JOURNAL OF DEVELOPMENT EFFECTIVENESS, 2017
VOL. 9, NO. 4, 519–542
https://doi.org/10.1080/19439342.2017.1388267

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

Can rigorous impact evaluations improve humanitarian assistance?

Jyotsna Puri[a], Anastasia Aladysheva[b], Vegard Iversen[c], Yashodhan Ghorpade[d] and Tilman Brück[e]

[a]Independent Evaluation Unit, Green Climate Fund, Songdo, South Korea; [b]Stockholm International Peace Research Institute (SIPRI), Stockholm, Sweden; [c]Indian Institute of Management Ahmedabad, Ahmedabad, India; [d]Social Protection and Jobs Unit, World Bank, Washington, DC, USA; [e]Leibniz Institute of Vegetable and Ornamental Crops (IGZ) and ISDC – International Security and Development Center, Berlin, Germany

ABSTRACT

Each year billions of US-dollars of humanitarian assistance are mobilised in response to man-made emergencies and natural disasters. Yet, rigorous evidence for how best to intervene remains scant. This dearth reflects that rigorous impact evaluations of humanitarian assistance pose major methodological, practical and ethical challenges. While theory-based impact evaluations can crucially inform humanitarian programming, popular methods, such as orthodox RCTs, are less suitable. Instead, factorial designs and quasi-experimental designs can be ethical and robust, answering questions about how to improve the delivery of assistance. We argue that it helps to be prepared, planning impact evaluations before the onset of emergencies.

ARTICLE HISTORY
Received 13 January 2016
Accepted 29 September 2017

KEYWORDS
Impact evaluation; methodology; research design; statistics; humanitarian emergency; humanitarian assistance; disaster; violent conflict; reconstruction; aid; development

> Understanding the impact of humanitarian assistance is an area where much work is needed … Linking impact measurement and accountability better to the funds agencies receive is a key recommendation of this review
>
> (Humanitarian Emergency Response Review, UK Government, March 2011).
>
> The evidence base proving which humanitarian responses are most effective is extremely lacking. Investments must be made in the consolidation of evidence about what works in response to different kinds of needs in different contexts
>
> (The Use of Evidence in Humanitarian Decision Making, ACAPS Operational Learning Paper, January 2013).

1. Introduction

In 2016, an estimated 164.2 million people were directly affected by humanitarian crises worldwide. While the international community responded by raising US$27.3 billion in funding, 40 per cent of the needs identified by the United Nations were not met.[1, 2]

In a context where lives are in danger and the demand for resources overwhelmingly exceeds supply, effective and efficient assistance and service delivery are key factors. However, and despite the countless ex-post evaluations conducted in the humanitarian sector, there is a dearth of *theory-based, reliable* evidence *causally* linking interventions to relevant outcomes. This paper critically examines whether and if so how, impact evaluation methods can provide reliable evidence to help improve the effectiveness and efficiency of humanitarian action.

CONTACT Jyotsna Puri jpuri@gcfund.org Independent Evaluation Unit, Green Climate Fund, Songdo, South Korea

So far, few studies have used impact evaluation methods to study humanitarian assistance.[3] In a review conducted at the end of 2013, 38 studies met our criteria for measuring attributable changes with statistical confidence, in outcomes and impact of humanitarian action. This shortfall partly reflects that evaluations of humanitarian assistance, especially during the relief phase, pose unique and important methodological challenges. Impact evaluations of rapid – and to a lesser extent slow – onset or protracted crises must deal with a mismatch between resources and needs, disruptions to everyday life, security concerns, the typical absence of baseline data, logistic, medical and other hurdles to sampling and data collection, with finding a valid counterfactual, and with ethical implications. The lack of impact evaluation experts in the humanitarian sector and the shortage of humanitarian expertise in the impact evaluation community further accentuate these challenges.

This paper presents methodological challenges and options associated with collecting and generating high-quality evidence that measures attributable change caused by humanitarian interventions, programmes or policies, with statistical confidence. We argue that evaluations need to use these methods to inform humanitarian action. Indeed, questions that such high-quality evidence should seek to answer include whether humanitarian assistance is reaching the right people at the right time, whether the intended improvements for beneficiaries are achieved or not (effectiveness), and whether assistance is delivered in the right doses and ways, and with manageable costs (efficiency). We also address whether these challenges can be overcome without compromising the ethical standards and principles that ought to guide humanitarian action and social science research practice.

This paper contributes to the development and the humanitarian practice literature in two ways. Firstly, we contribute to the literature on aid and development effectiveness. Just as impact evaluations helped improve our understanding of cash transfer policies and their potential effectiveness, we assess whether rigorous research methodologies could achieve the same for humanitarian assistance. In this regard, we speak to development researchers with experience of impact evaluations who are less familiar with humanitarian emergencies. Secondly, we also contribute to the humanitarian literature, which concerns itself with how to best organise responses to a variety of natural and man-made disasters. In this regard, we address an audience of humanitarian experts who may be less familiar with impact evaluation techniques. In fact, we posit that bridging the artificial divide between development and humanitarian sectors in practice and research should yield significant dividends. We begin, therefore, by presenting the particularities of the humanitarian sector in some detail, thereby motivating our methodological approach and findings.

In contrast to what some may construe as received wisdom (Few et al. 2014), we propose that it is possible to conduct rigorous impact evaluations in humanitarian emergencies. Drawing on real-life examples and findings from a small but growing academic literature, we show how impact evaluation methods can be used successfully and in an ethical manner to distil lessons about how to deliver humanitarian assistance effectively and efficiently. This can be achieved by adjusting research designs to programme and ground realities and through creative use of so-called factorial designs when crucial ethical principles are not compromised, and when not all delivery can be done in one fell swoop. In fact, there are many decisions that field staff need to make that can be usefully informed by impact evaluation findings.

We also show that impact evaluation methodologies may be used constructively, not only to understand impact, but to assess what *design* of programmes might be best suited to a humanitarian context, and to help understand *what method* of *delivery* might be most appropriate depending both on the phase during which humanitarian assistance is implemented and the context of the humanitarian disaster. There is therefore, we suggest, scope for improving practice in the humanitarian sector through learning based on impact evaluations (e.g. Humanitarian Emergency Response Review (HERR) 2011).

The rest of this article is structured as follows: Section 2 presents a taxonomy of humanitarian emergencies, while Section 3 expands Buttenheim's conceptual framework, which covers the phases and dynamics of sudden onset to also include slow onset emergencies. Sections 4 and 5 discuss some of the key and often unique methodological and practical challenges confronting

impact evaluation efforts in humanitarian settings. Section 6 discusses how to mitigate or overcome ethical concerns by suitably adapting research designs. Section 7, which is organised in line with the phases in Section 3, combines new suggestions with examples from the literature to illustrate how methodological, practical and ethical challenges can be addressed when designing impact evaluations of humanitarian assistance. Section 8 concludes the paper.

2. Humanitarian emergencies and humanitarian action

A humanitarian crisis is characterised by an exceptional and generalised threat to human life, health or subsistence. Crises may appear within the context of an existing lack of protection where a series of pre-existing conditions (poverty, inequality, lack of access to basic services), exacerbated by a natural disaster or armed conflict, are likely to aggravate destructive effects.[4]

Among practitioners there are two recognised definitions of humanitarian action (see for example Barnett 2005). In the narrower, *Dunantist* interpretation, humanitarian action is focused on saving lives, alleviating suffering and maintaining and protecting human dignity during and in the immediate aftermath of an emergency.[5] Accordingly, humanitarian action is short-term and distinct from development aid. The *Wilsonian* definition, in contrast, broadens the scope of humanitarian action to include responses to slow-onset and complex emergencies that may demand prolonged assistance to sustain human health, life and livelihoods. In this tradition, humanitarian action is not limited to providing immediate relief but also aids recovery and builds resilience. Our analysis below is relevant for both types of humanitarian actions; in either case, rigorous impact evaluation can help understand crises and improve humanitarian responses.

Scholars have proposed a variety of taxonomies to understand humanitarian emergencies. Buttenheim (2009) suggests five categories of disasters based on the immediate cause: (i) biological (epidemics, insect infestations, animal attacks), (ii) geophysical (earthquakes, volcanoes, dry mass movements), (iii) climatological (droughts, extreme temperatures, wildfires), (iv) hydrological (floods, wet mass movements), and (v) meteorological events (storms). We add a sixth category, violent conflict, to this list. Conflict differs from the others by being anthropogenic.[6]

We also contrast anticipated and unanticipated humanitarian emergencies. Few disasters, man-made or otherwise, are random and completely unanticipated events: we distinguish those that can be predicted with more than even odds. In addition, humanitarian crises can be sudden (e.g. an earthquake) or slow-onset (e.g. a famine). This latter distinction draws attention to contrasting operational and evidence requirements: severe drought conditions do not translate into a famine overnight, while pre-conflict tensions may simmer long before the outbreak of hostilities. An earthquake is, perhaps, the most compelling example of a sudden-onset emergency: meteorological events usually come with at least a short forewarning.

To direct effort and plan for assistance, many relief organisations closely monitor crisis hotspots. The World Bank maintains a list of fragile and conflict situations and another of natural disaster hotspots (see e.g. Dilley et al. 2005).[7] The access to such information can, as discussed below, crucially aid impact evaluation efforts.

Yet another taxonomy distinguishes preventive action (or action that helps to build long-term resilience and reduce the risk of future humanitarian emergencies), and humanitarian assistance provided in the immediate aftermath of an emergency, whether sudden or slow onset.[8] Over time, policies that have improved preparedness and increased resilience have reduced losses and casualties, and helped prevent the type of emergencies that were historically responsible for the largest number of casualties. A key insight from Devereux's (2000) analysis of twentieth century famines is that while famines during colonial times were tightly linked to droughts, post 1980 famines have typically occurred during conflicts (e.g. in Uganda, Sudan, Ethiopia, Eritrea, Niger). This is illustrated by records of casualties over a 200-year period in India. During the Great Bengal famine in 1769–1770, about one-third of the population of the province was wiped out (Menon 2013). Post 1974, there have been no famines in South Asia (Hussain 1995). Similarly, during the November 1970 cyclone which affected the

coastal belt of Bangladesh, approximately 300,000 people perished (World Bank 2010, 34). In 1997, during another powerful cyclone and after lessons from the 1970s cyclone had been duly absorbed, 111 people lost their lives in Bangladesh (Menon 2013).

These different typologies (types of humanitarian *emergencies* and types of humanitarian *assistance*) have implications for the way impact evaluations may be designed and planned in humanitarian settings. We next consider how humanitarian emergencies unfold, building on and expanding Buttenheim's (2009) analysis of sudden onset emergencies. This elaboration is important since a clear understanding of the dynamics of humanitarian emergencies is a prerequisite for the design of a rigorous impact evaluation.

3. The dynamics of humanitarian emergencies

Buttenheim (2009) divides a sudden onset emergency into four main phases:

(t_{-1}) Baseline: Pre-disaster phase. Most agencies assume that no data exists for this phase.[9]

(t_0) Emergency: The point (or period) in time when the disaster strikes or conflict breaks out.

(t_1) Relief phase: Relief is provided in the immediate aftermath of the disaster and as soon as access is restored. This phase often lasts for 3–6 months unless the crisis is (or becomes) protracted.

(t_2) Recovery phase: Longer-term assistance is provided to aid recovery to the pre-disaster 'condition' and to improve resilience. This phase may start about 6 months after the emergency.

We expand Buttenheim's (2009) framework to be able to cover other emergency events, phases and types of assistance:

(t_{-2}): The period before what we now define as an adverse event: to illustrate and starting at t-2, a gradual escalation of tension may trigger and ultimately create an emergency situation (conflict). Similarly, the gestation period between a severe drought (again starting at t-2) and famine conditions could be weeks, months or years (Devereux 2000). For such slow-onset emergencies, household welfare (e.g. health) deterioration and asset depletion may begin long before the emergency is declared.[10] This, then, presumes an emergency discontinuity or threshold beyond which negative health and other gradients deepen and risks to human life, health and other losses dramatically intensify.[11]

(t_{-1}): The period *immediately* before the emergency. This distinction is necessary since a **(t_{-1})** baseline will be straightforward for an earthquake, but not for a famine or a conflict. Since a gradual deterioration of welfare may begin long before the emergency, a **(t_{-1})** based estimate of 'normal' household welfare and asset holdings will be biased *downwards*: this introduces an *upward* bias in a **(t_{-1})** based assessment of the restoration of household welfare to 'normalcy' during the recovery phase.

(t_0): The time of the emergency. This raises the question of whether there is a 'correct' post- t_0 time to measure human welfare or asset holdings. As noted, emergency environments tend to be chaotic, with breakdown in public and other service delivery (e.g. water, sanitation, health care). In such environments, indicators or measures of health or human welfare may very rapidly deteriorate. An overarching objective of humanitarian assistance is to quickly arrest such slides.

(t_3): A point in time (much) after the disaster with activities directed at reducing the risk of future disasters/conflicts and the likelihood of large damages in the event of recurrence. The overall goal is thus to reduce vulnerability.

Figure 1 summarises these different phases of observation, intervention and outcomes. Using this diagram, impact evaluations may be used to measure attributable changes in a variety of outcome indicators in the following ways ('**X**' denotes a variable that measures welfare or health or any other indicator of household or individual well-being):

- **$Xt_{-2} - Xt_{-1}$** = household welfare loss induced during the pre-emergency phase which e.g. could be due to drought or escalation of tension: this may include gradual deterioration of health, quality of life and public services as well as depletion of assets.

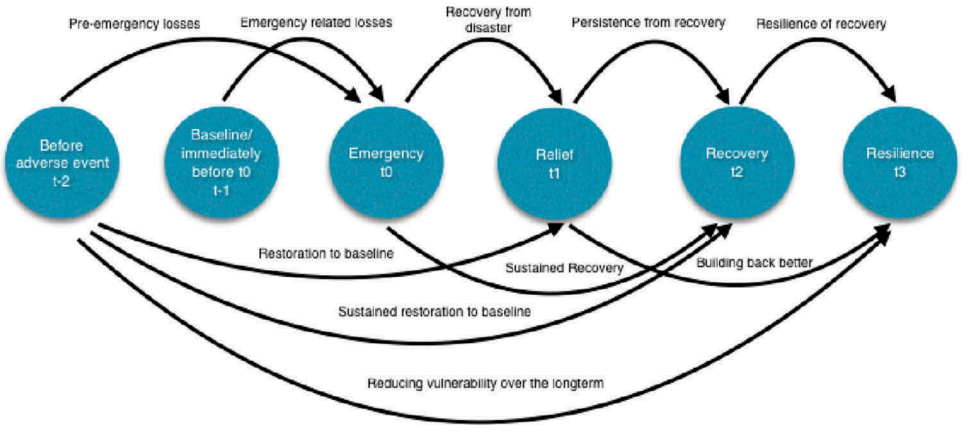

Figure 1. Stages of emergency.

- $\mathbf{Xt_{-1} - Xt_0}$ = household welfare loss induced by the emergency, which includes losses to health/life and/or asset loss/destruction.
- $\mathbf{Xt_{-2} - Xt_0}$ = total household welfare loss during the now analytically separate pre-emergency and emergency phases, which, accordingly, can be decomposed into $\mathbf{(Xt_{-2} - Xt_{-1}) + (Xt_{-1} - Xt_0)}$.
- $\mathbf{Xt_1 - Xt_{-1}}$ = extent of household recovery from the emergency.
- $\mathbf{Xt_1 - Xt_{-2}}$ = extent of household recovery from the adverse event.
- $\mathbf{Xt_2 - Xt_{-1}}$ = sustained restoration of households to baseline for sudden onset emergencies (t-2 = t-1).
- $\mathbf{Xt_2 - Xt_{-2}}$ = sustained restoration of households to the pre-adverse event baseline for slow onset emergencies, e.g. conflicts and famines (t-2 predates t-1).

4. Methodological challenges

Theory-based impact evaluations measure and help understand measurable changes in outcomes, outputs or long-term impact *causally* attributable to a policy, programme or intervention, with statistical confidence while relying on a pre-specified theory of change. We use this definition for theory-based impact evaluations, or impact evaluations in short, to argue that when carefully crafted and planned, impact evaluations can shed light on what, how and why changes as well as any unintended consequences occurred.

The ideal first step in an impact evaluation is to map a causal pathway – a theory of change – that links how activities are expected to relate to processes and outputs, articulating the assumptions required along this postulated pathway. Formative research (see also Section 7.3) and familiarity with the context are key ingredients in developing a persuasive theory of change. A counterfactual-based identification strategy is then required to attribute causal impact to the intervention, policy or programme.[12] There are, however, often stark contrasts between this ideal and the ground realities that will confront an impact evaluation effort in an emergency setting. In the present and the next two sections, we discuss, respectively, the methodological, practical and ethical challenges facing rigorous impact evaluations in such settings.

The methodological validity of an impact evaluation can be challenged by three types of biases: selection bias, information bias and contamination bias:

Selection bias occurs if attributes of the target group are likely to affect programme effectiveness. This will occur, for instance, if the most privileged receive emergency relief first, not accounting for other covariates such as their typically higher education and income, which likely affects how effective the relief programme will be for this group. Impact evaluations seek to establish what would have happened to disaster-stricken households who received relief

assistance, in the absence of this assistance. In humanitarian crises, selection bias may occur on three counts. Poorer, marginalised and more vulnerable households may (a) be more exposed (likely to experience a shock), e.g. by living closer to the flooding river; (b) suffer more damage and destruction once the shock occurs, e.g. because of lower quality housing; (c) lack the influence, networks and voice to secure access to disaster relief. This creates a 'triple' selection bias that represents a potential threat to the internal validity and thus the success of the impact evaluation.

Information bias occurs when respondent perceptions affect the accuracy of the information they provide, often in a manner correlated with their poverty, education and relief programme eligibility. In the absence of systematic baseline information, attempts to measure impacts of humanitarian action typically rely on respondent recall to assess changes in human welfare. However, during a crisis, respondents may not accurately remember details of e.g. their living conditions prior to the emergency. 'Recall error' may be further compounded if beneficiaries are interviewed by officials associated with the recovery effort, and expect levels of compensation to depend on their responses to a questionnaire. Furthermore, error and deviations in self-reporting may correlate with the severity and frequency of disaster exposure. Guiteras, Jina, and Mobarak (2015) find that households facing more regular flooding tend to under-report in self-assessments of whether or not they faced flooding at lower levels of flood exposure. In other words, they tend to internalise, and deem normal, a certain level of flooding exposure, unlike households experiencing a surprise flooding episode.

Contamination bias occurs if relief spills over to non-targeted (control) areas thereby affecting relief impact estimates. Contamination bias occurs in humanitarian assistance when people outside the intended targeted or treatment area receive benefits that were not meant for them (thus reducing the intensity and amount of the dose in a dose-response equation). Contamination bias can also occur if e.g. other organisations' relief efforts specifically decide to work in control areas because of the absence of relief programmes there, or specifically choose to work in treatment areas to realise operational and logistical synergies by piggybacking on existing relief efforts. This density of simultaneous relief efforts makes it harder to attribute impacts to a specific programme: it may still, however, and as discussed below, be possible to measure the *additional* and *differential* contribution of an intervention.

Furthermore, *non-random attrition* (that is, the better off or the most vulnerable may not be accounted for in less rigorous evaluations since they are the first to migrate or to perish during a natural disaster or conflict) and *non-random response* (the most accessible areas receive relief first but may also, other factors held constant, be disproportionately more likely to respond to surveys and be counted) can both be accounted for by impact evaluations.

Finally, it is relatively easy to establish the average effect of an intervention. A more challenging task is to identify if and how *heterogeneous effects* obtain, that is if an intervention has systematically different impacts on different groups of people, for example based on their asset endowments, gender, age or ethnicity. To give an example. three of the four evaluations undertaken by the Tsunami Evaluation Coalition (TEC) found that aid was disbursed disproportionately to areas that were easily served by transportation, rather than based on need, and that the old and disabled were often excluded from benefits because they were poorly informed about availability (Cosgrave 2007). At the same time, the TEC could not determine how much worse off these vulnerable groups were.

5. Practical challenges

Impact evaluations are harder to implement during humanitarian emergencies. The unique challenges that arise during humanitarian emergency settings depend (a) on the nature of the emergency (armed conflict involves additional and other practical challenges than an earthquake), and (b) on the precise post-emergency phase and outcomes that the evaluation focuses on. Practical challenges to impact evaluations in humanitarian contexts include:

Complex settings: Humanitarian crises are often unanticipated. Disasters disrupt the supply of basic services, and often result in (temporary) logistics and infrastructure breakdown, security and law and order concerns and so forth. Teams frequently find that baseline data do not exist. This occurs alongside the simultaneous presence of a multiplicity of actions and interventions. Further, agencies do not always know how long they will stay on, despite their response being precipitated by humanitarian crises (this has been the case, for example, in Somalia, Haiti, Pakistan, Rwanda, and the Democratic Republic of Congo).[13] The multitude of interventions, changing activities and outcomes, and blurred timelines make planning for an impact evaluation particularly challenging

Need for speed: speed and outreach are critically important in humanitarian assistance. Usually there is no time to train teams and plan and prepare an evaluation. This constraint is more binding for unanticipated, rapid-onset emergencies than for slow and protracted crises.

Multiplicity of actors: after a disaster, many international agencies, donor countries, foreign nationals, domestic and foreign non-governmental organisations (NGOs), national and local governments, their armed forces personnel, and others may provide financial and technical assistance. The World Bank (2010) reports that the number of international and domestic actors responding to disasters has increased in recent years, with roles not fixed and becoming more blurred over time. For example, during the Indian Ocean Tsunami, 42 international aid agencies responded: this number does not include national and local agencies and other private initiatives. This makes it harder to *ex ante* plan and subsequently implement a credible impact evaluation.

Attribution: the array of actors not only makes their coordination a challenge, but accentuates the difficulties with attributing impact to a particular humanitarian initiative (ibid).

High co-variability: since large areas are often affected during a humanitarian emergency, it may be difficult to find credible counterfactuals since few unaffected locations or population groups may be sufficiently similar to the population or populations that were affected but not targeted for reasons that are unrelated to their conditions.

Evaluations of preventive action: for humanitarian activities that are directed at prevention rather than post-emergency assistance, it can be difficult (and inconceivable), for ethical and technical reasons, to construct an explicit counterfactual for an intervention that e.g. seeks to prevent a severe drought from developing into a famine. A similar challenge is captured by a situation where escalation of tension threatens to develop into a full-blown conflict.

The emergency relief phase: As reiterated below, few impact evaluations have focused on the relief phase. This phase poses specific methodological challenges that have not been sharply articulated in the literature and that we therefore touch upon here. A unique characteristic of humanitarian settings, is that human well-being may rapidly deteriorate. This implies that e.g. health indicators and variable values will change more rapidly than in the more stable developmental contexts that impact evaluations usually focus on. Let h_{max} represent the maximum value of a health indicator h once the emergency sets in, while h_{min} is the lowest value h can drop to if permitted to deteriorate: such deterioration may be highly non-linear and involve critical thresholds.

Consider the case of child dehydration and diarrhoea. Suppose that humanitarian assistance arrives very soon after a disaster strikes and that a slide from h_{max} to h_q occurs in a 'treatment' group that receives assistance: the corresponding slide in a control group that assistance is unable to reach is from h_{max} to h_{qc}. Let $h_q > h_{qc}$ so that the received assistance affects child health and that a positive treatment effect thus is discernible.

If, instead, assistance arrives late, so that $h_l < h_q$ (i.e. health indicators further deteriorate because of the late arrival, which is realistic for the diarrhoea and dehydration example) and ditto for $h_{lc} < h_{qc}$: the impact estimate will be determined by $(h_l - h_{lc})$. There is a 'perverse incentive' problem if $(h_l - h_{lc}) > (h_q - h_{qc})$ since impact for late arrivals (and 'low quality' administration) will exceed the impact for early arrival (and 'high quality administration'). Hence, humanitarian assistance may generate larger positive impacts if conditions are allowed to deteriorate. Naïve application in settings where well-being indicators rapidly deteriorate thus involves the risk of ranking a highly effective and early arrival intervention as inferior to a less effective and late arrival

intervention, since the impact size of the latter, because of the uniquely rapid slide, is likely to be larger. This illustrates the need to pay very careful attention to measurement issues and timing during surveys and in analytical follow up of the emergency relief phase

6. Ethical challenges

Ethics are at the forefront of concerns about conducting impact evaluations of humanitarian assistance especially if a control or a comparison group means withholding assistance for this group. This may not be acceptable in an emergency situation. The Declaration of Helsinki (1964 and its subsequent updates) and the Belmont Report (1979) set ethical standards to protect human subjects participating in health and medical research and offer instructive guidance. In a humanitarian emergency, one option would be to translate the no harm principle as: '*the taken approach may significantly improve but will not worsen outcomes for emergency relief recipients*'. However and resonating with Angell's (1997) account of ethical standards for medical trials, this is not sufficient: if the alternative offered to the treatment group is preferable at the outset, withholding this best available alternative from the control group amounts to a severe ethical breach.

Distinguishing between the emergency and the subsequent recovery and resilience phases, ethical concerns are particularly pressing during the emergency phase. Facilitating rigorous comparisons while maintaining high ethical standards, factorial designs represent an attractive option by facilitating learning about the relative effectiveness of e.g. different ways of delivering emergency assistance, whether assistance as cash or food is preferable (e.g. Schwab, Margolies, and Hoddinott 2013) and so forth.

Another ethical concern, accentuated during the emergency phase, is whether research ethics protocol is at a higher than usual risk of being compromised. For respondents in a particularly vulnerable situation, voluntary and informed consent is likely to be harder to credibly ensure. Further, and as touched upon earlier, survey participant responses at odds with the interests of evaluators, local organisations or humanitarian organisation representatives may also be harder to obtain.[14]

For the subsequent recovery and resilience phases (see Section 3 and Figure 1), standard practices, like staggered roll-out (also called pipeline or phased in design) will often be acceptable. For example, in the study by IRC (2011) in Burundi, households were first randomly selected to participate in a village savings and loans association intervention. Of those selected, half of the households were randomly selected and assigned either to a waitlist control group (phasing in) that receive treatment in the future (first arm), or a treatment group that immediately receive the treatment (second arm). Of the households in the treatment group, half were selected to also participate in the family-based discussion group (third arm). Note, however, that few studies discuss explicitly the ethics of imposing such treatment alternatives (or of withholding them). In Puri et al. (2014), we provide some examples of how these ethical concerns may be addressed.

Employing other data sources can also help alleviate ethical concerns. The main requirement in impact evaluations is to introduce or exploit a variation that helps to either naturally or artificially create comparison groups and intervention groups that allow us to understand what would have happened in the absence of the intervention. This variation needs to be exogenous to the intervention being examined, i.e. not affected by it nor affecting it. RCTs create this exogenous variation by random selection. However, many other sources of variation can be exploited. One such opportunity, discussed in more detail below, is provided by spatially disaggregated or GIS data.

7. Lessons for conducting impact evaluations of humanitarian assistance

Drawing on the above typology of humanitarian contexts and phases and of our identification of the methodological, practical and ethical challenges that impact evaluations of humanitarian assistance pose, we next examine how a variety of questions that are relevant to policy makers and programme managers may be answered. To aid this task, we first searched and carefully

reviewed existing impact evaluations of humanitarian assistance. Our systematic search of databases of evaluations found that among more than 900 existing evaluations of humanitarian interventions to date, only 31 classify as impact evaluations.[15]

In line with our typologies in Sections 2 and 3, we organised the 31 impact evaluations according to the phase of humanitarian assistance, while distinguishing Dunantist and Wilsonian interpretations. Among these 31 studies, 20 were impact evaluations of peacebuilding and conflict prevention interventions, only one study was an impact evaluation in relation to an unanticipated disaster and two studies were impact evaluations of anticipated disasters covering medium and long term recovery and resilience periods. Only eight were impact evaluation studies of emergency phase relief.

Twenty of these studies report doing a formal test of intervention and control group balance; five studies mention doing power analysis when deciding on sample size; 15 studies had a narrative of underlying economic theory and hypothesised causal pathways; only four studies mention ethical approval or discuss ethical concerns; and most (23) studies used randomised control trial (RCT) as an identification method to select subjects for intervention and for control: the remaining eight studies used quasi-experimental methods.

For the eight impact evaluations of emergency phase relief, six mention the theory of change, four mention power analysis, five conducted a test of balance, and three mention ethics. Among the 23 recovery and resilience stage studies, nine mention the theory of change, 15 mention the test of balance, and one mentions power analysis: one study discusses ethics.

We now consider the issues of relevance to policy makers and programme managers, flagged above, on how critical concerns related to comparators, assessing causality, urgency, confounding, length of time periods, ethics and bias may be addressed.

7.1. *The emergency relief phase*

We begin with studies covering the acute emergency phase which largely overlaps with the Dunantist interpretation of humanitarian action. Studies that address the methodological and ethical concerns discussed above, primarily use a factorial (randomised) design: this does not require a *pure* control group. The critical question for humanitarian organisations is not *if and whether* action should be taken (i.e. comparing the effectiveness of a programme with doing nothing) but to determine *which*, among the available action options, is more effective. A pure control group is not necessary if an organisation wants to understand *which strategy* to implement. To illustrate, Schwab, Margolies, and Hoddinott (2013) investigate the relative effectiveness of two randomly assigned treatment arms in Yemen: cash and food transfers.

Crucially, factorial designs neutralise ethical concerns over assigning no-intervention to a control group. Accordingly, in an acute emergency situation, all groups will receive treatment – such as assistance in the form of a basic provision package – but some groups receive an incremental treatment. For example, Doocy and Burnham (2006) evaluate water-cleaning interventions in a refugee camp in Liberia. While both groups received improved water storage, one group was randomly assigned an additional water treatment intervention where team members were uncertain whether this additional intervention would add value or not. Another example is a study in Chad by Huybregts et al. (2012) which compared a general food distribution programme and ready-to-use supplementary food with a general food distribution (see Table A1 in the Appendix).

Overall, a wide variety of questions relating to logistics and how to best organise emergency relief can be answered by impact evaluations using factorial designs. These include questions about camp organisation; measuring the effects of crowding in settings where target populations are low-income and nutritionally vulnerable and where otherwise proven 'best practices' may not apply or even be harmful. Other examples include understanding the value *added* from complementing basic provisions with hygiene education, the use of provision packages that are more

reflective of local food habits and routines and the targeting of emergency provisions or the top-up of emergency provisions to women.

7.2. *Sudden-onset emergencies: recovery and resilience phases*

Most impact evaluations of humanitarian assistance measure programme effects during recovery or resilience stages, which are consistent with the Wilsonian interpretation and that may occur several years after the disaster or conflict. During this phase, we identified studies that, in order to overcome the above challenges, use RCTs, quasi-experimental methodologies, recall data or delayed treatment. As in the relief phase, the use of multiple treatment arms and factorial design is common.

Park and Wang (2017) evaluates the impact of governmental support to the households affected by the 2008 earthquake in western China that killed almost 70,000 people and brought substantial damage to the local infrastructure. In addition to the relief efforts provided by the government of China shortly following the disaster, in early 2009 the affected communities received a stimulus package from the state to rebuild infrastructure and social contributions from the NGOs and private individuals. Using the household survey conducted 10 months after the earthquake the authors found that the mean income per capita rose by 17.5 per cent and the poverty rate declined from 34 per cent to 19 per cent as a result of the assistance. The evaluation was based on instrumental variable method with random sampling and the survey consisted of a set of retrospective questions on the household welfare.

De Mel, McKenzie, and Woodruff (2012) investigate the recovery of private firms in Sri Lanka after the 2004 Indian Ocean tsunami. Using data from 209 enterprises, the authors employ a four-arm RCT with a delayed treatment control group. They randomly assign four types of treatments to firms: two values of monetary grants are distributed either as cash or in-kind. By comparing treated firms with comparable firms, they found a positive effect of grants on profits, representing a 9.9 per cent real monthly return on the treatment. They also find that direct aid is more important in the recovery of enterprises operating in the retail sector than for manufacturing and service sector firms. Table A2 in the Appendix provides the summaries of the studies, with methodology used and main findings.

7.2.1. *Post-conflict: recovery and resilience phases*

Many impact evaluations examining the recovery and resilience phase of humanitarian assistance focus primarily on the impact of community-driven peacebuilding and stabilisation initiatives in fragile states. Samii, Brown, and Kulma (2011) review 25 most recently completed or ongoing impact evaluations of stabilisation interventions in post-conflict countries. Building on the review by Samii et al. (2011), Gaarder and Annan (2013) explore further the impact evaluations of peacebuilding and stabilisation interventions. Finally, Cameron et al. (2015) conduct a thorough search on all recent impact evaluations with a view to identify research gaps. Building on the three reviews, we identified 20 impact evaluation studies related to humanitarian assistance from the following countries: Rwanda, Burundi, Liberia, DRC, Uganda, Cote d'Ivoire, Sri Lanka, Philippines, Nepal, Thailand and Palestine. Table A3 in the Appendix provides summaries of these studies, with methodology used and main findings.

As above, most studies employ RCT with delayed treatment control group, factorial designs or variations in treatment. Five studies use a quasi-experimental methodology. Kondylis (2008) and Scholte et al. (2011) use difference-in-difference with matching to compare treatment and control groups in Rwanda. Levely (2014) employs propensity score matching to match treatment and control groups to understand the effect of an ex-combatant reintegration programme in Liberia. Malhotra and Liyanage (2005) use a matching technique that compares participants and non-participants to understand the effectiveness of peace workshops in Sri Lanka. Finally, Mvukiyehe and Samii (2010) investigate the impact of a UN programme on conflict de-escalation and security by matching the treatment and control communities.

7.2.2. *Anticipated emergencies*

Two studies are impact evaluations of anticipated emergencies. The first, by Aker et al. (2011), focused on recurrent droughts in Niger – which have a long history – and with the most recent drought in 2010. Aker et al. (2011) study the effectiveness of providing cash transfers via mobile phones. The use of mobile technology may potentially reduce the costs to implementing agencies of providing cash to rural populations living in remote areas with few financial institutions. The authors employed a three-arm RCT with manual cash transfer, cash transfer via mobile phone (the 'zap' approach), and a placebo group (with manual cash transfer and a mobile phone). They found that the 'zap' approach reduced the variable costs to the implementing agency by 30 per cent and the recipients of accessing these transfers by saving almost 1 USD per household. It also led to a more diverse spending on food items among beneficiaries. Households in 'zap' villages purchased 0.78 more types of food and non-food items as compared with the cash group.

Another study investigates assistance after floods in Bangladesh. Just like droughts in Niger, these floods are seen as an anticipated emergency since these are recurrent events expected with high probability each year. Being located on the Ganges Delta, a large territory of the country is flooded each year leading to thousands of casualties and destroying millions of homes. The assistance was introduced in Bangladesh in 2004, in response to floods the same year, and allowed for rescheduling of savings and instalments in microfinance institutions. Shoji (2010) illustrates the use of recall data in impact evaluation of this assistance. By using the same 326 households before and after the policy was introduced, the author investigated the impact of the possibility of rescheduling on individual meal frequency. The policy acted as a safety net during the natural disaster by decreasing the probability that people skip meals during negative shocks by 5.1 per cent, with a more pronounced effect on women and the landless.

7.3. *Other challenges (and their solutions): heterogeneity, missing baseline data, IE limitations*

Humanitarian crises are seldom completely external to the social and political dynamics of communities and groups. Existing structural and systematic inequalities in terms of income, wealth, knowledge, power, and voice can, as noted above, predispose certain groups to experiencing crises, or to their more aggravated effects. Wisner et al. (1994) argue that such humanitarian crises are often 'the product of social, political and economic environments.' Impact evaluations of humanitarian (as indeed other) interventions must be cognizant of extant systemic and structural inequalities and incorporate a meaningful understanding of these factors in design. For example, in a setting where the coverage of regular social protection and social assistance programmes even in pre-crisis times among the poorest, or among a social group is low, it is not hard to imagine that their access to humanitarian assistance may also be limited. In such a setting, evaluators must attempt to examine impact heterogeneity – for instance by stratifying treatment and control groups to be able to track the uptake and impact of humanitarian assistance among the likely marginalised groups.

Quasi-experimental methodology – regression discontinuity designs in particular – may be especially useful when evaluating a programme post-facto and when baseline data are not available. Nielsen, Jahan, and Canteli (2012) evaluated effectiveness of food assistance provided by the World Food Programme (WFP) and the United Nations High Commissioner for Refugees (UNHCR) by comparing two groups of refugees in Bangladesh: registered refugees received assistance and were compared with unregistered refugees, a natural control group who did not receive assistance because of an exogenous policy event. Accordingly, the authors did not have to randomly assign the intervention, since treatment status was determined by a policy unrelated to the overall condition of the refugees. Lehmann and Masterson (2014) used a regression discontinuity design to evaluate cash assistance for Syrian refugees in Lebanon. Cash assistance was provided by the International Rescue Committee (IRC) for the households in high altitudes

(500 m above the sea level in the coldest areas). Households slightly above the altitude eligibility cutoff (treatment group) were then compared with the households living slightly below the eligibility cutoff (control group).

While relatively unexplored, geographic information databases have great potential for tackling methodological and ethical concerns. These are spatially explicit databases that contain data for every layer (variable) for each pixel (data point). GIS can contain physiographic data on, for example, weather, elevation, slope, location and distance. In an impact evaluation of protected areas in Thailand, the impact of protected areas (national parks and wildlife sanctuaries) on reducing deforestation might be overstated, after one accounts for the fact that areas that are usually protected are those with low agricultural productivity, and that the likelihood of them being cleared for cultivation is lower than elsewhere (Cropper, Puri, and Griffiths 2001). The study used physiographic attributes such as elevation, slope and location and found that including socio-economic factors such as population density and travel time weighted distance to the market, typically used to explain the opportunity cost of clearing land, did not affect the estimates. Exogenous variation in (easily available) physiographic features thus provided the opportunity to use instrumental variables. Currently, much data are being collected using either satellites or mobile phones, both of which represent cheap and quick methods for collecting rich, spatially disaggregated data that can be used for undertaking impact evaluations without raising ethical concerns.

In our own experience, one important research method for dealing with the complexities of humanitarian emergencies is to adopt an explicit mixed method approach (see for example Bamberger, Rao, and Woolcock 2010). In doing so, researchers do not compensate a weak research design of a theory-based impact evaluation with some qualitative research but rather augment both approaches to drawing from each other's strengths. For example, in a recent study of the effectiveness of a WFP programme to reduce Moderate Acute Malnutrition in young children in Niger, qualitative research helped to uncover likely impact channels that the quantitative method could not determine (Brück et al. 2017).

Another instance where mixed methods can be very helpful is if information bias is a serious concern (see Section 4). Mixed methods can help triangulate results through a variety of information collection techniques, such as satellite imagery, for example, and rely on more objective measures of disaster exposure (for e.g. in the case of floods, depth and duration of floodwaters in the house, rather than subjective assessments of whether or not a household was flooded). Furthermore, impact evaluations can also use other sources of quantitative data to achieve this aim of triangulation such as spatially explicit information, census data and other surveys, as discussed above.

8. Conclusion

This paper assesses how rigorous impact evaluations can help inform and improve interventions in humanitarian emergencies. Given the complexity of humanitarian contexts, the need for speed, the lack of baseline data, the multitude of actors, the requirements of coverage and capacity, and the significant ethical concerns about impact evaluations often expressed, it is usually assumed that theory-based impact evaluation methods cannot be used in such contexts.[16] This helps explain the scarcity of high-quality studies, especially of the relief phase. We argue that the need for learning in the context of humanitarian assistance is vast: one area where impact evaluations can add value without compromising ethical principles relates to the logistics, organisation and content of humanitarian assistance: as discussed above, factorial designs can provide sharp insights about the most effective mode of delivery which is of value to humanitarian organisations, donors and recipients alike.

In contrast to the more stable and gradual changes that are typical of environments that most impact evaluations cover, in humanitarian emergencies lives may be in immediate danger, variable values change fast, and effects may be heterogeneous and varied.

Taking account of and understanding the context is, hence, particularly important. While theory-based impact evaluation has the potential to help generalise lessons, because the analysis will

uncover why something did or did not work and for whom, theory based evaluations of humanitarian emergencies require solid investment in advance preparations.

The paper concludes that data requirements may be less onerous than often suspected. In many cases, researchers may draw upon pre-existing datasets that can help evaluations by providing evidence of balance, as well as providing insights into context. In the absence of baseline data, regression discontinuity designs may provide an attractive methodological option.

Furthermore, while ethical concerns about impact evaluations are valid, they can be sufficiently addressed, making impact evaluations feasible also from an ethical standpoint.

A key lesson is that it pays to be prepared. Much information is being collected these days about the risks of various emergencies unfolding, be they sudden or slow onset emergencies. Hence, national actors and international donors can prepare on three fronts:

(i) they can learn about where emergencies may unfold and where assistance may be required;
(ii) they can plan ahead and be prepared to intervene for when an emergency unfolds (including strengthening local resilience *ex ante*); and
(iii) they can prepare their impact evaluation designs in advance, drawing on the many insights into how to conduct successful impact evaluations offered in this paper and in the emerging literature on this topic.

Being prepared to conduct rigorous impact evaluations also includes building capacity at the national and local levels, and securing buy-in for impact evaluation among donors. Impact evaluations can answer some but not all questions that donors pose. They are less useful for fast learning about how to improve an ongoing intervention, even as implementing an impact evaluation can itself be a valuable learning experience. Yet given the dearth of rigorous causal evidence of what works and what does not work in the humanitarian sector, there is a high dividend to be earned from conducting more impact evaluations in emergency settings, and also much to be lost in not doing so. With a better-informed appreciation of the need, rationale and feasibility of impact evaluation in emergency settings, and with a growing evidence base of methods and techniques employed in such contexts, we expect there to be many more impact evaluations in the humanitarian sector in the years ahead.

Notes

1. Global Humanitarian Assistance. GHA Report 2017. http://www.globalhumanitarianassistance.org/reports.
2. 'Financial Tracking Service (FTS) Tracking Global Humanitarian Aid Flows.' Financial Tracking Service (FTS). UN Office for the Coordination of Humanitarian Affairs. Web. 10 August 2017.
3. See C. Bozzoli, T. Brück and N. Wald (2013) for an exception discussing impact evaluation in conflict settings.
4. The School for a Culture of Peace, Barcelona http://escolapau.uab.cat.
5. Defining Humanitarian Aid | Global Humanitarian Assistance. http://www.globalhumanitarianassistance.org/data-guides/defining-humanitarian-aid.
6. Conflicts may be exacerbated by natural disasters while natural disasters and consequent scarcities often create conflict, see e.g. Collier (2008) and Justino, Brück, and Verwimp (2013).
7. 'Fragility, Conflict and Violence.' *Fragile and Conflict Situations*. World Bank, n.d. Web. http://web.worldbank.org/WBSITE/EXTERNAL/PROJECTS/STRATEGIES/EXTLICUS/0,,contentMDK:22978911~menuPK:4168000~pagePK:64171540~piPK:64171528~theSitePK:511778,00.html.
8. The recent review of the UK's emergency relief placed much emphasis on prevention (HERR 2011).
9. However, pre-disaster administrative or survey data may be present and can be extremely useful.
10. An emergency declaration protocol declares an emergency.
11. This discontinuity may be sharp but can also, for instance in a conflict situation, be fuzzy.
12. Attribution refers to (i) ensuring that a causal pathway runs from a specific intervention or programme to the relevant set of outcomes, and (ii) accurately isolating and estimating the *contribution* of this intervention.
13. As noted above, an emergency declaration protocol declares an emergency. Most humanitarian agencies stay in the affected area for the first 3 or 6 months. For example, the International Rescue Committee (IRC) is constitutionally required to open a country desk if it stays for longer than 6 months in a country.

14. We are grateful to an anonymous reviewer for making this point.
15. This systematic search covered articles published in peer-reviewed academic journals, the ALNAP database, professional and academic networks in ICRC, IRC, Oxfam, WFP, UNICEF, UNCHR, UNOCHA, 3ie's databases of impact evaluations and systematic reviews and 'grey' literature using online search engines. To assess study quality, we used keywords such as 'theory of change', 'theory', 'economic theory', 'power', 'power analysis', 'sample', 'balance', 'test of balance', 'randomization', 'ethics', 'ethical', and other terms. The search of relevant studies was done in English and covered the period from 2001 to 2014. More details are available on request from the authors.
16. For other perceived limitations, see Woolcock (2013) and Basu (2014).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Department for International Development and International Initiative for Impact Evaluation.

Notes on contributors

Jyotsna Puri is currently the Head of the Independent Evaluation Unit of the Green Climate Fund. She is also an adjunct associate professor at the School of International and Public Affairs (SIPA), Columbia University, New York. Dr Puri's areas of work include policy impact analysis on poverty reduction, environment, agriculture, health and climate change, so forth. She has more than 22 years of experience in policy research and development evaluation at several organisations including the World Bank, Columbia University and the UN. Also, she has led evaluation-related work for UNDP, UNICEF, GEF and the MacArthur Foundation. Dr Puri also has extensive policy experience and provides advice on evidence-based climate change and environmental policy, evidence-based advocacy for behaviour change, community-based engagements for improving health and methods for evaluating humanitarian action. Her expertise is using mixed methods that use rigorous quantitative methods informed by high-quality qualitative approaches. Dr Puri's academic qualifications include a PhD and MSc in Resource Economics and a Masters in Development Economics.

Anastasia Aladysheva is a senior researcher at Stockholm International Peace Research Institute (SIPRI) in Sweden. She works with the Peace and Development Programme, focusing on impact evaluations of development and peace-building programmes in fragile states. Her primary research interests include impact evaluations, child health and impact of family programmes on fertility, social capital and its measurement using experimental methods, history and political economy of the Soviet policies, poverty and conflict. She currently works on the impact evaluation projects of peacebuilding programmes in Kyrgyzstan and a panel study Life in Kyrgyzstan. She joined SIPRI in 2013 after completing her PhD in international economics at the Graduate Institute of International Studies and Development in Geneva. Prior that, she studied economics at the Central European University in Budapest, and mathematics and computer programming at the Lomonosov Moscow State University in Russia. She has also worked as a consultant at the World Bank's Development Impact Evaluation Initiative for a community-driven development project in Burkina Faso and as an instructor and teaching and research assistant at the Graduate Institute in Geneva.

Vegard Iversen is a professor (economics area), Indian Institute of Management Ahmedabad and has been a senior research fellow (Hon) at Global Development Institute (former IDPM), University of Manchester since 2008. He conducts applied research on development microeconomics and completed his PhD from University of Cambridge in 2000. He spent 6 years as a tenured Faculty at School of Development Studies (now International Development), University of East Anglia and has been a research fellow at IFPRI, a visiting Faculty at Indian Statistical Institute and a professor and Vice Dean at Jindal School of Government and Public Policy. From 2015–2017, he was an (India-based) adjunct professor at Sanford School of Public Policy at Duke. His research interests span a wide thematic canvas, including affirmative action (political reservations for women from a novel, dynamic angle); persistence and change in caste-based inequality in rural India; the long-term effects of colonial policies on contemporary development; business history; citizen participation as a mechanism to improve public service delivery; social mobility in the Global South. His works were widely published and was awarded the annual Dudley Seers Prize for the best paper published in *Journal of Development Studies* in 2008. He has been the Editor in-Chief for the *Journal of South Asian Development* since 2012 and is on the Editorial Board of *Journal of Development Studies*.

Yashodhan Ghorpade is an economist in the Social Protection and Jobs Unit of the World Bank, Washington DC, working on the Middle East and North Africa region on safety net programmes in complex emergencies and labour market analyses. Yashodhan completed a PhD in Economics from the University of Sussex, UK. In his dissertation, he examined the impact of conflict on households' access to cash transfers and remittances after the 2010 floods in Pakistan. He was awarded the EU Marie Curie Initial Training Network Early Stage Researcher Award for his Doctoral studies (2011–2014), and the DfID Commonwealth Shared Scholarship for his MA degree in Development Economics at the University of East Anglia, UK (2005–2006). He has previously worked as a research fellow at the International Food Policy Research Institute, portfolio leader at Oxford Policy Management in New Delhi, Research Collaborator at the International Labour Organization (child labour programme) and as a consultant at the World Bank offices in India and Myanmar. His research interests include household behaviour under duress (in conflict and natural disasters), migration, impact evaluation and fragility.

Tilman Brück is Team Leader of Economic Development and Food Security at IGZ – Leibniz Institute of Vegetable and Ornamental Crops, and the Founder and Director of ISDC – International Security and Research Center in Berlin, Germany. He is also the Co-Founder and Co-Director of the 'Households in Conflict Network (HiCN)', a global research network studying the intersections of conflict and development at the micro-level, and the Principal Investigator of the Life in Kyrgyzstan Study, an individual-level, nationally representative panel study in Kyrgyzstan. His research interests focus on the economics of household behaviour and well-being in conflict-affected and fragile economies, including the measurement of violence and conflict in household surveys. He has published over 35 articles in peer-reviewed journals (including *Journal of Conflict Resolution*, *Journal of Peace Research*, *Journal of Comparative Economics*, *Economics of Transition*, *European Journal of Political Economy* and *World Development*) and edited over a dozen books and special issues of journals on the economics of conflict and insecurity. Tilman Brück was previously Visiting Professor at the London School of Economics and Political Science, Director of the Stockholm International Peace Research Institute (SIPRI), Professor of Development Economics at Humboldt-University of Berlin and Head of the Department of Development and Security at the German Institute for Economic Research (DIW Berlin). He has also worked as a consultant for the BMZ, European Commission, DFID, FAO, GIZ, ILO, KfW, OECD, UNDP, USAID and the World Bank. He studied economics at the Universities of Glasgow and Oxford University and obtained his doctorate in economics from the University of Oxford.

ORCID

Tilman Brück http://orcid.org/0000-0002-8344-8948

References

Aker, J. C. 2013. "Cash or Coupons? Testing the Impacts of Cash versus Vouchers in the Democratic Republic of Congo." Center for Global Development Working Paper 320.

Aker, J. C., R. Boumnijel, A. McClelland, and N. Tierney. 2011. "Zap It to Me: The Short-Term Impacts of a Mobile Cash Transfer Program." Center for Global Development Working Paper 268.

Angell, M. 1997. "The Ethics of Clinical Research in the Third World." *The New England Journal of Medicine* 337 (12): 847–849. doi:10.1056/NEJM199709183371209.

Annan, J., S. M. Murray, D. Kaysen, S. Griffiths, T. Cetinoglu, K. Wachter, L. K. Murray, and P. A. Bolton. 2014. "Controlled Trial of Psychotherapy for Congolese Survivors of Sexual Violence." *The New England Journal of Medicine* 368 (2): 182–2,191.

Annan, J., T. Bundervoet, J. Seban, and J. Costigan. 2013. *Urwaruka Rushasha: A Randomized Impact Evaluation of Village Savings and Loans Associations and Family-Based Interventions in Burundi*. Washington, DC: IRC.

Bamberger, M., V. Rao, and M. Woolcock. 2010. "Using Mixed Methods in Monitoring and Evaluation: Experiences in International Development." Policy Research Paper 5245, The World Bank.

Barnett, M. 2005. "Humanitarianism Transformed." *Perspectives on Politics* 3 (4): 723–740. doi:10.1017/S1537592705050401.

Barron, I. G., G. Abdallah, and P. Smith. 2013. "Randomized Control Trial of a CBT Trauma Recovery Program in Palestinian Schools." *Journal of Loss and Trauma* 18 (4): 306–321. doi:10.1080/15325024.2012.688712.

Bass, J., R. Neugembauer, K. F. Clougherty, H. Verdeli, P. Wickramaratne, L. Ndogoni, L. Speelman, M. Weissman, and P. Bolton. 2006. "Group Interpersonal Psychotherapy for Depression in Rural Uganda: 6-Month Outcomes." *The British Journal of Psychiatry* 188 (6): 567–573. doi:10.1192/bjp.188.6.567.

Basu, K. 2014. "Randomisation, Causality and the Role of Reasoned Intuition." *Oxford Development Studies* 42 (4): 455–472. doi:10.1080/13600818.2014.961414.

Bolton, P., J. Bass, T. Betancourt, L. Speelman, G. Onyango, K. F. Clougherty, R. Neugebauer, L. Murray, and H. Verdeli. 2007. "Interventions for Depression Symptoms among Adolescent Survivors of War and Displacement in Northern Uganda: A Randomized Controlled Trial." *JAMA* 298 (5): 519–527. doi:10.1001/jama.298.5.519.

Bozzoli, C., T. Brück, and N. Wald. 2013. "Evaluating Programmes in Conflict-Affected Areas." In *A Micro-Level Perspective on the Dynamics of Conflict, Violence and Development*, edited by P. Justino, T. Brück, and P. Verwimp. Oxford: Oxford University Press.

Brück, T., N. T. N. Ferguson, J. Ouédraogo, and Z. Ziegelhöfer. 2017. An Impact Evaluation of WFP Malnutrition Interventions in Niger. Berlin: ISDC – International Security and Development Center, unpublished report.

Buttenheim, A. 2009. "Impact Evaluation in the Post-Disaster Setting: A Conceptual Discussion in the Context of the 2005 Pakistan Earthquake. International Initiative for Impact Evaluation (3ie)." Working Paper no. 5.

Cameron, D. B., A. N. Brown, A. Mishra, M. Picon, H. Esper, F. Calvo, and K. Peterson. 2015. "Evidence for Peacebuilding: An Evidence Gap Map." 3ie scoping paper.

Catani, C., M. Kohiladevy, M. Ruf, E. Schauer, T. Elbert, and F. Neuner. 2009. "Treating Children Traumatized by War and Tsunami: A Comparison between Exposure Therapy and Meditation-Relaxation in North-East Sri Lanka." *BMC Psychiatry* 9 (22): 1–11. doi:10.1186/1471-244X-9-22.

Collier, P. 2008. *The Bottom Billion*. Oxford: Oxford University Press.

Cosgrave, J. 2007. *Synthesis Report: Expanded Summary. Joint Evaluation of the International Response to the Indian Ocean Tsunami*. London: Tsunami Evaluation Coalition.

Cropper, M., J. Puri, and C. Griffiths. 2001. "How the Location of Roads and Protected Areas Affects Deforestation in North Thailand." Policy, Research working papers; no. WPS 2583. Washington, DC: World Bank.

Crost, B., J. Felter, and P. B. Johnston. 2014. *Conditional Cash Transfers, Civil Conflict and Insurgent Influence: Experimental Evidence from the Philippines*. Sussex: HiCN.

De Mel, S., D. McKenzie, and C. Woodruff. 2012. "Enterprise Recovery following Natural Disasters." *The Economic Journal* 122 (559): 64–91. doi:10.1111/ecoj.2012.122.issue-559.

Devereux, S. 2000. "Famine in the 21st Century." IDS Working Paper. Brighton

Dilley, M., R. S. Chen, U. Deichmann, A. L. Lerner-Lam, M. Arnold, J. Agwe, P. Buys, O. Kjekstad, B. Lyon, and G. Yetman. 2005. *Natural Disaster Hotspots: A Global Risk Analysis*. Disaster Risk Management Series 5. Publication no. 34423. Washington, DC: World Bank, Hazard Management Unit.

Doocy, S., and G. Burnham. 2006. "Point-Of-Use Water Treatment and Diarrhea Reduction in the Emergency Context: An Effectiveness Trial in Liberia." *Tropical Medicine & International Health* 11 (10): 1542–1552. doi:10.1111/j.1365-3156.2006.01704.x.

Ertl, V., A. Pfeiffer, E. Schauer, T. Elbert, and F. 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): 503–512. doi:10.1001/jama.2011.1060.

Few, R., D. McAvoy, M. Tarazona, and V. M. Walden. 2014. *Contribution to Change: An Approach to Evaluating the Role of Intervention in Disaster Recovery*. Oxford: Practical Action Publishing.

Gaarder, M., and J. Annan. 2013. "Impact Evaluation of Conflict Prevention and Peacebuilding Interventions." World Bank Policy Research Working Paper, No 6496.

Guiteras, R., A. Jina, and A. M. Mobarak. 2015. "Satellites, Self-Reports, and Submersion: Exposure to Floods in Bangladesh." Maryland Population Research Center Working Paper PWP-MPRC-2015-001, and IGC Working Paper.

Hidrobo, M., J. Hoddinott, A. Peterman, A. Margolies, and V. Moreira. 2014. "Cash, Food, or Vouchers? Evidence from a Randomized Experiment in Northern Ecuador." *Journal of Development Economics* 107: 144–156. doi:10.1016/j.jdeveco.2013.11.009.

Hossain, M., C. Zimmerman, L. Kiss, T. Abramsky, D. Kone, M. Bakayoko-Topolska, J. Annan, H. Lehmann, and C. Watts. 2014. "Working with Men to Prevent Intimate Partner Violence in a Conflict-Affected Setting: A Pilot Cluster Randomized Controlled Trial in Rural CôTe d'Ivoire." *BMC Public Health* 14 (339): 1–13. doi:10.1186/1471-2458-14-339.

Humanitarian Emergency Response Review. 2011. "UK Government." March 28.

Hussain, A. 1995. "Introduction." In *The Political Economy of Hunger*, edited by J. Dreze, A. Sen, and A. Hussein. Oxford: Oxford Clarendon Press.

Huybregts, L., F. Houngbé, C. Salpéteur, R. Brown, D. Roberfroid, M. Ait-Aissa, and P. Kolsteren. 2012. "The Effect of Adding Ready-To-Use Supplementary Food to a General Food Distribution on Child Nutritional Status and Morbidity: A Cluster-Randomized Controlled Trial." *PLoS Medicine* 9 (9): e1001313. doi:10.1371/journal.pmed.1001313.

International Rescue Committee. 2008. *Getting down to Business: Women's Economic and Social Empowerment in Burundi*. New York: IRC.

International Rescue Committee. 2012. *Measuring Impact: Survivors' Social, Psychological and Economic Well-Being*. New York: IRC.

IRC. 2011. Urwaruka Rushasha: A Randomized Impact Evaluation of Village Savings and Loans Associations and Family-Based Interventions in Burundi.

Jordans, M. J. D., I. H. Komproe, W. A. Tol, B. A. Kohrt, N. P. Luitel, R. D. Macy, and J. T. V. M. De Jong. 2010. "Evaluation of a Classroom-Based Psychosocial Intervention in Conflict-Affected Nepal: A Cluster Randomized Controlled Trial." *Journal of Child Psychology and Psychiatry* 51 (7): 818–826. doi:10.1111/j.1469-7610.2010.02209.x.

Justino, P., T. Brück, and P. Verwimp. 2013. "Micro-Level Dynamics of Conflict, Violence and Development: A New Analytical Framework." In *A Micro-Level Perspective on the Dynamics of Conflict, Violence and Development*, edited by P. Justino, T. BrüCk, and P. Verwimp. Oxford: Oxford University Press.

Kondylis, F. 2008. "Agricultural Outputs and Conflict Displacement: Evidence from a Policy Intervention in Rwanda." *Economic Development and Cultural Change* 57 (1): 31–66. doi:10.1086/590446.

Lehmann, C., and D. Masterson. 2014. *Emergency Economies: The Impact of Cash Assistance in Lebanon*. Beirut: IRC.

Levely, I. 2014. "Measuring Intermediate Outcomes of Liberia's Disarmament, Demobilization, Rehabilitation and Reintegration Program." *Defence and Peace Economics* 25 (2): 139–162. doi:10.1080/10242694.2012.727065.

Malhotra, D., and S. Liyanage. 2005. "Long-Term Effects of Peace Workshops in Protracted Conflicts." *Journal of Conflict Resolution* 49 (6): 908–924. doi:10.1177/0022002705281153.

Menon, N. V. C. 2013. Disaster Management in India: Past, Present and Future.

Mvukiyehe, E., and C. Samii. 2010. Quantitative Impact Evaluation of the United Nations Mission in Liberia: Final Report. Unpublished. SSRN eLibrary.

Nielsen, N. S., F. Jahan, and C. Canteli. 2012. "The Contribution of Food Assistance to Durable Solutions in Protracted Refugee Situations; Its Impact and Role in Bangladesh: A Mixed Method Impact Evaluation." A series of UNHCR-WFP Impact Evaluations.

O'Callaghan, P., J. McMullen, C. Shannon, H. Rafferty, and A. 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 & Adolescent Psychiatry* 52 (4): 359–369. doi:10.1016/j.jaac.2013.01.013.

Park, A., and S. Wang. 2017. "Benefiting from Disaster? Public and Private Responses to the Wenchuan Earthquake." *World Development* 94: 38–50. doi:10.1016/j.worlddev.2016.12.038.

Puri, J., A. Aladysheva, V. Iversen, Y. Ghorpade, and T. Brück. 2014. *What methods may be used in impact evaluations of humanitarian assistance?* 3ie Working Paper 22. New Delhi: International Initiative for Impact Evaluation (3ie). doi:10.23846/WP0022

Roberts, L., Y. Chartier, O. Chartier, G. Malenga, M. Toole, and H. Rodka. 2001. "Keeping Clean Water Clean in a Malawi Refugee Camp: A Randomized Intervention Trial." *Bulletin of the World Health Organization* 79 (4): 280–287.

Samii, C., A. Brown, and M. Kulma. 2011. "Evaluating Stabilization Interventions." Working Paper, International Initiative for Impact Evaluation (3ie) and New York University.

Scholte, W. F., F. Verduin, A. M. Kamperman, T. Rutayisire, A. H. Zwinderman, and K. Stronks. 2011. "The Effect on Mental Health of a Large Scale Psychosocial Intervention for Survivors of Mass Violence: A Quasi-Experimental Study in Rwanda." *PLoS ONE* 6 (8): e21819. doi:10.1371/journal.pone.0021819.

Schwab, B., A. Margolies, and J. Hoddinott. 2013. Impact Evaluation of Cash and Food Transfers for the Seasonal Emergency Safety Net in Hajjah and Ibb Governorates: Yemen Endline Report. Washington, DC: International Food Policy Research Institute.

Shoji, M. 2010. "Does Contingent Repayment in Microfinance Help the Poor during Natural Disasters?" *The Journal of Development Studies* 46 (2): 191–210. doi:10.1080/00220380902952381.

Sim, A., J. Annan, E. Puffer, C. Salhi, and T. Betancourt. 2014. *Building Happy Families: Impact Evaluation of a Parenting and Family Skills Intervention for Migrant and Displaced Burmese Families in Thailand*. New York: IRC.

The Belmont Report : Ethical Principles and Guidelines for the Protection of Human Subjects of Research. 1979. United States, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Torrente, C., J. L. Aber, D. Witteveen, T. Gupta, B. Johnston, and A. Shivshanker. 2011. *Opportunities for Equitable Access to Quality Basic Education (OPEQ)*. New York: IRC.

Wisner, B., P. Blaikie, T. Cannon, and I. Davis. 1994. *At Risk: Natural Hazards, People's Vulnerability and Disasters*. Routledge: London and New York. doi:10.4324/9780203428764.

WMA's Declaration of Helsinki Serves as Guide to Physicians. JAMA. 1964; 189(13): 33-34. doi: 10.1001/jama.1964.03070130073046

Woolcock, M. 2013. "Using Case Studies to Explore the External Validity of 'Complex' Development Interventions." *Evaluation* 19 (3): 229–248. doi:10.1177/1356389013495210.

World Bank. 2010. *Natural Hazards – Unnatural Disasters. The Economics of Effective Prevention*. Washington, DC: World Bank.

Appendix

Table A1. Humanitarian relief.

No.	Study	Method, counterfactual and rigour	Main findings
1	Nielsen, Jahan, and Canteli (2012) – The contribution of food assistance to durable solutions in protracted refugee situations; its impact and role in Bangladesh: A mixed method impact evaluation *Country*: Bangladesh *Category*: food assistance *Disaster*: conflict	Method: Quasi-experimental *Counterfactual*: Natural control group *Rigour*: ToC, PA	• Improved dietary diversity and reduced frequency of negative coping strategies; • Positive impact on economic activity; • Improved self-reliance and security
2	Hidrobo et al. (2014) – Cash, food, or vouchers? Evidence from a randomised experiment in northern Ecuador *Country*: Ecuador *Category*: cash vs. food *Disaster*: conflict and poverty	*Method*: RCT *Counterfactual*: Random Assignment *Rigour*: –	• Value of per capita food consumption increased by 13 per cent, per capita caloric intake increased by 10 per cent, HDDS improved by 5.1 per cent, dietary diversity index (DDI) by 14.4 per cent, and food consumption score (FCS) by 12.6 per cent; • Vouchers lead to the largest gains in dietary diversity and food leads to the largest increase in caloric intake; • Did not lead to a significant change in haemoglobin levels or anaemia classifications (negative effects on food group); • Discrimination decreased by 6 percentage points and participation in groups increased by 6 percentage points, decrease in IPV
3	Huybregts et al. (2012) – The Effect of Adding Ready-to-Use Supplementary Food to a General Food Distribution on Child Nutritional Status and Morbidity: A Cluster-Randomised Controlled Trial *Country*: Chad *Category*: food assistance *Disaster*: conflict	*Method*: RCT *Counterfactual*: Random assignment with factorial model *Rigour*: PA, ToC, ToB, E	• Reduction in cumulative incidence of wasting (incidence risk ratio: 0.86); • Lower gain in height-for-age (+0.03 Z-score/mo); • Higher haemoglobin concentration (+3.8 g/l), thereby reducing the odds of anaemia (odds ratio: 0.52); • Lower risk of self-reported diarrhoea (229.3%) and fever episodes (222.5%).
4	Doocy and Burnham (2006) – Point-of-use water treatment and diarrhea reduction in the emergency context: An effectiveness trial in Liberia *Country*: Liberia *Category*: water cleaning *Disaster*: conflict	*Method*: RCT *Counterfactual*: Random assignment with factorial model *Rigour*: PA, ToC, ToB, E	• Improved storage reduced diarrhoea incidence by (90%) and prevalence by (83%), when compared with control households with improved water storage alone; • Among the intervention group, residual chlorine levels met or exceeded Sphere standards in (85%) of observations with a (95%) compliance rate

(*Continued*)

Table A1. (Continued).

No.	Study	Method, counterfactual and rigour	Main findings
5	Roberts et al. (2001) – Keeping clean water clean in a Malawi refugee camp: A randomised intervention trial *Country*: Malawi *Category*: water cleaning *Disaster*: conflict and poverty	*Method*: RCT *Counterfactual*: Random assignment *Rigour*: ToC, E	• (31%) less diarrhoeal disease in children under 5 years of age among the group using the improved bucket; • Less water contamination; • Proper chlorination is a less expensive and more effective means of water quality protection in comparison with the improved bucket, but was unpopular and rarely utilised by the camp inhabitants
6	Bolton et al. (2007) – Interventions for Depression Symptoms Among Adolescent Survivors of War and Displacement in Northern Uganda *Country*: Uganda *Category*: reconciliation *Disaster*: conflict	*Method*: RCT *Counterfactual*: Delayed treatment control group *Rigour*: PA, ToC, ToB	• Difference in change in adjusted mean score for depression symptoms between group interpersonal psychotherapy and control groups was 9.79 points; • Girls receiving group interpersonal psychotherapy showed substantial and significant improvement in depression symptoms compared with controls (12.61 points), and improvement among boys was not statistically significant (5.72 points); • Creative play showed no effect on depression severity (−2.51 points); • No statistically different improvements in anxiety and conduct problems or function scores in either intervention group
7	Aker (2013) – Cash or Coupons? Testing the Impacts of Cash versus Vouchers in the Democratic Republic of Congo. *Country*: DRC *Category*: vouchers vs. cash *Disaster*: conflict and poverty	*Method*: RCT *Counterfactual*: Random assignment (two arm treatment with no pure control group) *Rigour*: ToB, ToC	• Given that the value of the voucher was less than the household's normal total food expenditures, voucher households purchased more food items than they would have had they been provided with cash. • There was also no evidence that cash-transfer households were using their transfers for temptation goods.
8	Lehmann and Masterson (2014) – Emergency Economies: The Impact of Cash Assistance in Lebanon. *Country*: Lebanon *Category*: cash assistance *Disaster*: conflict	*Method*: Quasi-experimental *Counterfactual*: Matched Control Group *Rigour*: PA, ToB, ToC	• Cash assistance for beneficiaries at high altitude during winter months increased spending on heating supplies, however, almost half of beneficiaries report that heating supplies were often not enough to keep warm. • The programme did not have a meaningful impact on prices in local market. • Cash assistance increased access to school and reduced child labour.
9	Schwab, Margolies, and Hoddinott (2013) – Impact Evaluation of Cash and Food Transfers for the Seasonal Emergency Safety Net in Hajjah and Ibb Governorates: Yemen Endline Report. *Country*: Yemen *Category*: cash versus food transfers *Disaster*: conflict	*Method*: RCT *Counterfactual*: Random assignment (two-arm treatment with no pure control group) *Rigour*: ToB	• Cash transfers raised dietary diversity and quality more highly than food and were cheaper to deliver and administer • However, food beneficiaries consumed more calories overall

Table A2. (Un) anticipated natural disasters.

No.	Study	Method, counterfactual and rigour	Main findings
1	De Mel et al. (2012) – Enterprise recovery following natural disasters *Country*: Sri Lanka *Category*: Entrepreneurship *Disaster*: tsunami	*Method*: RCT *Counterfactual*: Delayed treatment control group *Rigour*: ToC, ToB	• Positive effect of grant programme on profits, representing a 9.9 per cent real monthly return on the treatment; • Direct aid is more important in the recovery of enterprises operating in the retail sector than for those operating in the manufacturing and service sectors; • The use of cash grants is more helpful than the use of in-kind, but only in limited cases
2	Shoji (2010) – Does contingent repayment in microfinance help the poor during natural disasters? *Country*: Bangladesh *Category*: Microfinance *Disaster*: floods	*Method*: Quasi-experimental *Counterfactual*: Before/after comparison *Rigour*: ToC	• Decreasing probability that people skip meals during negative shocks by 5.1 per cent, with a higher effect on landless and females; • The authors did not estimate the effects on nutritional outcomes, and no conclusions could be made about whether these households are better off nutritionally
3	Aker, Boumnijel, McClelland, and Tierney (2011) – Zap it to me: The short-term impacts of a mobile cash transfer programme *Country*: Niger *Category*: Cash transfer and information technology *Disaster*: drought	*Method*: RCT *Counterfactual*: Matched control group *Rigour*: ToB	• Mobile-phone based money transfer reduced the variable distribution costs to the implementing agency and the costs to the programme recipients in accessing their cash transfer. • Additional benefits for the programme recipients: households in *zap* (mobile money transfer system) villages spent the cash transfer on more types of items, sold fewer non-durable assets, consumed more diverse foods and cultivated more diverse crops as compared to those receiving a manual cash transfer.

Table A3. Peacebuilding

No	Study	Method, counterfactual and rigour	Main findings
1[b]	Kondylis (2008) – Agricultural Outputs and Conflict Displacement: Evidence from a Policy Intervention in Rwanda *Country*: Rwanda *Category*: Victims of War	*Method*: Quasi-experimental (DID) *Counterfactual*: Natural control group *Rigour*: ToC	• Returns to on-farm labour are higher for returnees (refugees returning to settlement) relative to stayers, although the evidence suggests that the policy contributed little additional effect to this differential; • These differentials suggest that, upon return from conflict-induced exile, returnees are more motivated to increase their economic performance
2	Levely (2014) – Measuring Intermediate Outcomes of Liberia's Disarmament, Demobilisation, Rehabilitation and Reintegration Programme *Country*: Liberia *Category*: Ex-Combatant Reintegration	*Method*: Quasi-experimental (PSM) *Counterfactual*: Matched Control Group *Rigour*: -	• A higher employment rate for those who complete the programme, although there is consistently no effect on income
3	Malhotra and Liyanage (2005) – Long-Term Effects of Peace Workshops in Protracted Conflicts *Country*: Sri Lanka *Category*: Consensus & Dialogue	*Method*: Quasi-experimental *Counterfactual*: Natural control group *Rigour*: ToC	• Compared with two control groups, the participant group showed greater empathy towards members of the other ethnicity, even one year after participation in the peace workshops; • Consistent with the attitudinal data on empathy, participants donated more money to help poor children of the other ethnicity than did nonparticipants
4	Mvukiyehe and Samii (2010) – Quantitative Impact Evaluation of the United Nations Mission in Liberia: Final report *Country*: Liberia *Category*: Ex-Combatant Reintegration, Peace Dividends	*Method*: Quasi-experimental *Counterfactual*: Cluster matched sampling *Rigour*: ToB, ToC	• Humanitarian community can contribute to consolidating the peace in Liberia by supporting the reintegration of newly resettled households; • Supporting efforts to foster social and community cohesion, especially among the newly resettled households; and • Providing electoral assistance to sustain political interest among ordinary citizens
5[a]	Torrente et al. (2011) – Opportunities for Equitable Access to Quality Basic Education (OPEQ) *Country*: DRC *Category*: Victims of War (children)	*Method*: RCT *Counterfactual*: Delayed treatment control group *Rigour*: ToC	*Only baseline report available*
6	IRC (2012) – Measuring Impact: Survivors' Social, Psychological and Economic Well-Being *Country*: DRC *Category*: Victims of War	*Method*: RCT *Counterfactual*: Random assignment with factorial model *Rigour*: -	*No final results available.* *Preliminary findings*: • While mental health symptoms decreased over time in both groups, the decrease in overall mental health symptoms was significantly greater for VSLA (Village Savings and Loan Associations) group members compared to control; • No significant difference in ability to complete daily tasks, e.g. farming or caring for children, between VSLA members and control, though both groups reported less difficulty over time

(*Continued*)

Table A3. (Continued).

No	Study	Method, counterfactual and rigour	Main findings
7[b]	IRC (2008) – Getting down to business: Women's economic and social empowerment in Burundi *Country*: Burundi *Category*: Victims of War (women)	*Method*: RCT *Counterfactual*: Random assignment with factorial model *Rigour*: –	*Preliminary results*: • Incidence of IPV decreased; • Women reported increased decision-making; • Use of negotiation skills increased; • Acceptance of violence decreased
8	Annan et al. (2013) – Urwaruka Rushasha: A Randomised Impact Evaluation of Village Savings and Loans Associations and Family-Based Interventions in Burundi *Country*: Burundi *Category*: Victims of War	*Method*: RCT *Counterfactual*: Random assignment with factorial model and delayed treatment control group *Rigour*: PA, ToC, ToB	• Increased assets and consumption; • Decreased harsh physical and verbal discipline in the home, improved communication between children and caregivers, and a decrease in family problems, including violence and intoxication of family members
9	Annan et al. (2014) – Controlled trial of psychotherapy for Congolese survivors of sexual violence. *Country*: DRC *Category*: Victims of War	*Method*: RCT *Counterfactual*: Random assignment *Rigour*: ToC, ToB	• Mean scores for combined depression and anxiety improved in the individual-support group, but improvements were significantly greater in the therapy group. • Similar patterns were observed for PTSD and functional impairment.
10	Barron, Abdallah, and Smith (2013) – Randomised control trial of a CBT trauma recovery programme in Palestinian *Country*: Palestine *Category*: Victims of War	*Method*: RCT *Counterfactual*: Random assignment *Rigour*: ToB	• Teaching Recovery Techniques (TRT) trauma programme has the potential to ameliorate children's trauma symptoms during situations of ongoing violence.
11	Bass et al. (2006) – Group interpersonal psychotherapy for depression in rural Uganda: 6-month outcomes. *Country*: Uganda *Category*: Victims of War	*Method*: RCT *Counterfactual*: Random assignment *Rigour*: ToB	• Participants who received treatment had mean depression symptom and functional impairment scores respectively 14.0 points and 5.0 points lower than the control group. • The rate of major depression among the treated population was significantly lower than that of the control group.
12	Bolton et al. (2007) – Interventions for depression symptoms among adolescent survivors of war and displacement in northern Uganda: A randomised controlled trial. *Country*: Uganda *Category*: Victims of War (children)	*Method*: RCT *Counterfactual*: Random assignment *Rigour*: ToB	• Group interpersonal psychotherapy was effective for depression symptoms among adolescent girls affected by war and displacement; • While improvement for boys was not statistically significant.

(*Continued*)

Table A3. (Continued).

No	Study	Method, counterfactual and rigour	Main findings
13	Catani et al. (2009) – Treating children traumatised by war and tsunami: A comparison between exposure therapy and meditation-relaxation in North-East Sri Lanka *Country*: Sri Lanka *Category*: Victims of War and Tsunami (children)	*Method*: RCT *Counterfactual*: Random assignment *Rigour*: -	• In both treatment conditions, PTSD (post-traumatic stress disorder) symptoms and impairment in functioning were significantly reduced at one month post-test and remained stable over time; • There was no significant difference between the two therapy groups in any outcome measure.
14	Crost et al. (2014) – Conditional cash transfers, civil conflict and insurgent influence: Experimental evidence from the Philippines *Country*: Philippines *Category*: Victims of War	*Method*: RCT *Counterfactual*: random assignment *Rigour*: ToB	• The programme caused a considerable reduction in the number of conflict-related incidents in the programme area. • The programme reduced insurgent influence in treated villages.
15	Ertl et al. (2011) – Community-implemented trauma therapy for former child soldiers in northern Uganda: A randomised controlled trial. *Country*: Uganda *Category*: Victims of War (children)	*Method*: RCT *Counterfactual*: random assignment *Rigour*: ToB	• PTSD (post-traumatic stress disorder) symptom severity was considerable improved in the narrative exposure therapy group than in the academic catch-up programme and waiting-list groups.
16	Hossain et al. (2014) – Working with men to prevent intimate partner violence in a conflict-affected setting: A pilot cluster randomised controlled trial in rural Côte d'Ivoire *Country*: Côte d'Ivoire *Category*: Intimate Partner Violence (IPV) against women	*Method*: RCT *Counterfactual*: random assignment *Rigour*: ToB	• Significant differences were found between men in the treatment and control groups' reported ability to control their hostility and manage conflict and participation in gendered household tasks.
17	Jordans et al. (2010) – Evaluation of a classroom-based psychosocial intervention in conflict-affected Nepal: A cluster randomised controlled trial *Country*: Nepal *Category*: Victims in War (children)	*Method*: RCT *Counterfactual*: random assignment *Rigour*: ToB	• The intervention reduced psychological difficulties and aggression among boys, increased prosocial behaviour among girls, and increased hope for older children. • The intervention did not result in reduction of psychiatric symptoms.
18	O'Callaghan et al. (2013) – A randomised controlled trial of trauma-focused cognitive behavioural therapy for sexually exploited, war- affected Congolese girls. *Country*: DRC *Category*: Victims in War (girls)	*Method*: RCT *Counterfactual*: random assignment *Rigour*: ToB	• The FT.-CBT (trauma-focused cognitive behavioural therapy) group experienced significantly greater reductions in trauma symptoms • Plus, the TF-CBT group showed a highly significant improvement in symptoms of depression and anxiety, conduct problems, and prosocial behaviour.

(*Continued*)

Table A3. (Continued).

No	Study	Method, counterfactual and rigour	Main findings
19	Scholte et al. (2011) – The effect on mental health of a large scale psychosocial intervention for survivors of mass violence: A quasi-experimental study in Rwanda. *Country*: Rwanda *Category*: Victims of War	*Method*: Quasi-experimental study (DID) *Counterfactual*: matched control group *Rigour*: ToB, ToC, E	• A large scale psychosocial intervention caused a lasting improvement of mental health in survivors of mass violence in Rwanda.
20	Sim et al. (2014) – Building happy families: Impact evaluation of a parenting and family skills intervention for migrant and displaced Burmese families in Thailand. *Country*: Thailand *Category*: Child Protection	*Method*: RCT *Counterfactual*: random assignment *Rigour*: ToB	• A significant impact was found on parenting practices, family functioning, and child behaviour. • Results showed some effects on harsh punishment and child psychosocial wellbeing, as reported by caregivers or children

[a]Ongoing studies, [b]preliminary results, PA: power analysis; ToC: theory of change; ToB: test of balance; E: ethics.