cost-effectiveness across the sector. We hope to provide a springboard for launching more rigorous evaluations of programming and interventions within the humanitarian sector.

This is a broad field; we have focused primarily on forms of short-term assistance and action designed to prevent and alleviate human suffering in crisis settings. We have also paid particular attention to actions that seek to deliver humanitarian protection outcomes, an area where there has been particularly little impact evaluation to date. Throughout the document, we refer to humanitarian 'action', 'relief', 'interventions' or 'assistance' as generic terms to encompass activities rooted in the humanitarian principles, but we recognise that not all actions that contribute to protection outcomes in particular are captured by these terms.

We have written this guide with at least two different audiences in mind: humanitarian practitioners and policymakers who are interested in better understanding how randomised evaluations can or cannot help them reach the learning and evaluation goals they may have, and researchers, evaluators and others who are well versed in the methods of rigorous impact evaluation but would like to better understand the policy questions and practical constraints of this field.

The document begins with a review of the programming and policy questions that humanitarian practitioners and policymakers are asking, collected during conversations and consultations in the autumn and winter of 2021–22. We then review some of the operational and conceptual challenges to designing and deploying randomised evaluations on these questions and propose potential solutions or alternative strategies based on the rich body of experience accumulated in the network of J-PAL-affiliated researchers.

To set out what is possible, both in terms of what we may learn and how we might structure compelling research to get us there, this document draws on a large and rapidly expanding body of experimental research in fragile settings as well as in adjacent fields. With new approaches to designing experiments, this research has broken new ground in bringing experimental methods to questions and settings where the approach once did not seem feasible. By reviewing the findings as well as the designs of these evaluations here, we try to provide inspiration and help practitioners and researchers make the case for randomised evaluations in the humanitarian space. The goal is not to provide an exhaustive summary—there is a longer history of using randomised evaluations in the public health field, for example, that we do not fully cover here. Instead, we present an illustrative range of experimental research that was deployed in fragile or humanitarian contexts or that may ask similar questions to those posed by practitioners (or faced similar challenges in their deployment), or that has asked similar questions in nonhumanitarian settings (tackling issues of targeting, norms change or information diffusion, for example) and that we believe may be worth borrowing from.

In conclusion, we found considerable demand for impact evaluation of protection programming in particular and present a case for the feasibility and potential of a dedicated research initiative for randomised evaluations in the humanitarian sector. This includes how it might alleviate obstacles—perceived or otherwise—that have limited the scope of rigorous evaluation in the sector so far as well as ideas on the knowledge and evidence research that it is likely to produce over the next five years.

BUILDING A NEW RESEARCH FOCUS: LAYING THE FOUNDATIONS FOR J-PAL'S HUMANITARIAN INITIATIVE

Building on the expertise of the J-PAL network in launching over a thousand randomised evaluations around the world, and with seed funding from the UK Foreign, Commonwealth & Development Office, J-PAL's Europe office is laying the foundations for a new research initiative to identify effective models of humanitarian action, with a particular focus on improving protection outcomes.

In addition to creating this learning agenda to outline the areas where demand and opportunities for more rigorous evidence exist, the foundational phase included activities to build a conducive environment for more rigorous evaluation in the humanitarian space from October 2021 onwards:

- In late 2021, J-PAL conducted its first online training on the basics of randomised evaluations tailored specifically to humanitarian practitioners, with just under 50 participants.

- In early 2022, J-PAL invited five selected organisations to a three-day evaluation design workshop to think through the key elements of a randomised evaluation for one of their own programmes.

REFUGEES FROM THE CENTRAL AFRICAN REPUBLIC HEADING TO A FOOD DISTRIBUTION SITE IN SOUTHERN CHAD. PHOTO: DOMINIQUE CATTON, © 2017 EUROPEAN UNION, CC BY-NC-ND 2.0

I. INTRODUCTION: WHY EVALUATE?

Meaningful impact evaluation begins with a careful investment in rigorously capturing the impact of a programme or policy, and it then extends to the consideration of how we can improve our action based on the results. If we have evidence of effectiveness, can we leverage this to expand the reach of a programme or to reduce its costs? If we find evidence of no change, or even of harm, can we scale down programming and redirect resources elsewhere? How can what we have learned drive further innovation or a shift in how we approach programming more generally?

The scale of humanitarian challenges demands a sophisticated understanding of the effectiveness of programming. The scale of these challenges continues to grow: an estimated 274 million people will be in need of assistance and protection in 2022 (UN Office for the Coordination of Humanitarian Affairs, 2021), a 17 percent increase over last year.

Rigorous evidence on effective programming in the humanitarian sector is limited. While the sector invests in research, learning and evaluation, only a small part of the existing research can establish causal links between programming and outcomes. Even in the burgeoning literature around the effectiveness of cash transfers in humanitarian settings, the most recent systematic review found just four randomised control trials (RCTs) (Doocy and Tappis 2017), while a recently released review focused on health outcomes found seven (van Daalen et al. 2022).

The lack of rigorous evidence is particularly marked with regards to protection programming. A 2013 review identified 'only a few sophisticated attempts at measuring the success of different types of protection interventions' and concluded that it was 'generally easier to find negative examples of humanitarian protection efforts than positive ones' (Reichhold, Binder and Nilan 2013). The situation does not appear to have changed much nearly a decade later.

We identify only a few randomised evaluations of protection interventions in a review of experimental research in humanitarian settings.[1] As always, the absence of evidence of effectiveness should not be confused with evidence of no impact. However, we have heard from several protection practitioners that investing in the evidence base for protection programming would help with more informed decision-making on which activities to pursue in resource-constrained environments, cut ineffective or at best symbolic activities, and fundraise for meaningful interventions.

WHAT DOES PROTECTION MEAN IN THE CONTEXT OF HUMANITARIAN ACTION?

Protection is concerned with the safety, dignity and rights of people affected by disaster or armed conflict. It is about taking action to keep people safe from violence, coercion and deliberate deprivation.

As such, it encompasses all efforts pursued by humanitarian and human rights actors to ensure that the rights of affected persons and the obligations of duty bearers under international law are understood, respected, protected and fulfilled without discrimination.[2]

In practice, there is often a priority set of protection concerns in any given humanitarian context. It is vital to understand and address these key risks to those affected if they are to be effectively tackled.

—*Adapted from the Sphere Association (2018)*

This learning agenda sets out a vision for how randomised evaluations can usefully facilitate learning in humanitarian programming, with a focus on humanitarian protection. With its concern for interrogating causal links between interventions and outcomes, rigorous impact evaluation offers a powerful tool for ensuring that the principles established in international humanitarian and human rights law are borne out in concrete outcomes for protected populations. Monitoring indicators can only deliver one part of the picture if we are uncertain of how different outcomes link to one another or deliver upon core protection principles. By helping us refine our understanding of how different components of assistance may or may not link to one another, we can better establish links between the forms of protection assistance provided and the outcomes we care about. If randomised evaluations are a powerful tool, they are also *precise* and *actionable*. The research described in this learning agenda also seeks to answer how programming may impact different groups (e.g. by age, gender, disability) in different ways and to examine what different bundles of treatment or delivery mechanisms generate the greatest impact, at the lowest costs.

Rigorous evaluation is also a way to help validate the piloting and testing of innovative approaches to long-standing problems so that proven innovations can be scaled. These may derive from a variety of sources: technological advances, shifts in operational constraints (such as those recently imposed by the Covid-19 pandemic), structural changes to the way that humanitarian assistance is designed and delivered and new ideas born out of new thinking to address evolving problems. By rigorously testing these innovations before they move to scale, practitioners can leverage evidence of impact to marshal resources for scaling. In addition, as scaling innovative approaches following proof of concept usually necessitates some degree of adaptation, accompanying this process with rigorous evaluation also helps practitioners make informed decisions along the way.

In the consultations and trainings led by J-PAL during this foundational phase of the Humanitarian Initiative, we undiscovered considerable demand for integrating randomised evaluations into humanitarian programming and policy. The next chapter summarises this demand, and we explore in subsequent chapters how we think randomised evaluations can be structured to provide the right answers.

---

[1] Protection covers a broad field of activities (c.f. box on page 4). Existing evaluations pertaining directly to protection interventions are primarily from the public health literature, where we identified some 20 randomised evaluations focused on interventions to remedy the psychological harm stemming from violence in humanitarian settings. A limited number of other studies look at reducing individual or community-level violence in societies affected by conflict or examine measurement strategies to capture protection outcomes, but these are largely conducted outside humanitarian settings and were not conceived of as evaluations of protection outcomes. Chapter IV showcases some of the ways in which these studies, as well as experimental research from adjacent fields, can inform protection programming and humanitarian action more broadly.

[2] The Inter-Agency Standing Committee (IASC) defines protection as 'all activities aimed at obtaining full respect for the rights of the individual in accordance with the letter and the spirit of the relevant bodies of law (i.e. International Human Rights Law, International Humanitarian Law, International Refugee Law).' For a summary of the applicable normative frameworks, refer to the IASC's policy on protection in humanitarian action (2016, pp. 16-24).

A REFUGEE WORKING ON GROWING HIS OWN VEGETABLES TO INCREASE FOOD SECURITY AND SELF-RELIANCE IN MAURITANIA. PHOTO: JOSE CENDON, © 2016 EUROPEAN UNION, CC BY-NC-ND 2.0

II. WHAT DO HUMANITARIAN ACTORS WANT TO KNOW?

To maximise the impact that randomised evaluations can have on improving outcomes for affected populations, they must engage with—and answer—questions that practitioners and funders are asking. In conversations with a range of humanitarian practitioners and funders in late 2021 and early 2022 (see list of organisational affiliations at the end of this document), we sought to gather input on what sorts of learning and evaluation questions were top of mind for practitioners, without reference to methodological concerns (i.e. whether or not randomised evaluations were the best tool). Our focus was upon understanding protection-related questions, but we also captured broader questions about improving humanitarian assistance generally.

We summarise the issues raised in these conversations in two different categories below. First, we present a set of questions that apply to the evaluation of humanitarian programming more broadly—questions about how programmes are designed and delivered, how and whom to target, how to measure outcomes over different time frames, how to deliver assistance within constrained budgets and how to safeguard beneficiaries. Many of these questions are common to rigorous impact evaluations that have been launched in other settings (and explored in Chapter V below).

A second set of questions focuses on the substance and theory of protection-related programming. These reflect an interest in understanding more concretely how to deliver protection outcomes for the most vulnerable populations, with a particular focus on conflict settings. These concerns are generally newer to impact evaluation, and we explore potential solutions to challenges they may throw up in Chapter IV.

A. HOW CAN WE DESIGN MORE EFFECTIVE HUMANITARIAN PROGRAMMES?

How, on what timeframe, with what intensity, and in what configuration programming is delivered are classic concerns of impact evaluation and they came up frequently in the conversations we had with practitioners.

Targeting: What are the most effective strategies to identify beneficiaries and target assistance?

Most humanitarian action—driven by observance of the principle of impartiality as well as important resource constraints—relies on the targeted allocation of support to those in most need. But the markers of greatest need and vulnerability are not always readily observable, especially when it comes to protection risks. There may not be sufficient resources (e.g. time, money, staffing) for collecting exhaustive data, or there may be lingering uncertainty about the appropriate targeting thresholds.

Inclusion: How does the impact of my programme vary for different subgroups of the target population and are we reaching those most in need?

Designing effective targeting strategies requires a nuanced understanding of how needs may differ among different subgroups. Questions about understanding differential effects to better design and target interventions come in different forms: Does a programme manage to reach people with disabilities, and if so, is it equally successful in moving relevant outcome indicators for this subgroup? How can we deliver assistance in a way that does not exclude those who need the support most? Which member of a household should be the primary beneficiary, both to reduce harm and maximise a programme's impact?

Spillovers: Looking beyond direct beneficiaries, how does my programme affect the wider community?

Is the programme affecting the broader community through positive or negative spillovers? Are harm minimisation strategies needed to address negative spillovers? What are the trade-offs between blanket distribution and a more targeted approach?

Relative effectiveness: What type of assistance serves our target population best and how intense should the support or interaction with programme participants be to create meaningful impact?

The broadest category of questions we heard from practitioners were concerned with the relative effectiveness of different approaches to address a specific need. Another related set of questions looks at evaluating the intensity of assistance, particularly in thinking about how to either incorporate cash assistance or offer cash assistance as an alternative to traditional programming.

WHAT ARE THE HUMANITARIAN PRINCIPLES, AND WHAT DOES THEIR APPLICATION IMPLY?

Humanitarian action is guided by four widely accepted principles:

- **Humanity**: Human suffering must be addressed wherever it is found. The purpose of humanitarian action is to protect life and health and ensure respect for human beings.

- **Impartiality**: Humanitarian action must be carried out based on need alone, giving priority to the most urgent cases of distress and making no adverse distinction on the basis of nationality, race, gender, religious belief, class or political opinion.

- **Independence**: Humanitarian action must be autonomous from political, economic, military or other objectives that any actor may hold with regard to areas where humanitarian action is being implemented.

- **Neutrality**: Humanitarian actors must not take sides in hostilities or engage in controversies of a political, racial, religious or ideological nature.

The application of the humanitarian principles is essential to distinguish humanitarian action from other forms of activities that may pursue similar goals or operate in the same environments. For researchers, it is important to keep in mind that, in practice, this may place important normative or operational constraints on the kind of variation that experimental evaluations can seek to introduce and will require careful consideration in addition to the established standards for ethical research.[3]

— Adapted from CHS Alliance, Group URD and Sphere Project (2014, p. 8)

Delivery: How does the way in which programming is delivered affect uptake and impact?

This question takes different forms depending on the type of programming or intervention intended, but it concerns identifying the trade-offs in terms of effectiveness or costs to delivering a programme through different channels. For example, practitioners may wonder whether cash, mobile money or in-kind transfers of vouchers all deliver similar outcomes. Or whether, when seeking to remedy psychological harm from conflict, how group counselling sessions may compare to peer support groups or individual therapy sessions with a counsellor.

---

[3] See J-PAL's online resources on "Ethical conduct of randomised evaluations" for a discussion of the principles guiding ethical research (J-PAL 2021). The International Rescue Committee's Airbel Impact Lab has also created resources developed specifically for humanitarian contexts (International Rescue Committee 2022).

Bundling interventions with cash: Can adding cash potentially extend or broaden impacts?

There is considerable interest in exploring the ways that adding small cash transfers alongside other kinds of programming (mental health or nutrition interventions, training and other interventions that seek to give participants new skills or capacities) can potentially help extend their impacts, either directly or by creating positive spillovers on households and communities.

Cost-effectiveness: Can we deliver more for less?

Organisations raised questions around how to deliver programming within budget constraints or whether programmes could potentially deliver in more cost-effective ways. These considerations may be particularly important when examining how to adapt resource-intensive programmes that have been proven to deliver results in noncrisis settings. Practitioners are interested in understanding which parts of the interventions are most cost-effective to be able to focus their attention on delivering those components well in more complex environments.

Sustainability: Can we measure impact beyond the ttlife cycle of the project to see if impacts are durable and persist? Are there long-term effects that we should consider when designing short-term assistance?

Designed to deliver life-saving assistance, humanitarian programmes are usually delivered on short time frames. But while project time frames may be short, organisations we spoke to were interested in learning about outcomes over both the short and long run. Some of these questions focused on outcomes that may straddle traditional divisions between humanitarian and development programming, but a well-designed impact evaluation can yield robust impact estimates for short-term impacts and to test whether these persist. Increasing the research in this area may also be a way to better understand how to develop approaches that build beneficiary resilience/independence from aid.

Accountability and safeguarding: How can we improve programming based on robust feedback mechanisms and make sure to better prevent and remedy abuse and risks?

Working with highly vulnerable populations, some organisations raised their interest in learning how to improve processes for ensuring accountability to beneficiaries and the safeguarding of programme participants and staff. These include questions surrounding the design of effective feedback and complaint mechanisms as well as evaluating policies on the treatment of serious allegations against staff

members. In establishing effective feedback mechanisms, there is also the potential to create more participatory programme design and implementation processes.

Two other topics emerged from our conversations that concern humanitarian action more broadly but are less 'classical' concerns of programme evaluation. Instead, they are rooted in active debates within the sector at present over how the financing and operationalisation of assistance should be structured.

Anticipatory action: Should we provide assistance before it is needed?

As our ability to model and forecast shocks advances, particularly in the face of increased threat of climate-related shocks, there is growing interest in anticipatory action consisting of providing funding and assistance before a shock. Can this kind of funding, even if it may potentially introduce a locally inefficient allocation of resources (if forecast shocks do not occur), mitigate human suffering before it occurs and thus deliver greater impact and better outcomes on average on a broad scale? Can it be more cost-effective?

Beyond weather and food shocks, there is increasing interest in trying to find ways to anticipate and forecast the outbreak of conflict. Are there ways to develop and harness early warning tools to drive effective anticipatory action?

Localisation: Can localising humanitarian response improve the quality and efficiency of aid?

A broad alliance of stakeholders have called for an increased localisation of humanitarian assistance—that is, for assistance to be structured and delivered by organisations built in affected countries or led by the vulnerable communities being served (see, e.g., Barbelet et al. 2021). One subject for careful evaluation is to consider the effects of this on delivery and protection outcomes. Does staff diversity and/or the involvement of affected populations at all levels of humanitarian interventions produce better outcomes? Do local actors and networks produce better outcomes than international actors? If so, how can international actors provide support?

B. HOW CAN WE IMPROVE HUMANITARIAN PROTECTION OUTCOMES?

As acknowledged above, protection interventions have been the subject of very limited impact evaluation so far. We heard a range of different questions from organisations about how to conceive, conceptualise and measure the impact of protection-related programming.

Support for well-being and remedying the effects of violence, coercion and deliberate deprivation

People affected by a crisis are faced with shocks and stressors of varying intensity, including the experience of loss, physical and emotional trauma, and economic shocks. Established support systems are overstretched or otherwise unable to provide remedy and support healing, while a high level of general violence in society may also lead to a spike in violence among intimate partners and peers. In these situations, humanitarian actors step in to provide medical, psychological and psychosocial support. Identifying those most in need of support and delivering remedial treatment to, for example, survivors of sexual violence effectively and without causing harm remains an area where practitioners are looking for evidence-backed practice. They are also looking at how to effectively provide well-being services and teach well-being practices to help children and young people recover from trauma (and reinstate their learning ability after traumatic events).

Recognition, protection and uptake of rights

Activities to promote and protect rights in crisis situations can take different forms. Questions in this area range from the impact of making changes to the legal framework and policies in a given country to the effectiveness of providing legal assistance to vulnerable people or running campaigns to help raise awareness of rights and individuals' ability to claim these rights. For example, how does refugee status recognition impact individuals or households? Does lowering legal barriers to formal work impact target populations' levels of income or well-being and reduce work-related abuse? What is the effect of improving housing, land and property rights provisions that help clarify claims to legal ownership and strengthen tenants' rights? And what effect do information campaigns on rights and obligations targeted at those at risk of abuse, at possible perpetrators or other actors have?

Improving detention conditions and preventing detention-related abuse

Detention facilities present important protection-related challenges where the levers for effectively influencing outcomes may not be clear, particularly in conflict settings. Environment-building activities by humanitarian actors can include efforts to prevent abuses, such as torture, disappearances and extrajudicial killings; to maintain access to essential health services; or to establish contact between detainees and their families. Questions in these settings include how to structure dialogue with prison authorities and what sorts of training activities may be effective as well as how prison governance can be improved more broadly.

Preventing (the escalation of) conflict

A fundamental approach to promoting protection at the broadest level can entail promoting the resolution or de-escalation of conflict as well as efforts to prevent conflict from erupting. Humanitarian actors may explore ways to promote access to nonviolent livelihoods for (potential) combatants in an effort to lessen the incidence of violence. Interventions to provide economic security (e.g. through insurance products for pastoralists), to resolve latent community conflict (e.g. through more effective resolution of land conflicts), or to build social capital, increase social cohesion, and mitigate intergroup intolerance among communities might ultimately deliver protection-related outcomes. Similarly, there are questions of whether protective actions (e.g. following up and resolving disappearances) may lay the groundwork for more effective peacebuilding efforts.

Promoting restraint among armed actors

One area of protection programming is concerned with the promotion of restraint among armed actors. Can direct dialogue with armed actors (privately or in public) reduce the use of violence? Are training activities effective in promoting adherence to international humanitarian law? How can humanitarian actors leverage the influence of third parties (e.g. religious groups or influential elders)? Is public or private criticism more effective?

Prevention of sexual- and gender-based violence

There is a growing evidence base on interventions to reduce sexual- and gender-based violence (SGBV), but the relevance of these approaches to conflict settings and how they may need to be adapted is an area for further research. Practitioners said they are uncertain about how these findings may apply, in part because the drivers behind such violence in conflict settings may differ (Hossain and McAlpine 2017, p. 1–24) and thus require a different approach. Examining what elements of successful violence reduction strategies can be imported into conflict settings and how to make these successful amid resource and capacity constraints in a crisis setting is an important issue. There is also interest in exploring how economic interventions that target either victims or those who commit such violence can reduce incidence.

Concluding this chapter, we take from our conversations with practitioners that there is a clear need for more rigorous evidence in the humanitarian space, including in specific protection as a sector. The conversations have also given us a sense for the questions that practitioners have about the applicability and feasibility of randomised evaluations as a methodology in humanitarian settings, which we discuss in the next chapter.

WOMEN COLLECTING WATER IN AREAS AFFECTED BY SEVERE DROUGHT IN ETHIOPIA. PHOTO: © EUROPEAN UNION/ECHO/MELAKU ASEFA

III. CONSIDERING RANDOMISED EVALUATIONS IN HUMANITARIAN CONTEXTS: CHALLENGES AND POTENTIAL SOLUTIONS

All research conducted in crisis or fragile settings can pose important challenges, and randomised evaluations are no exception. The need to work with sufficiently large samples, the complexity of design processes, and the need to establish and maintain a robust comparison group can also present challenges that may be unique (at least in scale) to randomised evaluations. In this section, we review some of the logistical, methodological and ethical issues that have made impact evaluation difficult in this sector in the past.

A. ANTICIPATING COMMON IMPLEMENTATION CHALLENGES

Timelines

Humanitarian crises—both their onset and evolution—are often difficult to predict, and humanitarian assistance is designed to offer short-term, immediate relief or protection. As a result, programmes are typically designed and rolled out quickly, without much time to consider drawing up customised research or evaluation designs from scratch.

The funding available for projects is generally short term (less than a year), and the outcomes on which actors are asked to deliver and assessed upon focus on time frames of less than one year. In crisis contexts, it is difficult to foresee frequent changes in the context that may require additional resources.

All of these time pressures can make the careful design and implementation of randomised evaluations difficult and can limit our ability to capture medium- and long-term outcomes in addition to shorter-term impacts.

STRATEGIES TO CONSIDER

Predesigned evaluations. When seeking to evaluate programmes that follow a relatively standardised model (e.g. cash transfers in emergency assistance, psychosocial support modules for survivors of violence), organisations could work with researchers to design the bulk of an evaluation design in advance (required sample size, predesign survey questions, pre-identify existing data sources that could be leveraged).

Leverage pre-established knowledge. Despite quick funding and deployment cycles for humanitarian assistance, humanitarian crises are often prolonged in duration and humanitarian organisations are active in a specific space for multiple years over which they have set up staff and infrastructure and have developed in-depth knowledge of the context. Leveraging this can enable organisations and researchers to design and implement randomised evaluations more quickly and effectively (Quattrochi et al. 2020).

Funding mismatches

Compounding the challenge of quick project cycles is the issue of limited evaluation budgets that may often be embedded within project budgets and thus compete with implementation costs for funding. Organisations (and particularly 'localised' efforts in low- and middle-income countries) will struggle to set aside programme budgets and expertise to focus on research when these resources are perceived to come at the expense of investment in assistance to vulnerable populations.

In addition to the question of the *scale* of funding available is the challenge of the *timeframes* for funding: if research and evaluation budgets are tied to the project cycle, then it is impossible to gather follow-up survey data to understand how durable impacts are.

STRATEGIES TO CONSIDER

Generating dedicated funding pools for research. These must be allocated above the programme level because few actors will be able to justify moving individual programme budgets into research at the expense of implementation. This makes sense since the learning benefits of evaluation in the sector will accrue far above the level of individual programmes.

Pre-allocating funding for evaluations. Earmarked funding for evaluation can ensure swift deployment when crises occur and can make strategic up-front investment in evaluation design and partnerships possible.

Isolation of programme impact from other influencing factors

To isolate the real impacts of a programme with a randomised evaluation, researchers must be able to account for other factors that could influence the outcome of interest and may affect treatment and comparison groups unequally.

Humanitarian settings and crises contexts pose extraordinary challenges to this task as there are often multiple actors providing different forms of assistance to crisis-affected populations. In principle, if a given intervention is assigned randomly, treatment and control communities will, on average, benefit from the other interventions in the same proportions. The additional impact of the intervention of interest is thus well isolated. There is, however, a risk that other programmes will react to the randomisation and concentrate their action on control communities. This would make it difficult for researchers to single out the impact of a specific programme.

STRATEGIES TO CONSIDER

Improving communication and, where possible, coordination between humanitarian actors. The importance of coordinating humanitarian action is widely acknowledged, first and foremost to ensure gaps and overlaps in assistance are minimised. Coordination structures that do exist (e.g. the cluster system) may be one forum for communicating research plans and how other programming may affect the study population. Practitioners involved in randomised evaluations in humanitarian settings have flagged the importance of adequate staffing for coordination to see evaluations through to completion.

CHILDREN PLAYING UNDER THE RAIN IN THE BALUKHALI REFUGEE CAMP IN COX'S BAZAR, BANGLADESH. PHOTO: ALLISON JOYCE, UN WOMEN, CC BY-NC-ND 2.0

Scale

Randomised evaluations require large samples to yield statistically precise results. Sample size requirements for a given intervention can also be further inflated by the characteristics of the programme design/implementation (e.g. by low rates of take-up of the intervention or high levels of heterogeneity among the programme beneficiaries). In humanitarian interventions, sample size may be limited by the intervention's scale and the cost of data collection.

Even when in contexts where the intervention can enrol enough beneficiaries for an adequately powered evaluation, downstream impacts on the evaluation costs (e.g. data collection, tracking and following up with beneficiaries in the case of mobile populations) may discourage practitioners from conducting an evaluation.

STRATEGIES TO CONSIDER

Keep scale in mind from the outset when considering randomised evaluations. Some humanitarian programming will never reach the scale necessary for a well-powered randomised evaluation, and programmes that intend to only target very limited populations should probably rule out an RCT.

Reducing the burden of data collection to limit costs when scale is achievable. Designing light survey instruments to capture essential data only or relying on administrative data when available, if the costs associated with a sufficiently powered evaluation are prohibitory.

Localisation

Localising the implementation and delivery of humanitarian response is an emerging priority in the sector. Innovations in humanitarian response may be more impactful when designed through a bottom-up approach (Bloom and Betts 2013). This puts local humanitarian operators in a unique position to develop new approaches for delivering humanitarian assistance. But they may lack the organisational capacity to test them rigorously since implementing randomised evaluations requires specific expertise and is resource intensive (time, planning, survey costs), which likely makes it prohibitive for smaller organisations running locally led programming.

STRATEGIES TO CONSIDER

Dedicated external evaluation funding. Understanding the impact of a shift towards locally led efforts will likely require pairing organisations with external researchers and ensuring dedicated resources are assigned to this question, even more so than for international actors.

Investment in local evaluation capacities. Reducing the burden of programme teams in humanitarian settings is important to make randomised evaluations work in humanitarian settings (Quattrochi et al. 2020). This can be done by investing in training (or identification of) dedicated local impact evaluation staff.

Evaluating outcomes that are several degrees removed from the intervention

Many protection interventions, including those focused on risk reduction and restraint, are based on theories of change with several levels of causal links (i.e. multiple intermediate outcomes that build on each other before resulting in the final outcome of interest). This could, for example, be an activity targeted at one set of actors (e.g. a community) that is designed to influence their behaviour towards a different group (e.g. engage in a dialogue with armed actors in their vicinity when relevant) so that the latter agree to change their course of action (e.g. respect a local peace agreement). The outcomes that we care about are among the vulnerable population (e.g. fewer civilian casualties) but there are important intermediary stages that need to materialise for the intervention to be successful. In the case of international humanitarian law (IHL) trainings for members of armed forces that aim at promoting

restraint, an important intermediary step may be to determine whether armed forces attend these trainings, digest the material, find it salient, form the intention to prioritise the protection of civilian populations and, lastly, exercise restraint in the future. All of these are elements to measure that are separate from the outcome we care about most, which is whether there is an observable decline in civilian casualties.

The example of IHL training raises a further challenging 'separation' with which evaluations of this type of intervention must contend, and that is the question of *when* the downstream protection outcomes should be measured. IHL trainings are rarely conducted on a regular schedule exactly three weeks before combat operations, allowing us to collect outcome measurements one month later. The intended restraint may not emerge until months or even years later. In this example, then, the first step towards establishing impact may be to examine the first stage: does the training lead to changes in reported intentions and understanding of IHL?

STRATEGIES TO CONSIDER

Breaking up evaluations into different stages. It may be useful to first test and establish whether an intervention such as a training or developing an action plan has an impact on reported attitudes and intended behaviour before evaluating the second stage of whether this in turn leads to changed behaviour (and reduced violence).

Considering lab-in-field experiments as a first step. Evaluations of measures designed to promote restraint might first be piloted in a lab-in-field environment before being tested in a conflict setting.

B. RESPONSIBLE RANDOMISATION DESIGNS FOR HUMANITARIAN SETTINGS

Randomisation itself can also pose challenges, particularly when it may seem to be at odds with humanitarian principles such as impartiality, which dictates that treatment must be provided purely based on need.

Several different options for randomising responsibly exist: the suitability of the following solutions is in part a function of how the assistance being provided can be structured and what the most salient learning goals for the evaluation are.

FIGURE 1. A/B TESTING

Treatment group A   ◉ Cash transfer   Treatment group B   🌾 In-kind transfer
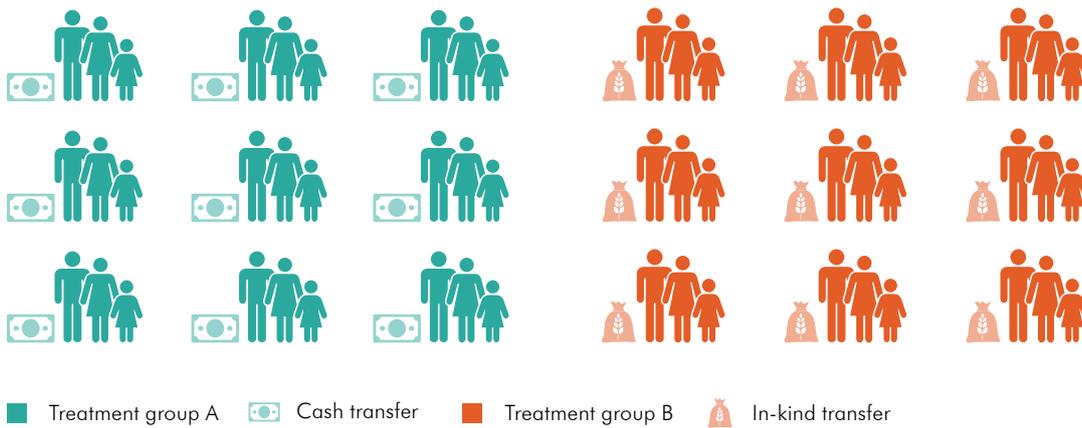
A/B testing

Description: A/B testing allows us to relatively nimbly (and quickly) evaluate the difference between two or more approaches, without creating a pure control group (i.e. a comparison group that receives nothing).

Use: This can be an effective way of comparing the effect of two or more delivery mechanisms or models of treatment when we need answers quickly to inform operational decisions while ensuring that all potential beneficiaries receive some assistance. In addition, this form of evaluation can be particularly useful for piloting and validating new models of delivery or treatment before scaling them. As many humanitarian projects are based on standard models of intervention, comparing adaptations or alternatives to the standard model may be more policy relevant than comparing the effect of an innovative intervention against no assistance. In some situations, and building on the necessary preparations, this study design can generate evidence relatively quickly to inform operational decisions.

Limitations: Because there is no 'pure' control group, we do not estimate the impact of the treatment but rather the differential impact of one treatment method over the other.

Cross-cutting treatments

Description: An impact evaluation can be set up to include multiple treatment arms which each including different combinations of treatment components (e.g. cash only, cash plus training, cash plus training plus psychosocial support) to test which approach is most effective and to reveal potential trade-offs.

Use: These forms of evaluations are useful when comparing different bundles of complex interventions to determine what may be most effective. They can be particularly useful in rigorously testing how to generate the greatest cost-effectiveness—that is, to ensure you are generating rigorous impacts while also reducing costs.

Limitations: Without a pure control group, these evaluations will again yield estimates of the relative impact of adding or subtracting components—but this is often the most relevant question. With more treatment arms comes a need for larger samples.

FIGURE 2. CROSS-CUTTING TREATMENTS

FIGURE 3. ENCOURAGEMENT DESIGNS

Treatment ■   Control ■   Encouragement ⓘ
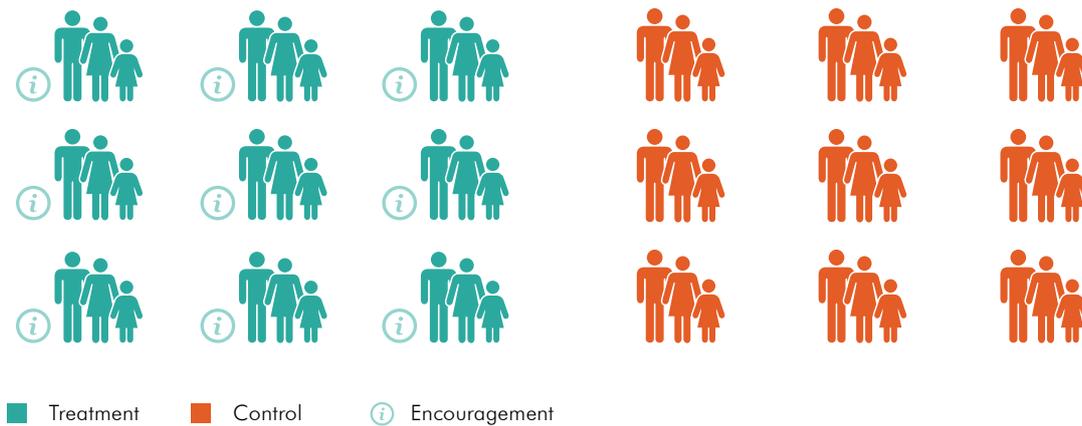
Encouragement designs

Description: Rather than randomising treatment or access to a programme, we can randomly assign some kind of encouragement or small incentive to enrol or take up treatment. This might be something simple like an SMS reminder or extra assistance in enrolling for a service. Effective encouragement leads to higher take-up of the programme in the treatment group, compared to the control group, from which the impact of the programme can be derived.

Use: This can be an appropriate design for evaluating programmes that do not lend themselves to randomisation (such as the effect of a right being recognised or not).

Limitations: These evaluations measure the impact of *receiving an encouragement* to take up a programme, from which one can derive the impact of the programme on those on which the encouragement was effective (the 'compliers'). There are two main caveats: compliers may be a specific subgroup such that the impact of the programme in the whole population may be different, and encouragement designs may suffer from a lack of statistical power unless the encouragement has a strong impact on programme take-up. Testing and refining the encouragement will therefore be key for this study design work.

Randomising 'in the bubble'

Description: Where a programme or intervention is administered conditional on an eligibility threshold (e.g. income, food insecurity), it is possible to randomise only among those who are just at the limit of eligibility.

Use: This approach to randomisation allows implementers to treat all those who very clearly meet the criteria without worrying about randomising treatment among them. Instead, they can essentially expand treatment to those who are either barely or almost eligible and randomise among this group to understand the impact of the treatment.

Limitations: These designs require a large sample size around the eligibility threshold to have a sufficiently powered study. As the sample only consists of participants with borderline eligibility, the impact we estimate is no longer representative of the full target population. We learn the impact of the programme on those who were borderline eligible, not on the group as a whole.

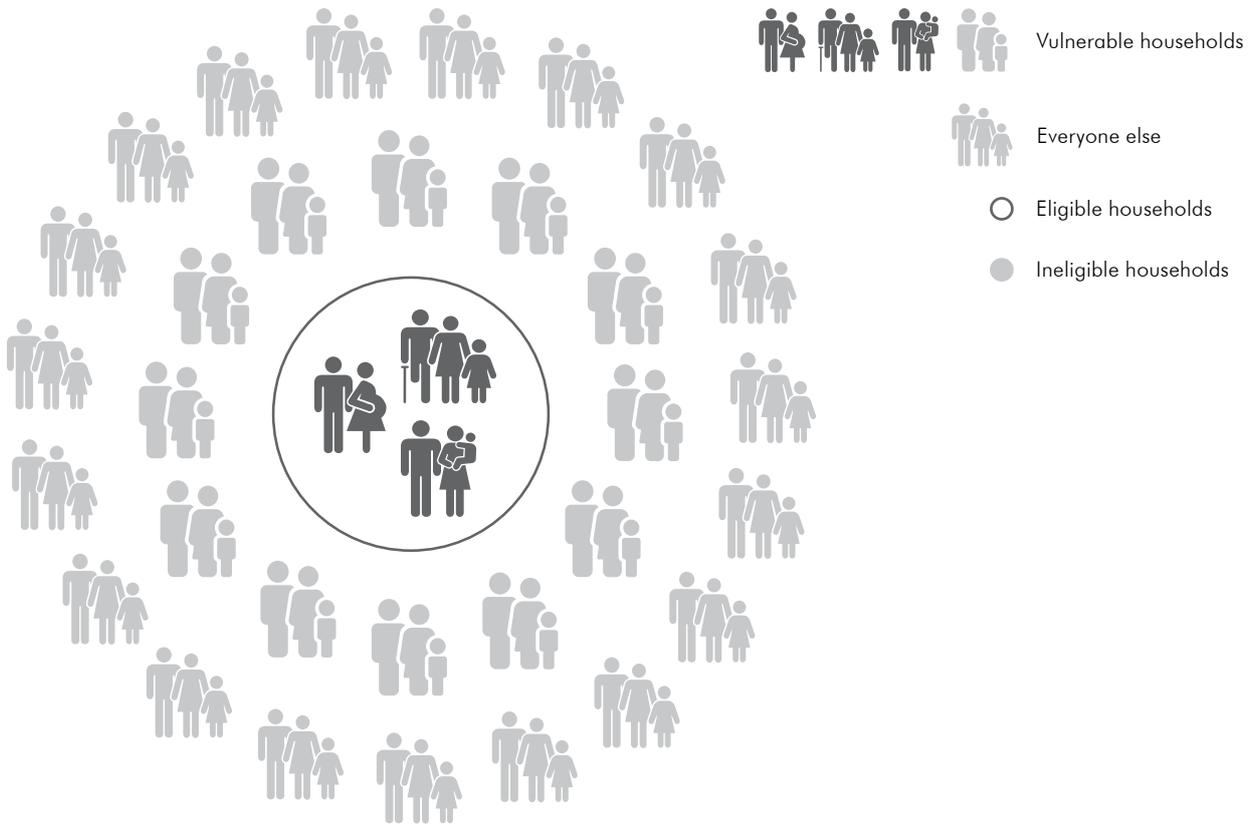**FIGURE 4A.** RANDOMISING IN A BUBBLE, INITIAL ELIGIBILITY CRITERIA (STRICT)

FIGURE 4B. RANDOMISING IN A BUBBLE, AFTER EXPANDING ELIGIBILITY CRITERIA

FIGURE 5. PHASE-IN RANDOMISATION

Phase 1
Study sample: 25% treatment, 75% control

Phase 2
Study sample: 50% treatment, 50% control

Phase 3
Study sample: 75% treatment, 25% control

Phase 4
Study sample: 100% treatment

■ Treatment ■ Control ⊡ Cash assistance

Phase-in randomisation

Description: This method of randomisation entails generating waves of consecutive treatment spread over planned intervals and comparing outcomes between groups who were administered the treatment at different points in time.

Use: This method is most useful when you cannot withhold treatment from any part of your population, but it is not possible to roll the programme out to everyone at the same time. Randomising rollout will, in this case, make it possible to compare between groups who already participated in the programme and those still on the waitlist and may in some cases also be a fair way to allocate assistance given the capacity constraints to reach everyone at the same time.

Limitations: Phase-in designs only allow you to measure impact over a short time period (until the last group participates in the programme). Anticipation of treatment among those who are in later waves may affect behaviour and make it more difficult to measure the true impact.

REFUGEES CONFIRMING THEIR IDENTITY TO COLLECT FOOD VOUCHERS IN CHAD. PHOTO: DOMINIQUE CATTON, © 2018 EUROPEAN UNION, CC BY-NC-ND 2.0

IV. WHAT DIRECTIONS CAN WE TAKE FROM EXISTING RESEARCH?

Answering many of the questions outlined above in Chapter III will require randomised evaluations to break new ground. But drawing on what research has been conducted in humanitarian contexts as well as existing research in several adjacent fields can provide both valuable insights for designing new programming in this area and models for designing robust, new evaluations.

We organise the insights from existing research around two broad categories: first, those that are relevant to improving the delivery of humanitarian assistance more generally, and second, those that may be of particular relevance for designing and evaluating protection-related programming.

A. DESIGNING MORE EFFECTIVE HUMANITARIAN PROGRAMMES

There are many ways to approach the question of improving the effectiveness of humanitarian programming. Among the many elements that practitioners raised in this regard, some are the need to consider the targeting of interventions, their timing and intensity, what combinations of bundled interventions are most effective, how to deliver more cost-effectively and how to deliver short-term assistance with an eye to long-term outcomes. A comprehensive review of existing evidence and approaches to these questions would be beyond the scope of this report, but here we sketch out some of the ways they have been approached in recent studies.

Targeting

In many crisis and conflict-affected settings, census and administrative data may be limited, quickly outdated, or unavailable and may thus provide little support for targeting assistance. How best to target this assistance remains an important question, both to produce fewer errors (by including too many or too few) and to find cost- (and time-) effective ways to target accurately.

Recent research from targeting strategies of social assistance programmes in other contexts has shown that community-based targeting may lead to better beneficiary identification than traditional targeting methods such as proxy-means testing (PMT).[4] PMTs are viewed as useful in capturing hidden income and as a way of avoiding including potential beneficiaries who would normally not qualify for the programme. One danger in fragile contexts is that PMTs often do not capture recent shocks to household consumption. Community methods, in which local leaders or a community council make decisions about eligible beneficiaries, may better capture these shifts. In a large evaluation of a social protection programme in Indonesia where researchers varied the targeting method used at the subvillage level (between PMT, a community-based approach or a hybrid), Alatas et al. (2012)*[5] found that while the PMT outperformed the hybrid and community-based approaches in correctly identifying beneficiaries, the difference was small and the community-based approach better reflected the community's own definition of vulnerability. The authors argued that the community's preferences 'appear to be informed by a better understanding of factors that affect a household's earning potential or vulnerability, such as the returns to scale within the family, as compared to relying purely on consumption'. This may be something for humanitarian actors to take into account in targeting decisions.

Community targeting, which relies in different ways on community inputs for identifying beneficiaries, may carry some risks related to local elite capture for certain types of programmes. There is some evidence to suggest that local elite capture may be less of a risk for interventions targeting the most vulnerable segments of the population. For example, Alatas et al. (2019)* varied the role that village leaders played in identifying beneficiaries for a conditional cash transfer scheme that targeted some of Indonesia's most vulnerable families and compared these to estimates of elite capture from other social programmes. They found no evidence for elite capture in the conditional cash transfer programme and estimated that capture in other social protection programmes was low. They also argued that 'this type of elite capture is not economically large, and in fact is small relative to the targeting error resulting from limited administrative capabilities'.

While these results may not be directly transposable to a humanitarian context, the studies give a sense for how the trade-offs between different targeting strategies can be framed and compared to inform and improve targeting mechanisms that minimise elite capture as well as administrative targeting errors.

> ### AVENUES FOR FUTURE RESEARCH
>
> **Comparing different targeting mechanisms.** What targeting approaches best identify those most in need of assistance, accounting for local perceptions of vulnerability? In situations where social networks may have been disrupted, a larger share of society is in need, and administrative capabilities are particularly stretched, do the advantages of community targeting uphold? What targeting strategy helps best prevent elite capture and reach those most in need?
>
> **Beneficiary targeting.** How can technology, including mobile phone data and machine learning, be leveraged to improve beneficiary targeting (see, e.g., Aiken et al. 2021)?
>
> **Markers of greatest need and/or vulnerability.** Which observable characteristics are effective in helping to identify those to whom assistance should be allocated?

---

[4] Proxy means testing is a way of estimating income based on reported information about different assets held by a household (such as a fridge or television).

[5] Asterisks following a citation indicate that the study presents the findings of a randomised evaluation.

Determining modalities: The example of cash and voucher assistance

Randomised evaluations have played an important role in strengthening our general understanding of the impacts of cash and voucher assistance. Importantly, they have helped buttress a growing movement to benchmark other forms of assistance against cash, to ask whether desired impacts cannot be delivered more cost-effectively simply through providing cash (Doocy and Tappis 2017).

The choice of how direct assistance, such as cash or voucher transfers, is delivered has important implications for its impact. Cash provides beneficiaries with a broader range of choices and can have a wider range of impact on household-level outcomes beyond food security (e.g. education, health) thanks to its flexibility. In a study in the Democratic Republic of Congo, Aker (2017)* compared the effectiveness of cash transfers against food vouchers in promoting food security in an internally displaced persons (IDP) camp; cash transfers required setting up a bank account at a bank in a village at some distance (a three-hour round trip), whereas the vouchers could be spent at a local fair. While both cash and vouchers increased food security and asset ownership, cash transfer households purchased a more diverse bundle of food and nonfood items.

In Niger, Aker et al. (2016)* examined the relative effectiveness of two forms of distributing an unconditional transfer to drought-affected households during the five-month 'hungry season' before the annual harvest. They randomly varied whether households received an electronic transfer by mobile phone or by cash handed out in person. The mobile transfers led households to adopt more diverse diets, with children eating more, and may have improved women's bargaining power in the household as the flexibility of mobile transfers alleviated certain logistical concerns and freed up time.

Other studies have compared the effect of cash and voucher assistance to food assistance of equal value for the recipients and found that households' investment decisions varied significantly, depending on which type of assistance they received (Schwab 2020)*, or compared the effect of livelihood trainings or an integrated nutrition programme to cost-equivalent cash transfers (McIntosh and Zeitlin 2020*, 2018*). While only a part of these evaluations has occurred in humanitarian situations, they illustrate the range of questions and learnings that can be derived from studies that rigorously compare different assistance delivery modalities.

Increasing cost-effectiveness

In a resource-constrained environment, assessing interventions' cost-effectiveness can provide valuable insights on the drivers of cost and impact for a better understanding of implementation approaches that offer 'the greatest value for money' (Dhaliwal et al. 2013).

For instance, in evaluating an economic support programme for internally displaced households in the Democratic Republic of Congo, Aker (2017)* determined that cash transfers were more cost-effective than equivalent-valued voucher transfers, for comparable levels of impacts on food consumption and household well-being. In Yemen, in the context of a cash and in-kind transfers programme for vulnerable households in rural communities, Schwab (2020)* found that while both modalities improved security among beneficiary households, cash transfers were cheaper to deliver than food.

Insights from cost-effectiveness analyses are dependent on the programme's objectives (i.e. the outcomes of interest to programme implementers) (Doocy and Tappis 2017). For instance, the cost-effectiveness analysis of a cash, voucher and in-kind assistance programme in northern Ecuador suggests that there was no difference between cash and voucher transfers in improving food consumption levels; however, voucher transfers were the most cost-effective modality for improving dietary diversity (Hidrobo et al. 2014)*.

Cost-effectiveness of bundled interventions. The relative cost-effectiveness of interventions aiming at delivering on the same outcome will in many cases be salient. Building cost-effectiveness analysis into evaluations of projects that deliver packages of complementary activities may be particularly relevant as this may help isolate the aspects that drive impact the most.

Timing: Anticipatory action

Anticipatory action is an emerging area of both practice and evaluation; until recently, most evaluations of this form of assistance were relatively simple ex post evaluations (WFP 2019). A recent quasi-experimental evaluation by Pople et al. (2021) of a single anticipatory cash transfer to households forecasted to experience extreme floods in Bangladesh found that it had a significant, positive impact on household welfare three months after the flood, increasing both adult and child food consumption and reducing asset loss.

Anticipatory vs. responsive action. Is an intervention before a shock occurs more effective in supporting households to mitigate the worst effects of an extreme drought than an intervention after the shock, whether this be in terms of food security, asset loss, well-being or levels of SGBV?

Refining anticipatory action. This new logic of structuring humanitarian action comes with a host of new questions: At what distance to the projected shock can households maximise the effect of the assistance? What combination of cash, in-kind support or information dissemination is most effective? What targeting criteria help identify and effectively support those who stand to suffer most from the forecasted crisis?

Intensity and dosage

Impact evaluations are well positioned to help us answer questions related to the relationship between the intensity (sometimes referred to as 'dosage') at which an intervention is delivered and the magnitude of its impacts (Puri et al. 2015). This carries key implications for assessing a programme's cost-effectiveness and scalability.

In the evaluation of a dispute resolution training intervention targeting communities recovering from conflict in Liberia, Blattman, Hartman and Blair (2014)* randomly varied the intensity of the programme in some communities (by increasing the enrolment threshold among the adult population) and found no significant incremental treatment effect on the programme's main areas of impact (unresolved land conflicts and property destruction).

As humanitarian interventions rely more and more on cash and in-kind delivery modalities, understanding the general equilibrium effects of large economic transfers is increasingly important. Rigorous evaluations can help answer such questions, even when the intensity/dosage of an intervention is not part of the randomisation process. In the randomised evaluation of a government in-kind transfers programme in Mexico, Cunha, De Giorgi and Jayachandran (2011)* determined the incremental treatment effect on local market prices by constructing measures of treatment intensity based on the size of the transfers. Using a quasi-experimental design to measure the impact of an emergency cash transfer programme in a refugee camp in Turkey, Özler et al. (2020) found that the programme's impact on household food consumption, relative to the size of the transfer, was diluted by the redistribution of the transfers among members of nontreated households (movement of primarily school-age children from larger ineligible households to smaller eligible ones). Such findings may point towards the need of recalibrating the size of economic support programmes to account for the potential spillover effects and ensure that the households found to be most vulnerable do not carry an additional burden as a result of the programme.

Comparing treatment intensities. What share of a community should be targeted by an intervention— be this through distribution of assistance, training or dissemination of information—to maximise the programme's impact? Could supporting a wider range of households in a situation with high levels of need be in the interest of the most vulnerable households?

Exploring the longer-term effects of short-term programming

When properly designed and implemented, randomised evaluations can provide insights on a programme's long-term impact, even beyond the project's life cycle, as randomisation ensures that treatment and comparison groups are statistically comparable before the evaluated intervention, and differences in outcomes post-intervention are attributable to the programme. Even a phase-in randomisation design, which may be preferred in a humanitarian context but is less conducive to capturing long-run impact, can allow the detection of long-term effects under certain assumptions (Bouguen et al. 2019).

For instance, in the Democratic Republic of Congo, Mvukiyehe and Van der Windt (2020)* measured the long-term effects of a community-driven development programme three years after its conclusion, leveraging the same randomisation design employed for a preceding evaluation of the same intervention (Humphreys, Sanchez de la Sierra and Van der Windt 2019)*. In Malawi, Baird et al. (2019)* evaluated the long-term effects of a cash transfer programme targeting adolescent girls more than two years after the programme's activities had ceased, leveraging the same randomisation design used to measure the programme's effectiveness in the short term (Baird, McIntosh and Özler 2011)*.

On a shorter timeline, Caria et al. (2020)* found no measurable effect of a small, one-off labelled cash transfer to support the job searches of refugees in Jordan six weeks after the transfer, but they identified a significant impact on employment and earnings at the two- and four-month mark. Together, these studies show that evaluating programmes beyond the closure of activities is possible and, in some cases, important to not miss the outcomes of interest.

AVENUES FOR FUTURE RESEARCH

Measuring longer-term effects to inform short-term programming. How long do the effects measured at the project closure persist? Is there a lag until the effects materialise? Or, on the contrary, do the effects decay as soon as an organisation's support or presence ends?

B. IMPROVING PROTECTION OUTCOMES

Given the very limited rigorous evaluation of protection-related humanitarian programming, we draw more heavily on research in adjacent fields to highlight insights from rigorous evaluations of other forms of programming that either touch on important elements of protection or provide models for thinking about how we might achieve protection outcomes. These include research into how to diffuse information, norms and awareness of rights that may contribute to protection; delivering remedial support, such as mental health interventions, in resource-constrained settings; and curbing violence and promoting restraint.

Diffusing information and promoting claiming of rights

One important element of protection programming involves efforts to raise awareness about their rights among people at risk of harm. This includes spreading information about rights as well as the services intended to support their protection, that is, services intended to reduce the barriers that keep them from enjoying their rights. Here, recent literature on how to most effectively diffuse information and awareness (e.g. health-promoting behaviours) through different communities, or to promote positive norms, offers insight.

Information spreads in different ways through communities depending in part on who is selected as messenger, and this can have important implications for programming that seeks to diffuse vital information to vulnerable populations. Banerjee et al. (2019)* examined whether seeding information on an immunisation campaign in the Indian state of Haryana through individuals nominated by their fellow villagers as good at transmitting information ('gossips', in the researchers' shorthand) is more effective than randomly selecting messengers. They find large impacts: villages where messengers were nominated by the community see 22 percent higher vaccination rates than those where the messengers were selected randomly.

In the Haryana study, 'gossips' were revealed to be effective at spreading messages regardless of whether they were respected or not. But there is also evidence that influential and respected figures may be able to spread information that leads to behaviour change more effectively. In a large-scale randomised evaluation that involved Covid-19 messaging sent to 25 million individuals in West Bengal, India, researchers found that an SMS with a link to a 2.5-minute video of Abhijit Banerjee (the Nobel laureate researcher, well known in his native state) promoting just one health-preserving behaviour (distancing or hygiene; this varied between recipients) led to a doubling in reporting of symptoms to local health workers and a large reduction in how much recipients travelled outside their village, and also led to significant spillover effects in behaviour among those not directly receiving the messages (Banerjee et al. 2020)*.

It is unclear how well this evidence will transport to humanitarian settings, where local network dynamics may be less predictable due to the volatility and mobility introduced by crises. Given the link between the structure of social networks and individual-level protection outcomes, understanding how social networks function, and how these dynamics can be affected by the presence of humanitarian operations, is of key importance, particularly in settings where the size, composition and sources of social networks can be very fluid (Cachia and Ramos 2020).

For instance, there exists some evidence that relying upon social networks can risk excluding marginalised members of the community. In a randomised evaluation of whether well-networked farmers in Mali were better at disseminating information about how to compost effectively, researchers found that female farmers were less likely to receive training from these farmers even though they had just as much contact with them as men, on average (Beaman and Dillon 2018)*. These results suggest that while disseminating information through socially influential people may diffuse information cost-effectively, there is a risk that this comes at the expense of reaching those who are less well networked.

Remedying harm: Delivering mental health and psychosocial support in resource-constrained settings

Another core form of protection interventions are efforts to remedy harm. One mechanism for achieving this is through mental health and psychosocial support interventions (MHPSS). A large body of public health literature has evaluated the impact of MHPSS and has considered the question of how to provide such services outside of clinics and in resource-constrained settings, including by drawing on the capacity of local organisations to adapt interventions, and by training laypersons to deliver simple care for less severe conditions, thus allowing trained professionals to focus on patients in need of more specialised care—the so-called task shifting approach.

Much of the literature has focused only on short-term effects—a 2019 systematic review of MHPSS programmes in humanitarian emergencies found that only one of 35 studies included in the review assessed impacts after one year (Bangpan, Felix and Dickson 2019).However, with this caveat, these studies have shown some evidence of effectiveness in reducing the strain of post-traumatic stress disorder (PTSD) and improving functioning. Bangpan, Felix and Dickson (2019) highlighted the need for evaluating the effect of this programming on broader outcomes beyond mental health to include resilience and well-being, arguing that these insights would help better anchor conversations around the cost-effectiveness of such programming.

Two randomised evaluations in this area have specifically focused on including local community organisations in the delivery and adaptation of MHPSS interventions, although it is worth noting that both worked with small samples. O'Callaghan et al. (2013)* evaluated a culturally adapted trauma-focused cognitive behavioural therapy programme that worked with young girls who had been witnesses or survivors of sexual abuse in the Democratic Republic of Congo. While the evaluation did not explore the question of how this adaptation of the programming may have improved its impacts, it did find reduced trauma-related symptoms and increased prosocial behaviours among participants three months later.[6] In a separate study in northern Uganda, Ertl et al. (2011)* evaluated a short-term community-based programme that worked with formerly abducted youth through individual narrative exposure therapy to target symptoms of PTSD. One year after the intervention, the severity of PTSD symptoms was lower among participants.

Randomised evaluations can also highlight the potential psychological costs of programmes that aim at improving the mental health of victims of conflict. In a randomised evaluation of a low-cost postconflict reconciliation intervention in rural villages of Sierra Leone, Cilliers, Dube and Siddiqi (2016)* found that while the programme increased social capital in treated villages and led to more forgiveness of perpetrators, these positive outcomes came at the cost of individual well-being, with increased anxiety, depression and PTSD. These impacts arose in villages with limited or no existing psychosocial support services, and the authors noted that a more effective strategy could be to bundle the reconciliation programme with sustained counselling services.

AVENUES FOR FUTURE RESEARCH

While focused on MHPSS interventions and outcomes, the underlying questions about longer-term impacts, tiered intervention models and bundled interventions are also relevant in considering the effectiveness of legal assistance, mediation or other types of interventions to remedy harm arising from violence, coercion and deliberate deprivation:

Capturing longer-term impacts. Recent research has shown the benefits of investing in methods and resources to track study participants over several years and analyse the long-term effects of intervention (Mvukiyehe and Van der Windt 2020*; Baird, McIntosh and Berk Özler 2019*; Millán et al. 2019). Such approaches could be useful to capture the longer-term impact of MHPSS interventions and other activities responding to the harm people suffer in conflicts.

Refining tiered intervention models. Building on the concept of task shifting, what tiers of support are possible and effective? What is the absolute minimum of support to which every conflict-affected person should have access? What sort of outreach, case management and targeting criteria are best suited to maximise outcomes within resource-constrained settings?

Identifying purposeful combinations of different types of assistance. Building on the effectiveness of bundled interventions in other settings (e.g. Blattman, Jamison and Sheridan 2017*), can cash and voucher assistance or skills training support healing in combination with an MHPSS programme? Are improvements in well-being more persistent when combined with other types of assistance? When such services are not available, can financial support prevent the development of more serious symptoms of anxiety and depression?

---

[6] Participants were randomly assigned to a treatment group (participating in the programme) and a wait-list comparison group that would benefit from the programme after the evaluation period.

Protection mainstreaming: The case of cash transfers and SGBV/Intimate Partner Violence (IPV)

As outlined under the heading of *Accountability and safeguarding* in Chapter III above, practitioners are interested in better understanding measures to prioritise the safety and dignity of the people they support and avoid causing harm. By way of example, there is a general agreement among practitioners on the potential for cash and voucher assistance in affecting multisectoral outcomes in humanitarian settings, including protection-related outcomes, making it one of the 'fastest growing evidence-based interventions in the humanitarian community' (Global Protection Cluster 2020). Poverty alleviation programmes, in the form of cash transfers, can indirectly address the issue of SGBV and IPV, but the evidence base on the impact of such interventions on the incidence of SGBV and IPV in humanitarian settings remains ambiguous (Cross, Manell and Megevand 2018).

An evaluation of a 30-month pilot unconditional cash transfer programme in the rural areas of Togo by Briaux et al. (2020)*, focused on improving children's health, nutrition and education in the first three years, uncovered significant and positive effects on women's experience of physical IPV. In Ecuador, an economic assistance programme providing either cash, in-kind or voucher transfers to Colombian refugees and vulnerable Ecuadorian households reduced the probability of experiencing controlling behaviours and/or sexual violence. Further analyses indicate that in-kind assistance reduced physical and sexual violence, while cash transfers reduced controlling behaviours; voucher assistance was effective in reducing both indicators (Hidrobo, Peterman and Heise 2016)*.

More rigorous evidence however is needed to better understand the causal link between economic support interventions in the form of transfers and protection-related outcomes such as gender-based violence (GBV) (Cross, Manell and Megevand 2018). For instance, there is evidence to suggest that individual-level characteristics (e.g. education levels) can determine the nature of the effect of cash transfer interventions on GBV and IPV outcomes. In Ecuador, an unconditional cash transfer programme targeting women reduced women's experience of emotional violence among those who had less schooling than their partner, while the programme had the opposite effect on women who had at least the same number of schooling years as their partner (Hidrobo and Fernald, 2013)*.

Reducing conflict, promoting restraint

Another core element of protection is preventing violence against protected populations—often concretely operationalised through efforts to reduce the risk that civilians are exposed to harm. This category of protection interventions may focus on promoting restraint among armed actors, and reducing the use of violence targeting civilian populations, or aim at curbing the prevalence of SGBV in crisis settings. This section draws on a rapidly expanding experimental literature around peacebuilding and conflict prevention programming and highlights in particular potentially relevant insights from media interventions (including 'edutainment') designed to reduce conflict; interventions to reinforce and leverage communities' own capabilities to prevent violence; livelihood programming that targets (former) combatants; and intergroup contact programmes that seek to reduce the salience of communal differences and build social cohesion.

Media interventions

There is some evidence that radio and television programming (including 'edutainment' interventions that seek to embed educational content within entertaining media formats like soap operas) can play a role in shifting social norms around violence, building social capital and even discouraging rebel recruitment. This is an encouraging area for more exploration; however, it is worth keeping in mind that the media's power to shift norms around violence can work in both directions—for example, retrospective quasi-experimental research has established the role played by radio broadcasts in Rwanda in driving more killing during the 1994 genocide (Yanagizawa-Drott 2014).

Perhaps most relevant for protection-related efforts, Armand, Atwell and Gomes (2020) established, using quasi-experimental methods, the impact of radio programming in Uganda that offered targeted defection messaging to members of the Lord's Resistance Army (LRA). The messaging increased defections, reduced LRA-related violence and reduced violence against both civilians and security forces.

Blair et al. (2021)* evaluated the impact of radio messaging promoting acceptance of returned Boko Haram fighters in Maiduguri, Nigeria. They randomly assigned listeners to a placebo message or one delivered by a religious leader promoting the value of forgiveness. This latter message increased support for the reintegration of former fighters and led listeners to report greater willingness to interact with such fighters.

In earlier research, including two randomised evaluations, radio programmes have generated mixed results in reducing conflict-related attitudes and behaviours. Paluck (2010)* evaluated the impact of accompanying a weekly radio soap opera in eastern Democratic Republic of Congo about a fictional conflict with a talk show designed to promote perspective-taking about intergroup conflict and cooperation. After one year, listeners of the talk show reported increased levels of group discussion but were also *less* tolerant of outsiders, as measured through behavioural games. Discussing group conflict may have made differing identities more salient. In another randomised evaluation, Paluck and Green (2009)* showed that a postgenocide radio soap opera in Rwanda that sought to promote individual thinking and collective action in problem-solving had little impact on listeners' attitudes towards other groups but *did* promote a willingness to express dissent and to resolve problems through active negotiation and discussion of sensitive subjects.

In another setting, researchers found that a video campaign aired during a local film festival in rural Uganda succeeded in shifting norms around reporting violence against women and even reduced actual incidence in treatment communities even as it had no impact on community members' attitudes generally towards violence against women (Green, Cooper and Wilke 2020)*. In this case, the researchers argued that the public community viewing of these videos may have been an important component of shifting norms.

Promoting community capacities for dispute resolution

Another area of interest to practitioners are interventions to increase communities' capacity to ensure their own protection in situations of heightened violence. One way to build bulwarks against violence and to strengthen communities' self-protection capacities may be to build support for local-level dispute resolution. Blattman, Hartman and Blair (2014)* evaluated a programme in Liberia that provided alternative dispute resolution training to local communities with the aim of helping them settle land disputes, an important source of local-level conflict and potential renewed instability following the end of the civil war. One year after the training, there was less violence in treated communities and land disputes were resolved at a faster rate; these effects persisted after three years (Hartman, Blair and Blattman 2020)*. The intervention did, however, lead to an increase in extrajudicial punishment (including witch trials and trials by ordeal), highlighting an alarming side effect that may have arisen from the focus on informal resolution.

Promoting alternative livelihoods

Another potential mechanism for reducing violence may be to provide economic opportunities that offer an attractive alternative to violence, dissuading some individuals from taking up arms and leading others to exit armed violence. Research has shown that targeted economic assistance can shape support for rebel forces in conflict-affected settings, but we do not yet know enough about the exact mechanisms behind this impact.

Blattman and Annan (2016)* evaluated the impact of an intensive agricultural training programme, paired with agricultural supplies and psychosocial counselling, for Liberian ex-combatants. The training led to increased employment in agriculture and average wealth among participants and reduced their time spent on illicit activities, although many did not completely give these up. When conflict broke out in neighbouring Côte d'Ivoire, programme participants were less likely to take part.

In Kandahar, Afghanistan, a randomised evaluation of a vocational training programme for at-risk youth had no impact on the participants' economic livelihoods (Lyall, Zhou and Imai 2020)*. A one-time cash transfer (~USD $75) evaluated in the same study initially led to increased support for the government but this impact dissipated over time and reversed into increased support for the Taliban. However, for individuals who received both the training and the cash transfer, the treatment marginally increased support for the government and reduced support for the Taliban. The study highlights dynamics between humanitarian assistance and conflict that seem important to better understand for informed decision-making in the sector.

Building intergroup contact to reduce potential for conflict

A growing body of research examines whether bringing groups from opposing sides of a divide into constructive forms of contact can be useful in shifting the perspectives and attitudes they hold about one another and potentially reduce violence and support collaboration. Mousa (2020)* evaluated the effect of generating mixed soccer leagues between Muslim and Christian youth in the ISIS-affected cities of Erbil and Qaraqosh in Iraq. She found that Christians who were randomly assigned to play with Muslim teammates were more likely to engage with them after the soccer league had ended, although this did not translate into improved levels of tolerance towards Muslim strangers.

Alan et al. (2021)* examined whether a pedagogical intervention in Turkish schools that promotes perspective-taking among students can help promote integration between Turkish-born and Syrian refugee students and reduce classroom violence. The intervention was shown to not only reduce ethnic segregation and improve trust levels among students but also measurably reduced within-school peer violence. The intervention also led to an improvement in the Turkish language skills of refugee children even though it did not have a focus on strengthening academic skills.

CHILDREN PLAYING IN A TEMPORARY LEARNING FACILITY IN THE HAJJ ALI REFUGEE CAMP IN IRAQ. PHOTO: PETER BIRO, © 2017 EUROPEAN UNION, CC BY-NC-ND 2.0

V. CONCLUSION: WHAT DO WE HOPE TO HAVE LEARNED IN THE NEXT FIVE YEARS?

This learning agenda has sought to provide an overview of what demand exists for more rigorous evaluation around humanitarian assistance, with a particular focus on humanitarian protection programming, and to identify where and how randomised evaluations will be well placed to meet this demand.

As discussed in Chapters II and III, our consultations have suggested that the primary obstacle has not been a lack of interest in deploying rigorous evaluations in this sector. Instead, two important challenges that have generated high 'barriers to entry' in this field have been 1) the difficulty of identifying research questions and feasible experimental designs that will support practitioners' learning aims and 2) a need for flexible resources that allow organisations to incorporate (time and labour-intensive) evaluations into projects that are often short term and short-staffed. Neither of these challenges are specific to the methodology but may be particularly salient for this type of research.

We believe that research initiatives can help alleviate the challenges of producing rigorous evaluation in this field in a few different ways. First, by providing dedicated funding to launch evaluations that will not come at the expense of programming budgets. Because well-designed evaluations take time to launch and programmes are often launched quickly, multiyear research funding pools will allow humanitarian organisations to plan how to build evaluation into future assistance efforts and develop required staff capacity. Second, by creating the flexibility for researchers and practitioners to work together to design evaluations that answer the most relevant questions. When evaluation funding is tied to specific interventions, it can be difficult to produce studies that may address the broadest questions. Third, evaluation funding should be targeted to research that will yield insights on a broad range of programming and not just the specific project evaluated. These are important elements of research initiatives that seek to produce evidence as a public good that can shift thinking and programme design across the sector.

One important outcome of investing more in rigorous evaluation in the sector is bringing attention and academic expertise to the question of how we measure the success of humanitarian programming. This is a question that touches on the indicators we use to track this success and how we understand the theories of change behind this programming. For organisations that explore the idea of a randomised evaluation of some of their programming but ultimately decide not to proceed, there is still value in having considered how to measure impacts and outcomes more carefully as well as how to draw on existing research. This is perhaps particularly true in the area of humanitarian protection, where the theoretical links between outputs and outcomes may be less clear.

Where do we expect to learn the most in the next five years from randomised evaluations in the humanitarian sector? This is a huge field, in which the number of experimental evaluations is already growing. It would be difficult to single out likely progress on particular sectoral areas of programming. But more broadly, we expect that randomised evaluations will play a role in advancing our thinking on the following issues, which often cut across sectors:

- Evaluating the persistence of key outcomes and the durability of impacts of humanitarian programming beyond a six-month time frame. This will involve, in many cases, evaluating outcomes beyond the humanitarian programming cycle. But insights into the longer-term impacts of interventions such as emergency livelihoods assistance or nutrition support can inform better humanitarian and development programming alike.

- Generating evidence on when and how it is best to provide humanitarian relief, including on whether anticipatory action and other forecast-based assistance mechanisms may be more effective.

- Identifying more optimal measures for targeting assistance and programming.

- Better understanding how to promote resilience in long-standing emergency contexts or in contexts where humanitarian operations are ending and families may need help in adapting to an end to assistance.

- Exploring how social networks either transform or develop in humanitarian settings, including how they respond to interventions of many types.

- Identifying complementarities between different interventions and assessing which elements are critical in integrated interventions (by 'unbundling' impacts), including the role of cash assistance in either extending or deepening the impacts of other interventions.

- Continuing to expand research with mobile populations, drawing on technological advances that have made it easier to deliver a range of forms of assistance as well as to track outcomes over time.

With specific reference to humanitarian protection, where there has been very little experimental evaluation to date, we think the following are promising areas for new contributions to the field, with more possibilities included in Chapter IV:

- Capturing how to support the spread of information among protected populations, including about how to exercise and access rights.

- Identifying measures that improve mental health and psychosocial support for populations in conflict and displacement settings.

- Exploring how to shift norms around violence, including sexual violence, in conflict and displacement settings.

- Evaluating the effectiveness of interventions designed to support and strengthen communities' capacity to assess and manage protection risks.

- Promoting alternative livelihoods for ex-combatants or potential combatants as a pathway to reducing violence.

- Understanding the role that host communities and local actors play in shaping the effectiveness of protection of refugees and displaced persons, and managing potential negative reactions from these communities.

Launching more rigorous evaluations conceived expressly to capture the impact of protection programming will shed more light on the mechanisms that lead to protection outcomes, an understudied topic. These findings will guide action more broadly in conflict and post-conflict settings, by helping practitioners unpack the mechanisms through which we might promote restraint, generate alternative livelihoods to demobilise armed actors, effectively remedy harm, and promote meaningful resilience among affected populations. With the necessary resources, we believe this evidence can improve efforts to reduce violence that otherwise threatens the lives, livelihoods and well-being of crisis and conflict-affected populations. An investment in more rigorous evaluation of protection programming should thus contribute to strengthening our understanding of how to guide both humanitarian action and measures designed to manage and prevent conflict and improve the long-term outcomes of people affected by crisis and conflict.

SOURCES CITED

Abdul Latif Jameel Poverty Action Lab (J-PAL). 2021. "Ethical Conduct of Randomized Evaluations." Research Resources. Accessed 30 March 2022. https://www.povertyactionlab.org/resource/ethical-conduct-randomized-evaluations#section-ethical-principles-and-the-belmont-report.

Aiken, Emily, Suzanne Bellue, Dean Karlan, Christopher R. Udry, and Joshua Blumenstock. "Machine Learning and Mobile Phone Data Can Improve the Targeting of Humanitarian Assistance." NBER Working Paper #29070, July 2021. https://doi.org/10.3386/w29070.

Aker, Jenny C. 2017. "Comparing Cash and Voucher Transfers in a Humanitarian Context: Evidence from the Democratic Republic of Congo." *The World Bank Economic Review* 31 (1): 44–70. https://doi.org/10.1093/wber/lhv055.

Aker, Jenny C., Rachid Boumnijel, Amanda McClelland, and Niall Tierney. 2016. "Payment Mechanisms and Antipoverty Programs: Evidence from a Mobile Money Cash Transfer Experiment in Niger." *Economic Development and Cultural Change* 65 (1): 1–37. https://doi.org/10.1086/687578.

Armand, Alex, Paul Atwell, and Joseph F. Gomes. 2020. "The Reach of Radio: Ending Civil Conflict through Rebel Demobilization." *American Economic Review* 110 (5): 1395–429. https://doi.org/10.1257/aer.20181135.

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay. 2021. "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking." *The Quarterly Journal of Economics* 136 (4): 2147–94. https://doi.org/10.1093/qje/qjab009.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review* 102 (4): 1206–40. https://doi.org/10.1257/aer.102.4.1206.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, Ririn Purnamasari, and Matthew Wai-Poi. 2019. "Does Elite Capture Matter? Local Elites and Targeted Welfare Programs in Indonesia." *AEA Papers and Proceedings* 109: 334–39. https://doi.org/10.1257/pandp.20191047.

Baird, Sarah, Craig McIntosh, and Berk Özler. 2011. "Cash or Condition? Evidence from a Cash Transfer Experiment." *The Quarterly Journal of Economics* 126 (4): 1709–53. https://doi.org/10.1093/qje/qjr032.

Baird, Sarah, Craig McIntosh, and Berk Özler. 2019. "When the Money Runs Out: Do Cash Transfers Have Sustained Effects on Human Capital Accumulation?" *Journal of Development Economics* 140: 169–85. https://doi.org/10.1016/j.jdeveco.2019.04.004.

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. 2019. "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials." *The Review of Economic Studies* 86 (6): 2453–90. https://doi.org/10.1093/restud/rdz008.

Banerjee, Abhijit, Marcella Alsan, Emily Breza, Arun G. Chandrasekhar, Abhijit Chowdhury, Esther Duflo, Paul Goldsmith-Pinkham, and Benjamin A. Olken. 2020. "Messages on COVID-19 Prevention in India Increased Symptoms Reporting and Adherence to Preventive Behaviors among 25 Million Recipients with Similar Effects on Non-Recipient Members of Their Communities." NBER Working Paper #27496, July 2020. https://doi.org/10.3386/w27496.

Bangpan, Mukdarut, Lambert Felix, and Kelly Dickson. 2019. "Mental Health and Psychosocial Support Programmes for Adults in Humanitarian Emergencies: A Systematic Review and Meta-Analysis in Low and Middle-income Countries." *BMJ global health* 4 (5): 1–11. https://doi.org/10.1136/bmjgh-2019-001484.

Barbelet, Veronique, Gemma Davies, Josie Flint, and Eleanor Davey. 2021. Interrogating the Evidence Base on Humanitarian Localisation: A Literature Study. *HPG Literature Review*. London: ODI, 1–85. https://odi.org/en/publications/interrogating-the-evidence-base-on-humanitarian-localisation-a-literature-study/.

Beaman, Lori, and Andrew Dillon. 2018. "Diffusion of Agricultural Information Within Social Networks: Evidence on Gender Inequalities from Mali." *Journal of Development Economics* 133: 147–161. https://doi.org/10.1016/j.jdeveco.2018.01.009.

Blair, Graeme, Rebecca Littman, Elizabeth R. Nugent, Rebecca Wolfe, Mohammed Bukar, Benjamin Crisman, Anthony Etim, Chad Hazlett, and Jiyoung Kim. 2021. "Trusted Authorities Can Change Minds and Shift Norms during Conflict." *Proceedings of the National Academy of Sciences* 118 (42): 1–5. https://doi.org/10.1073/pnas.2105570118.

Blattman, Christopher, Alexandra C. Hartman, and Robert A. Blair. 2014. "How to Promote Order and Property Rights under Weak Rule of Law? An Experiment in Changing Dispute Resolution Behavior through Community Education." *American Political Science Review* 108 (1): 100–120. https://doi.org/10.1017/S0003055413000543.

Blattman, Christopher, and Jeannie Annan. 2016. "Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State." *American Political Science Review* 110 (1): 1–17. https://doi.org/10.1017/S0003055415000520.

Blattman, Christopher Julian C. Jamison, and Margaret Sheridan. 2017. "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia." *The American Economic Review* 107 (4): 1165–206. https://doi.org/10.1257/aer.20150503.

Bloom, Louise, and Alexander Betts. "The Two Worlds of Humanitarian Innovation." RSC Working Paper #94, August 2013. https://www.rsc.ox.ac.uk/files/files-1/wp94-two-worlds-humanitarian-innovation-2013.pdf.

Bouguen, Adrien, Yue Huang, Michael Kremer, and Edward Miguel. 2019. "Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics." *Annual Review of Economics* 11: 523–61. https://doi.org/10.1146/annurev-economics-080218-030333.

Briaux, Justine, Yves Martin-Prevel, Sophie Carles, Sonia Fortin, Yves Kameli, Laura Adubra et al. 2020. "Evaluation of an Unconditional Cash Transfer Program Targeting Children's First 1,000 Days Linear Growth in Rural Togo: A Cluster-Randomized Controlled Trial." *PLoS Medicine* 17 (11): 1–29. https://doi.org/10.1371/journal.pmed.1003388.

Cachia, Romina, and Daniel Holgado Ramos. 2020. "Network Analysis as a Tool for Humanitarian Protection: Research and Practice." *Journal of International Humanitarian Action* 5 (1): 1–14 https://doi.org/10.1186/s41018-020-00071-7.

Caria, Stefano, Maximilian Kasy, Simon Quinn, Soha Shami, and Alex Teytelboym. 2020. "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan." CESifo Working Paper #8535, September 2020. https://www.cesifo.org/node/57859.

Chaaban, Jad, Nisreen Salti, Hala Ghattas, Wael Moussa, Alexandra Irani, Zeina Jamaluddine, and Rima Al-Mokdad. "Multi-Purpose Cash Assistance in Lebanon: Impact Evaluation on the Well-Being of Syrian Refugees." Paper submitted to ERF 26th Annual Conference, 2020. https://erf.org.eg/app/uploads/2020/09/1600007407_774_1507764_84_alexandrairani.pdf.

CHS Alliance, Group URD and Sphere Project. 2014. *Core Humanitarian Standard on Quality and Accountability.* Geneva: CHS Alliance, Group URD, and Sphere Project. https://corehumanitarianstandard.org/files/files/Core%20Humanitarian%20Standard%20-%20English.pdf.

Cilliers, Jacobus, Oeindrila Dube, and Bilal Siddiqi. 2016. "Reconciling after Civil Conflict Increases Social Capital but Decreases Individual Well-Being." *Science* 352 (6287): 787–94. https://doi.org/10.1126/science.aad9682.

Cross, Allyson, Tenzin Manell, and Melanie Megevand. 2018. "Humanitarian Cash Transfer Programming and Gender-Based Violence Outcomes: Evidence and Future Research Priorities." Collected Papers on Gender and Cash Transfer Programmes in Humanitarian Contexts. https://www.calpnetwork.org/wp-content/uploads/2020/03/genderandctpwrcirc-1.pdf.

Cunha, Jesse, Giacomo De Giorgi, and Seema Jayachandran. 2011. "The Price Effects of Cash Versus In-kind Transfers." NBER Working Paper #17456, September 2011. https://doi.org/10.3386/w17456.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2013. "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education." Edited by Paul Glewwe, University of Chicago Press. *Education Policy in Developing Countries* 17: 285–338. http://hdl.handle.net/1721.1/116111.

Doocy, Shannon, and Hannah Tappis. 2017. "Cash-Based Approaches in Humanitarian Emergencies: A Systematic Review." *Campbell Systematic Reviews* 13 (1): 1–200. https://doi.org/10.4073/csr.2017.17.

Ertl, Verena, Anett Pfeiffer, Elisabeth Schauer, Thomas Elbert, and Frank Neuner. 2011. "Community-Implemented Trauma Therapy for Former Child Soldiers in Northern Uganda: A Randomized Controlled Trial." *The Journal of the American Medical Association* 306 (5): 503–12. https://doi.org/10.1001/jama.2011.1060.

Global Protection Cluster. 2020. *Cash and Voucher Assistance for Protection: Taking Stock of Cash and Voucher Assistance to Achieve Protection Outcomes in the Protection Sector in Humanitarian Settings.* Geneva: Global Protection Cluster https://www.globalprotectioncluster.org/wp-content/uploads/GPC_Stock-Taking_2020_ENGLISH.pdf.

Green, Donald P., Anna Wilke, and Jasper Cooper. 2020. "Countering Violence against Women by Encouraging Disclosure: A Mass Media Experiment in Rural Uganda," *Comparative Political Studies*, 23 (14): 2283–2320. https://doi.org/10.1177/0010414020912275.

Hartman, Alexandra C., Robert A. Blair, and Christopher Blattman. 2020. "Engineering Informal Institutions: Long-Run Impacts of Alternative Dispute Resolution on Violence and Property Rights in Liberia." *The Journal of Politics* 83 (1): 381–89. https://doi.org/10.1086/709431.

Harvey, Paul, and Sara Pavanello. 2018. "Multi-Purpose Cash and Sectoral Outcomes: A Review of Evidence and Learning." Geneva: UN Refugee Agency. https://www.unhcr.org/5b0ea3947.pdf.

Hidrobo, Melissa, and Lia Fernald. 2013. "Cash Transfers and Domestic Violence." *Journal of Health Economics* 32 (1): 304–19. https://doi.org/10.1016/j.jhealeco.2012.11.002.

Hidrobo, Melissa, Amber Peterman, and Lori Heise. 2016. "The Effect of Cash, Vouchers, and Food Transfers on Intimate Partner Violence: Evidence from a Randomized Experiment in Northern Ecuador." *American Economic Journal: Applied Economics* 8 (3): 284–303. http://dx.doi.org/10.1257/app.20150048.

Hidrobo, Melissa, John Hoddinott, Amber Peterman, Amy Margolies, and Vanessa Moreira. 2014. "Cash, Food, or Vouchers? Evidence from a Randomized Experiment in Northern Ecuador." *Journal of Development Economics* 107: 144–56. http://dx.doi.org/10.1016/j.jdeveco.2013.11.009.

Hossain, Mazeda, and Alys McAlpine. 2017. *Gender Based Violence Research Methodologies in Humanitarian Settings: An Evidence Review and Recommendations*, Gender Based Violence Report. Cardiff: Elrha, 1–24. http://www.elrha.org/wp-content/uploads/2017/09/Final-for-Web-ElrhaR2HC-GBV-Report_0817-v6-web.pdf.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter Van der Windt. 2019. "Exporting Democratic Practices: Evidence from a Village Governance Intervention in Eastern Congo." *Journal of Development Economics* 140: 279–301. https://doi.org/10.1016/j.jdeveco.2019.03.011.

Inter-Agency Standing Committee. 2016. *Inter-Agency Standing Committee Policy on Protection in Humanitarian Action.* Geneva: Inter-Agency Standing Committee. https://interagencystandingcommittee.org/iasc-protection-priority-global-protection-cluster/iasc-policy-protection-humanitarian-action-2016.

International Rescue Committee. 2022. "Humanitarian Research Toolkit." Humanitarian Research Toolkit. https://www.humanitarianresearch.rescue.org/context.

Lyall, Jason, Yang-Yang Zhou, and Kosuke Imai. 2020. "Can Economic Assistance Shape Combatant Support in Wartime? Experimental Evidence from Afghanistan." *American Political Science Review* 114 (1): 126–43. https://doi.org/10.1017/S0003055419000698.

McIntosh, Craig and Andrew Zeitlin. 2018. "Benchmarking a Child Nutrition Program against Cash: Experimental Evidence from Rwanda." Working paper, 2018. https://gps.ucsd.edu/_files/faculty/mcintosh/mcintosh_research_child_nutrition.pdf.

McIntosh, Craig, and Andrew Zeitlin. 2020. "Using Household Grants to Benchmark the Cost Effectiveness of a USAID Workforce Readiness Program." Working paper, September 2020. https://arxiv.org/pdf/2009.01749.pdf.

Millán, Teresa Molina, Tania Barham, Karen Macours, John A. Maluccio, and Marco Stampini. 2019. "Long-Term Impacts of Conditional Cash Transfers: Review of the Evidence." *The World Bank Research Observer* 34 (1): 119–59. https://doi.org/10.1093/wbro/lky005.

Mousa, Salma. 2020. "Creating Coexistence: Intergroup Contact and Soccer in Post-ISIS Iraq." *Science* 369 (6505): 866–870. https://doi.org/10.1126/science.abb3153.

Mvukiyehe, Eric and Peter Van der Windt. 2020. "Assessing the Longer Term Impact of Community-Driven Development Programs: Evidence from a Field Experiment in the Democratic Republic of Congo." World Bank Policy Research Working Paper #9140, July 2020. http://hdl.handle.net/10986/33296.

O'Callaghan, Paul, John McMullen, Ciarán Shannon, Harry Rafferty, and Alastair Black. 2013. "A Randomized Controlled Trial of Trauma-Focused Cognitive Behavioral Therapy for Sexually Exploited, War-Affected Congolese Girls." *Journal of the American Academy of Child and Adolescent Psychiatry* 52 (4): 359–69. https://doi.org/10.1016/j.jaac.2013.01.013.

Özler, Berk, Çigdem Çelik, Scott Cunningham, and P. Facundo Cuevas. 2020. "Children on the Move: Progressive Redistribution of Humanitarian Cash Transfers among Refugees." Policy Research Working Paper #9471, November 2020. https://openknowledge.worldbank.org/bitstream/handle/10986/34772/Children-on-the-Move-Progressive-Redistribution-of-Humanitarian-Cash-Transfers-among-Refugees.pdf?sequence=1&isAllowed=y.

Paluck, Elizabeth Levy. 2010. "Is It Better Not to Talk? Group Polarization, Extended Contact, and Perspective Taking in Eastern Democratic Republic of Congo." *Personality and Social Psychology Bulletin* 36 (9): 1170–185. https://doi.org/10.1177/0146167210379868.

Paluck, Elizabeth Levy, and Donald P. Green. 2009. "Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda." *American Political Science Review* 103 (4): 622–644 https://doi.org/10.1017/S0003055409990128.

Pforr, Tobias, Lena Weingärtner, and Emily Wilkinson. 2020. *The Evidence Base on Anticipatory Action*. Rome: World Food Programme. https://www.wfp.org/publications/evidence-base-anticipatory-action.

Pople, Ashley, Ruth Hill, Stefan Dercon, and Ben Brunckhorst. 2021. "Anticipatory Cash Transfers in Climate Disaster Response." Centre for Disaster Protection Working Paper #6, April 2021. https://www.disasterprotection.org/publications-centre/anticipatory-cash-transfers-in-climate-disaster-response.

Puri, Jyotsna, Anastasia Aladysheva, Vegard Iversen, Yashodhan Ghorpade, and Tilman Brück. "What Methods May Be Used in Impact Evaluations of Humanitarian Assistance?" IZA Discussion Paper #8755, January 2015. https://www.iza.org/publications/dp/8755/what-methods-may-be-used-in-impact-evaluations-of-humanitarian-assistance.

Quattrochi, John, Jenny C. Aker, Peter van der Windt, and Maarten Voors. 2020. "Contributions of Experimental Approaches to Development and Poverty Alleviation: Field Experiments and Humanitarian Assistance." World Development 127 (104830): 1–3. https://doi.org/10.1016/j.worlddev.2019.104830.

Reichhold, Urban, Andrea Binder, and Norah Niland. 2013. *Scoping Study: What Works in Protection and How Do We Know?*, GPPi Report. Berlin: Global Public Policy Institute, 1-70. https://gppi.net/media/GPPi_2013_DFID_scoping-study-protection_180830_110813.pdf.

Schwab, Benjamin. 2020. "In the Form of Bread? A Randomized Comparison of Cash and Food Transfers in Yemen." *American Journal of Agricultural Economics* 102 (1): 91–113. https://doi.org/10.1093/ajae/aaz048.

Sphere Association. 2018. *The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response*, 4th ed. Geneva. Spere Association. www.spherestandards.org/handbook.

Van Daalen, Kim Robin, Sara Dada, Rosemary James, Henry Charles Ashworth, Parnian Khorsand, Jiewon Lim, Ciaran Mooney et al. 2022. "Impact of Conditional and Unconditional Cash Transfers on Health Outcomes and Use of Health Services in Humanitarian Settings: A Mixed-Methods Systematic Review." *BMJ Global Health* 7 (1): 1–25. https://doi.org/10.1136/bmjgh-2021-007902.

Yanagizawa-Drott, David. 2014. "Propaganda and Conflict: Evidence from the Rwandan Genocide." *The Quarterly Journal of Economics* 129 (4): 1947–94. https://doi.org/10.1093/qje/qju020.

LIST OF ORGANISATIONS CONSULTED

In conversations with a range of humanitarian practitioners, researchers and funders in late 2021 and early 2022, we gathered input on what sorts of learning and evaluation questions were most relevant for them. We are grateful for their time and the thoughts they shared so generously with us, both during dedicated meetings and on the margins of the training activities we organised as part of launching this new research focus. While we list below the organisations to which they were affiliated at the time of our engagement, the content of the learning agenda should not be construed to reflect their positions.

- Action Contre la Faim (ACF)

- Christian Blind Mission (CBM)

- Consilient Research

- Crisis and Support Centre of the French Ministry for Europe and Foreign Affairs (CDCS)

- Danish Refugee Council (DRC)

- Enhancing Learning & Research for Humanitarian Assistance (Elrha)

- European Commission's Directorate-General for European Civil Protection and Humanitarian Aid (DG ECHO)

- FHI 360

- Global Public Policy Institute (GPPi)

- Global Women's Institute, George Washington University (GWI)

- Immigration Policy Lab (IPL)

- IMPACT Initiatives

- Innovations for Poverty Action (IPA)

- InterAction

- International Committee of the Red Cross (ICRC)

- International Rescue Committee (IRC)

- Médecins Sans Frontières (MSF)

- Mercy Corps

- Nonviolent Peaceforce

- Norwegian Agency for Development Cooperation (Norad)

- Norwegian Church Aid (NCA)

- Norwegian Refugee Council (NRC)

- Overseas Development Institute (ODI)

- Refugee Wellbeing and Integration Initiative (RWII)

- Save the Children

- The New Humanitarian

- UK Foreign, Commonwealth & Development Office (FCDO)

- UN Children's Fund (Unicef)

- UN Office for the Coordination of Humanitarian Affairs (OCHA)

- UN Refugee Agency (UNHCR)

- Bureau for Humanitarian Assistance, USAID (BHA)

- Women in Crisis Response (WiCR)

- World Bank Group (WB)

- World Food Programme (WFP)

- ZOA International

APPENDIX I: GLOSSARY

Assumptions/Risks: Assumptions are the factors that need to hold for your theory of change to be valid. Risks describe the potential for any of your assumptions to fail.

Attribution: The extent to which the observed change in outcome is the result of the intervention, having allowed for all other factors which may also affect the outcome(s) of interest.

Attrition: Either the drop out of subjects from the sample during the intervention or the failure to collect data from a subject in subsequent rounds of a data collection. Either form of attrition can result in biased impact estimates.

Baseline: Pre-intervention, ex-ante. The situation before an intervention, against which progress can be assessed or comparisons made. Baseline data is collected before a programme or policy is implemented to assess the 'before' state.

Bias: The extent to which the estimate of impact differs from the true value as a result of problems in the evaluation or sample design.

Causality: The capacity of one variable to influence another. The first variable may bring the second into existence or may cause the incidence of the second variable to fluctuate.

Cluster: A cluster is a group of subjects who are similar in one way or another. For example, in a sampling of school children, children who attend the same school would belong to a cluster, because they share the same school facilities and teachers and live in the same neighbourhood.

Cluster sample: Sample obtained by drawing a random sample of clusters, after which either all subjects in selected clusters constitute the sample or several subjects within each selected cluster are randomly drawn.

Comparison group: A group of individuals whose characteristics are similar to those of the treatment groups (or participants) but who do not receive the intervention. Comparison groups are used to approximate the counterfactual. In a randomised evaluation it consists of those randomly chosen not to receive access to a programme.

Confidence level: The level of certainty that the true value of impact (or any other statistical estimate) will fall within a specified range.

Confounding factors: Other variables or determinants that affect the outcome of interest.

Contamination: When members of the control group are affected by either the intervention (see 'externalities/ spillover effects') or another intervention that also affects the outcome of interest. Contamination is a common problem as there are multiple development interventions inmost communities.

Correlation: Indicates the extent to which two variables tend to increase or decrease in parallel. However, correlation by itself does not imply causation. There may be a third factor, for example, that is responsible for the fluctuations in both variables.

Cost-benefit analysis: An approach to comparing the costs and benefits of a programmes in which all the different benefits of a programme are translated into a one scale (usually a monetary scale) and then compared to the costs of the programme.

Counterfactual: The counterfactual is an estimate of what the outcome would have been for a programme participant in the absence of the programme. By definition, the counterfactual cannot be observed, and therefore it must be estimated using comparison groups.

Dependent variable: A variable believed to be predicted by or caused by one or more other variables (independent variables). The term is commonly used in regression analysis.

(Sources: 3ie and the World Bank)

Difference-in-differences (also known as double difference or D-in-D): The difference between the change in the outcome in the treatment group compared to the equivalent change in the control group. This method allows us to take into account any differences between the treatment and comparison groups that are constant over time. The two differences are thus before and after and between the treatment and comparison groups.

Endline: Measurement at the end of a study.

Evaluation: Evaluations are periodic, objective assessments of a planned, ongoing or completed project, programme, or policy. Evaluations are used to answer specific questions often related to design, implementation and/or results.

Evidence-based policy: Public policy or social programmes informed by rigorously established objective evidence.

Ex ante evaluation design: An impact evaluation design prepared before the intervention occurs. Ex ante designs are stronger than ex post evaluation designs because of the possibility of considering random assignment and the collection of baseline data from both treatment and control groups. Also called prospective evaluation.

Ex post evaluation design: An impact evaluation design prepared once the intervention has started, and possibly been completed. Unless the programme was randomly assigned, a quasi-experimental design must be used.

External validity: The extent to which the causal impact discovered in the impact evaluation can be generalised to another time, place, or group of people. External validity increases when the evaluation sample is representative of the universe of eligible subjects.

Externalities/Spillover effects: When the intervention has an impact (either positive or negative) on units not in the treatment group. Ignoring spillover effects results in a biased impact estimate. If there are spillover effects then the group of beneficiaries is larger than the group of participants.

Follow-up survey: Also known as 'post-intervention' or 'ex-post' survey. A survey that is administered after the programme has started, once the beneficiaries have benefited from the programme for some time. An evaluation can include several follow-up surveys. Hawthorne effect: The "Hawthorne effect" occurs when the mere fact that you are observing subjects makes them behave differently.

Hypothesis: A specific statement regarding the relationship between two variables. In an impact evaluation the hypothesis typically relates to the expected impact of the intervention on the outcome.

Impact: Any change in outcome that is caused by a programme; the difference between an outcome with the programme and the outcome that would have been seen in the absence of the programme.

Impact evaluation: An impact evaluation tries to make a causal link between a program or intervention and a set of outcomes. It also tries to answer the question of whether a program is responsible for changes in the outcomes of interest. Contrast with 'process evaluation'.

Independent variable: A variable believed to cause changes in the dependent variable, usually applied in regression analysis.

Indicator: An observable signal used to measure inputs, outputs, outcomes, and impacts.

Inputs: The financial, human, and material resources used for the intervention's development.

Intention to treat (ITT) estimate: The average treatment effect calculated across the whole treatment group regardless of whether they actually participated in the intervention or not. Compare to 'local average treatment effect estimate'.

Internal validity: The acceptability of an evaluation's results in terms of the causal impact of the intervention. To say that an impact evaluation has internal validity means that it uses a valid comparison group, that is, a comparison group that is a valid estimate of the counterfactual.

Intra-cluster correlation: Intra-cluster correlation is correlation (or similarity) in outcomes or characteristics between subjects that belong to the same cluster. For example, children who attend the same school would typically be similar or correlated in terms of their area of residence or socio-economic background.

John Henry effect: The 'John Henry effect' happens when comparison subjects work harder to compensate for not being offered a treatment. When one compares treated units to those '"harder-working"' comparison units, the estimate of the impact of the programme will be biased: we will estimate a smaller impact of the programme than the true impact we would find if the comparison units did not make the additional effort.

Literature review: It seeks to summarize the information that existing studies have gathered on the context, process or impact of a specific program

Local Average Treatment Effect (LATE) estimate: The treatment on the treated estimate is the impact (average treatment effect) only on those who actually received the intervention. Compare to intention to treat.

Logical model (also referred to as Theory of Change): Describes how a programme should work, presenting the causal chain from inputs, through activities and outputs, to outcomes. While logical models present a theory about the expected programme outcome, they do not demonstrate whether the programme caused the observed outcome. A theory-based approach examines the assumptions underlying the links in the logical model.

Meta-analysis: The systematic analysis of several pre-existing studies of one intervention to produce a quantitative estimate of effect size. Meta-analyses also use the techniques of literature review to decide which studies are included in the analysis

Midline: A measure taken in the middle of a study.

Minimum desired effect: Minimum change in outcomes that would justify the investment that has been made in an intervention, accounting for not only the cost of the programme and the type of benefits that it provides but also the opportunity cost of not having invested funds in an alternative intervention. The minimum desired effect is an input for power calculations: evaluation samples must large enough to detect at least the minimum desired effects with sufficient power.

Needs assessment: Research that carefully collects descriptive information, both qualitative and quantitative, about problems that may exist and the needs of a population a programme is designed to serve.

Null hypothesis: A hypothesis that might be falsified based on observed data. The null hypothesis typically proposes a general or default position. In evaluation, the default position is usually that there is no difference between the treatment and control group, or in other words, that the intervention has no impact on outcomes.

Outcome: A variable that measures the impact of the intervention. Can be intermediate or final, depending on what it measures and when.

Output: The tangible, immediate, and intended products or consequences of the intervention that are still within the implementer's control (i.e. they do not depend on the end client).

Partial Compliance, Imperfect Compliance: When only a fraction of the individuals who are offered the treatment take it up. Conversely, some members of the comparison group may receive the treatment.

Power calculation: A calculation of the sample required for the impact evaluation, which depends on the minimum effect size that we want to be able to detect (see 'minimum desired effect') and the required level of confidence.

Pre-post comparison: Also known as a before and after comparison. A pre-post comparison attempts to establish the impact of a programme by tracking changes in outcomes for programme beneficiaries over time using measures both before and after the programme or policy is implemented.

Process evaluation: Process evaluation, also known as implementation assessment or assessment of programme process, analyses the effectiveness of programme operations, implementation, and service delivery. When process evaluation is ongoing it is called programme monitoring (as in Monitoring and Evaluation). Contrast with 'impact evaluation'.

P-value: The probability of obtaining outcomes such as those produced by the experiment had the null hypothesis been true.

Quasi-experimental design: Impact evaluation designs that create a control group using statistical procedures (and do not use random assignment). The intention is to ensure that the characteristics of the treatment and control groups are identical in all respects, other than the intervention, as would be the case in an experimental design.

Random assignment: An intervention design in which members of the eligible population are assigned at random to either the treatment group (receive the intervention) or the control group (do not receive the intervention). That is, whether someone is in the treatment or control group is solely a matter of chance, and not a function of any of their characteristics (either observed or unobserved).

Random sample: The best way to avoid a biased or unrepresentative sample is to select a random sample. A random sample is a probability sample where each individual in the population being sampled has an equal chance (probability) of being selected.

Randomised evaluation (RE) (also known as randomised controlled trial, or RCT): An impact evaluation design in which random assignment is used to allocate the intervention among members of the eligible population. Since there should be no correlation between participant characteristics and the outcome, the differences in outcome between the treatment and control can be fully attributed to the intervention, i.e. there is no selection bias. However, REs may be subject to several types of bias and so they need to follow strict protocols. Also called 'experimental design'.

Regression analysis: A statistical method which determines the association between the dependent variable and one or more independent variables.

Selection bias: A possible bias introduced into a study by the selection of different types of people into treatment and comparison groups. As a result, the outcome differences may potentially be explained as a result of pre-existing differences between the groups rather than the treatment itself.

Significance level: The significance level is usually denoted by the Greek symbol, α (alpha). Popular levels of significance are 5 percent (0.05), 1% (0.01) and 0.1 percent (0.001). If a test of significance gives a p-value lower than the α-level, the null hypothesis is rejected. Such results are informally referred to as 'statistically significant'. The lower the significance level, the stronger the evidence required. Choosing the level of significance is an arbitrary task, but for many applications, a level of 5 percent is chosen, for no better reason than that it is conventional.

Stratified sample: Obtained by dividing the population of interest (sampling frame) into groups (e.g. male and female) and then by drawing a random sample within each group. A stratified sample is a probabilistic sample: every unit in each group (or strata) has the same probability of being drawn.

Treatment group: The group of people, firms, facilities or other subjects who receive the intervention. Also called participants.

Unobservables: Characteristics which cannot be observed or measured. The presence of unobservables can cause selection bias in quasi-experimental designs.