WPS6496

Policy Research Working Paper 6496

Impact Evaluation of Conflict Prevention and Peacebuilding Interventions

Marie Gaarder
Jeannie Annan

The World Bank
Independent Evaluation Group
Public Sector Evaluation Department
June 2013

Policy Research Working Paper 6496

Abstract

The international community is paying increased attention to the 25 percent of the world's population that lives in fragile and conflict affected settings, acknowledging that these settings represent daunting development challenges. To deliver better results on the ground, it is necessary to improve the understanding of the impacts and effectiveness of development interventions operating in contexts of conflict and fragility. This paper argues that it is both possible and important to carry out impact evaluations even in settings of violent conflict, and it presents some examples from a collection of impact evaluations of conflict prevention and peacebuilding interventions. The paper examines the practices of impact evaluators in the peacebuilding sector to see how they address evaluation design, data collection, and conflict analysis. Finally, it argues that such evaluations are crucial for testing assumptions about how development interventions affect change—the so-called "theory of change"—which is important for understanding the results on the ground.

This paper is a product of the Public Sector Evaluation Department, Independent Evaluation Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at mgaarder@worldbank.org.

Impact evaluation of conflict prevention and peacebuilding interventions

Marie Gaarder and Jeannie Annan[1]

Keywords: Impact evaluation, conflict prevention, peacebuilding, conflict analysis, fragile states, ex-combatants, experimental design, quasi-experimental design, encouragement design

JEL Classification: C93, D03, D74, H43, O22

Sector Board: Social Protection

---

[1] Gaarder: Independent Evaluation Group (IEG), World Bank, mgaarder@worldbank.org; Annan, International Rescue Committee, jeannie.annan@rescue.org. We would like to thank a number of people for their support. Ole Winckler Andersen, Denmark's development cooperation (Danida), Beate Bull, Norwegian Agency for development cooperation (Norad) and Megan Kennedy-Chouane, DAC Network on Development Evaluation, editors of the forthcoming book *Evaluation Methodologies for Aid in Conflict*, have been a driving force behind this paper (an updated version of which will appear as a book chapter). Christoph Zürcher, Professor at University of Ottawa, Cyrus Samii, Assistant Professor at New York University, Macartan Humphreys, Professor at Columbia University and Julia Leininger, German Development Institute have reviewed and provided invaluable comments to draft versions of the paper. Anette Brown, Deputy Director and Daniela Barba, Research Assistant at the International Initiative for Impact Evaluation (3ie) kindly gave us access to the database and literature survey which formed the basis for the Samii, Brown, and Kulma (2012) survey. Finally, a number of the lead investigators of the impact evaluations discussed in the paper took the time to respond to our survey and we are very grateful for being able to draw on their experiences.

1. Introduction

The international community is paying increased attention to the 25 percent of the world's population that lives in fragile and conflict affected settings, acknowledging that these settings represent daunting development challenges. While rising levels of resources go into these contexts, results have proven to be difficult to achieve and sustain -- no fragile state has yet to reach any of the MDGs (OECD, 2012; WDR, 2011). To deliver better results on the ground, it is necessary to improve the understanding of the impacts and effectiveness of development interventions operating in contexts of conflict and fragility. While impact evaluations are increasingly used as a tool to establish what works, why and under what circumstances in a variety of development sectors, doubts have been voiced as to the feasibility and desirability of carrying out impact evaluation —evaluation that accounts for the counterfactual in order to attribute impact—in these situations. Some evaluators and practitioners in this field raise four main concerns: (i) it is unethical to identify a comparison group in situations of conflict and fragility; (ii) it is too operationally difficult to do so; (iii) impact evaluations do not address the most important evaluation questions; and (iv) they are too costly. This paper argues that it is both possible and important to carry out impact evaluations even in settings of violent conflict, and it presents some examples from a collection of impact evaluations of conflict prevention and peacebuilding interventions. The paper examines the practices of impact evaluators in the peacebuilding sector to see how they address evaluation design, data collection, and conflict analysis. Finally, it argues that such evaluations are crucial for testing assumptions about how development interventions affect change—the so-called "theory of change"—which is important for understanding the results on the ground.

2. Defining impact evaluations

Impact evaluation, as defined in this paper, refers to evaluations that draw from a set of methods designed to establish a counterfactual or valid comparison, to the intervention in question. The objective is to measure the net impact of the intervention, which in theory is the difference in outcomes for those receiving the intervention compared to what the outcomes would be for the same participants without the intervention. Since it is not possible to measure this difference in practice, impact evaluation methods are all designed to create a comparison group that resembles the participant group as closely as possible. This methodology can be used to explore attribution at any level throughout the results matrix, be it outputs or short- and long term outcomes and impacts.

Impact evaluation methods include experiments—randomized controlled trials (RCTs) in which subjects are randomized to receive a certain version of an intervention (the 'treatment' as it is known from medical trials) or not, and cluster randomized controlled trials in which groups of subjects (as opposed to individual subjects, such as schools, villages or households) are randomized —and "quasi-experiments". The key difference between experiments and quasi-experiments is the use of random assignment of the target population to treatment or control in experiments. Instead of random assignment, quasi-experimental designs typically allow the researcher to establish a comparison group that is as similar as possible to the intervention group using either program eligibility criteria or statistical methods to control for confounding variables. One important practical implication of this difference is that randomization requires planning and starting the evaluation prior to the initiation of the intervention for which we want to measure impact, as units of assignment can only be randomized ex ante, which may not always be politically feasible or operationally realistic. Quasi-experimental approaches do not have this specific limitation but suffer from other shortcomings, such as potential selection bias due to differences in time-varying unobservable characteristics (Baker, 2000; White, 2011; World Bank, 2013).

The impact evaluations we discuss in the first part of this paper are large *n* impact evaluations, meaning that the design is based on data collected over a large sample, usually of individuals. In the second part, we will briefly address the subject of small *n* attribution analysis, and what it may imply for evaluations in conflict prevention and peacebuilding.

On its own, a large *n* impact evaluation explores the effect (or lack thereof) of a certain intervention or activity. The counterfactual quantitative analysis of impact should be supplemented by factual analysis, such as of program beneficiary targeting effectiveness, implementation and process documentations, and qualitative data (which can be derived from a large variety of methods) in order to help develop the initial theory of change, dissect the differences in findings between different settings and to further understand why the results were what they were. The importance of including both counterfactual and factual analysis is exemplified by the International Initiative for Impact Evaluation's (3ie) requirement of the use of mixed methods for the evaluations they fund (White, 2009).

3. Are impact evaluations of interventions in conflict prevention and peacebuilding feasible?

Conflict-affected settings make conducting impact evaluations challenging. To address the objections that impact evaluation of peacebuilding interventions cannot be done, and to document the type of methodologies that have been more prominently used, Samii, Brown, and Kulma (2012) conducted a thorough literature search to identify all the impact evaluations that have been conducted (and including some ongoing studies) of what they call stabilization interventions. Their search and review covers impact evaluations of peacebuilding/stabilization interventions by any donor or government. They found that there are roughly two dozen impact evaluations, some ongoing, across seven categories of stabilization interventions. While the search did not fulfill all the criteria to qualify as a systematic review according to the Campbell Collaboration guidelines[2], it was extensive, covering multiple databases as well as direct contact with the researchers to identify ongoing studies. The largest number of impact evaluations has been of ex-combatant reintegration programs and of peace dividends (community-driven reconstruction) programs. The impact evaluations they found were conducted in Afghanistan, Burundi, Democratic Republic of Congo, the Aceh region of Indonesia, Israel and Palestine, Liberia, Rwanda, Sierra Leone, Sri Lanka, and Northern Uganda.[3]

The Samii et al. search results demonstrate that impact evaluation is indeed possible of peacebuilding interventions in conflict-affected settings in a number of circumstances. This insight then brings us to a second type of concerns relating to the worth of impact evaluations. What sort of insights can these types of evaluations bring that other types cannot?

This paper addresses major concerns and questions about the feasibility, value added and ethics of impact evaluations in conflict-affected settings. It builds among others on the review of the two dozen screened studies from the Samii et al. paper, as well as insights from a survey we administered to the authors of the included studies. The paper explores (i) evaluation design issues in conflict-affected situations; (ii) evaluations as interventions, and the implications for the risks and reliability of results; (iii) the importance and value-added of impact evaluations; and (iv) ethical concerns about impact evaluations in conflict prevention and peacebuilding.

[2] http://www.campbellcollaboration.org/systematic_reviews/index.php

[3] Annex A lists the included studies.

4. Evaluation design issues in conflict-affected situations

4.1. Establishing the counterfactual

In designing an impact evaluation in fragile or unstable contexts, it is important to carefully consider how to establish a counterfactual, analyzing what is ethical and feasible in the particular context. While one might have expected to see mainly quasi-experimental designs used for impact evaluations of conflict prevention and peacebuilding interventions, given that these methodologies avoid many of the challenges of randomization, the majority of impact evaluations of these interventions still use experimental designs (Samii et al., 2012). This section provides a few illustrative examples of how different researchers have established a counterfactual using experimental and quasi-experimental designs.

The first example is of individual randomization. In Blattman and Annan's (2011) study of a reintegration program for ex-combatants in Liberia, demand exceeded supply of spaces in the program so registrants were admitted to the program by individual lottery. The program team publicized the intervention in their target communities to identified 'risky' populations and screened people interested in registering, identifying 1,330 eligible participants. The random assignment was stratified by gender, military rank, and location, using a computer program. From an ethical point of view, given that space in the program was limited, the equal chances of participating that the lottery awarded within each stratum was arguably the fairest and most transparent approach. An exception to the random assignment was made for those who previously held a rank of general in an armed group. Because they were considered high-risk by the program implementers, all who met this criterion were assigned to the program and were hence excluded from the study.

The second type of example is of group-based randomization. Many peacebuilding interventions are implemented in groups or communities, which requires group- instead of individual-based randomization. For example, for a community-driven program aiming to improve social cohesion, economic welfare and democratic governance in Liberia studied by Fearon, Humphreys and Weinstein (2008), the NGO randomly assigned 42 communities to receive the program of 83 eligible communities. The lottery was conducted in a public place, with chiefs representing each community in attendance. In a similar community-driven project in Sierra Leone, a pool of communities was selected from two districts that had regional, political and ethnic diversity, high levels of poverty and little NGO presence. From those districts, an eligible pool of communities of the appropriate size for the project were chosen

and then randomly assigned into treatment (118) and control (118) communities using a computerized random number generator (Casey, Glennerster & Miguel, 2011).

An additional example of group-based randomization is Paluck's evaluation of a reconciliation radio program in Rwanda, which used matched-pair randomization at the level of listening groups. Communities were first sampled to represent political, regional and ethnic breakdowns. Then communities were matched to the most similar community "according to a number of observable characteristics, such as gender ratio, quality of dwellings, and education level. Then, one community in each pair was randomly assigned to the reconciliation program and the other to the health program. This stratification of sites helped to balance and minimize observable differences between the communities ex ante" (Paluck, 2009a, pgs. 577-78).

The third type of examples is of quasi-experimental designs. Where randomization is not feasible or ethical, quasi-experimental designs may be used to create a suitable counter factual. For example, to examine the impact of a reintegration program on ex-combatants in Burundi, Gilligan, Mvukiyehe, and Samii (2010) used a disruption in the roll out of a program to construct a counterfactual. Three NGOs were given contracts in three different regions to provide benefits to ex-combatants. However, due to external factors, one of the NGOs delayed providing services for a year. Because the disruption was unrelated to choice of entry by participants or implementers, this comparison group theoretically avoids the traps of self-selection or targeting bias. However, the participants in the delayed area may be systematically different from individuals in the other two areas. To account for the potential imbalance on important covariates, the authors matched the 'treatment' and 'control' groups on individual (e.g., age, economic variables, and combatant variables) and community characteristics (e.g., war violence and population density) as well as propensity score[4].

To estimate the effects of the Demobilization, Disarmament, Rehabilitation and Reintegration (DDRR) program in Liberia on incomes and chances of employment, Levely (2012) used propensity-score matching based on age, gender, rank and county. As pointed out by the authors, propensity-score matching does not entirely solve the identification problem, as it does not account for potential self-selection on unobservable characteristics. Nevertheless, it does provide a more accurate estimate by accounting for observable variables.

[4] A propensity score is the probability of a unit (a person, a school, a community etc.) being assigned to receive an intervention (the 'treatment') given a set of observed covariates, were the treatment to be made available.

Sometimes the experimental conditions are determined by nature or by other factors beyond the control of the experimenters but imitate a randomized process to the extent that they are called natural experiments. In an evaluation of peace workshops for youth in Sri Lanka (Malhotra, 2005), those who came from the same schools as workshop participants and had been nominated to attend the workshops but had not been able to participate due to budget cuts that year were treated as a natural control group.

An underused quasi-experimental design is the Regression Discontinuity Design (RDD) which uses program eligibility criteria (e.g., an eligibility cutoff score such as a poverty-line) to establish the counterfactual. The Samii et al. search uncovered no existing RDD impact evaluations in the fields of conflict prevention and peacebuilding. One could however imagine a scenario where a program in this field had rated districts in a country by a fragility index, or by some index related to risk of (re)outbreak of violence. If the program decided that only districts with a score above a certain level qualified for their program, then the districts that were close to, but below, the cutoff point and hence did not take part in the program would be very similar to those who were just above it and hence received it, and could act as a control group. The advantage of this approach is that, as long as individuals cannot manipulate the assignment variable it is as good as an experiment, but only around the cutoff point. A challenge is to have a large enough sample of observations close to cutoff. It is also important to note that causal conclusions are limited to units close to the cutoff and that extrapolation beyond this point (whether to the rest of the sample or to a larger population) requires additional assumptions. (Lee and Lemieux, 2010).

We have described ways of establishing a counterfactual when eligible individuals are excluded from the treatment or the treatment is delayed or rationed. However, the peacebuilding interventions whose effectiveness we would like to measure are often of a nature that does not easily permit the identification of a control or comparison group because in theory they should be available to everyone at the same time. This could be the case, for example, when using the media to deliver peace messages, as above, or when a service such as social reintegration services for ex- combatants is in theory available to all. As long as the uptake of the service or intervention is less than 100 percent, there still exists the possibility to create a comparison group. This method is called an "encouragement design" because it requires that a randomly-selected group of beneficiaries receive additional encouragement, typically in the form of additional information or incentives, to take up the offered service (or use more of it). As long as information on relative uptake is available along with the measured outcomes, the encouragement design allows estimation of the effect of the intervention as well as the effect of the encouragement itself. The creation of listening groups has already been mentioned (Paluck,

2009a) as one type of encouragement. Other types could include an informational brochure about a service which is available or subsidizing the service-fee or sign-up costs of a service for a limited period.

4.2. Adaptation and flexibility throughout the evaluation process

The unpredictability of the situation in which many peace building and conflict prevention impact evaluations take place sometimes calls for flexibility in the design and implementation of the evaluation.

Despite serious challenges to data collection in conflict-affected environments, all but one of the impact evaluations summarized by Samii et al. (2012) involved collecting primary data. It appears that the dearth of useful administrative data in these settings leaves little option but to collect primary data. The Samii et al. review did report that relative to the comparison group of impact evaluations carried out in other sectors, the impact evaluations appeared to be based on smaller average sample sizes than the comparison group of impact evaluations (2.5 – 4 times smaller), which may limit the analysis, for example of differential effects on sub-groups.

We asked the researchers of the conflict prevention and peacebuilding studies whether and how the research teams adapted the data collection methods for the conflict-affected settings. Of the survey-responses that reported some adaptations, the types of modifications can be roughly divided into four categories: 1) adaptations to the sample; 2) timing; 3) question formulation and focus group composition; and 4) the enumerators' experience and training.

First, adaptation of sample size, either by design or due to unforeseen events, was a recurring response. In the evaluation of the Community Development Fund in Sudan, the researchers reported that they lost 60% of the sample communities due to the (re)outbreak of war (Paluck, 2009b), whereas in the impact evaluation of Afghanistan's National Solidarity Programme (NSP), the districts in which the security of the enumerators and participants was at risk were excluded from the intervention and evaluation (Beath et al., 2010). The fact that the research was not being done in the hostile Pashtun communities clearly affects and limits the generalizability of the findings, and so it is important to be careful about how one reads the evidence. As described above, in the evaluation of ex-combatant agriculture and psychological training program in Liberia, the team decided to exclude high ranking commanders from the evaluation in order to avoid potential conflict caused by randomizing them into both treatment and control. The program was concerned that commanders who were randomized into the control group may cause problems for the overall program. Therefore, all commanders were

provided access into the program and were excluded from the evaluation. The validity of the evaluation findings is therefore limited to the ex-combatants of lower ranks.

Second, the timing of surveys is one of the most commonly-cited adjustments made. The researchers involved in the evaluation of the peace education program for Israeli and Palestinian youth reported having had to adjust the timing of data collection due to the conflict (Biton and Solomon, 2006). In the evaluation of the Rwandan radio program, the team had planned follow-up interviews in prisons which were among the experimental sites. The timing of these had to be changed due to a sudden move to release prisoners (Paluck, 2009a). Similarly, in an ongoing evaluation of the community monitoring for better health program in Burkina Faso (World Bank, 2012), the research team had to halt data gathering in the Sahel because of problems with Tuaregs who were engaged in violent conflict in neighboring Mali. They were however able to gather the data at a later stage. It is worth reflecting on the fact that the measured size of effects is likely to change over time, and may take a non-linear shape, hence a great deal of caution is necessary when interpreting the findings for policy-making purposes. This will be particularly important in situations when the window of opportunity for data collection is limited.

Third, researchers described issues over what questions could be asked due to conflict-related sensitivities. In the evaluation of the Community Driven Development interventions in Sierra Leone, they explored whether they could ask about ongoing tensions, or directly about people's role in the conflict. The team spent time discussing with those working in the communities and piloting questions. They found little reluctance to talk about the conflict and found that it did not seem to raise tensions. However, they decided not to ask about some areas of current tensions, such as marital infidelity, as they were warned that this could spark tensions (Casey et al., 2011). In the civic education program in Southern Sudan, the focus groups were designed to prevent more conflict. Where the social divisions were based on sect, single-sect discussion groups were organized. Where conflict was based on the affiliation to ethnic/tribal groups, the groups included members of only one ethnic group (Paluck, 2009b).

Finally, researchers frequently mentioned the experience and background of the enumerators as a factor that had been taken into account when designing data-collection strategies. For both the studies of the Burundi ex-combatant reintegration program (Gilligan et al. 2010) and of the peacebuilding and democracy promotion efforts in Liberia (Mvukiyehe and Samii, 2010, 2011), the authors reported having recruited specially trained enumerators who had either done social work or human rights advocacy. It was deemed important that the research staff were sensitive to issues of trauma and trained to handle themselves in sensitive situations. For the evaluation

of an IRC Community Driven Reconstruction program in northern Liberia, the authors reported that the use of staff from a local organization, consisting almost entirely of ex-combatants, as enumerators had been helpful. In the case of the Afghanistan evaluation, female enumerators who were able to decide the most appropriate means of selecting participants carried out the focus groups and interviews among the female population. In the case of the evaluation of peacekeeping in Cote d'Ivoire (Mvukiyehe and Samii, 2009), the enumerators were intensively trained in human subjects and survey techniques for a week. For the evaluation of the reconciliation radio program in Rwanda (Paluck, 2009a), the research assistants represented both Hutu and Tutsi backgrounds, which in itself gave a message of tolerance and may have helped in downplaying ethnicity issues when approaching the communities.

4.3. Evaluations as interventions: Implications for reliability and risk

All evaluations in which primary data are being collected through human interaction could in themselves be seen or perceived as a type of intervention. This fact has potential implications both for the reliability of the evaluation results and for the safety of the evaluation personnel and those being evaluated. In addition, the perceived or real threat to safety is likely to be negatively correlated with the reliability of the results, as has been acknowledged in the new OECD DAC guidance for *Evaluating Peacebuilding Activities in Settings of Conflict and Fragility* (OECD/DAC, 2012). The guidance states that "evaluations of interventions in the field of conflict prevention and peacebuilding expose – in contrast to almost all forms of evaluation – both evaluators and evaluated to real risk". The guidance goes on to discuss the implications: "First, the threat of violence may constrain the evaluators' ability to raise issues, collect material and data, recruit and retain local staff, meet interlocutors, publish findings, and disclose sources. Defending the integrity of evaluation findings in highly politicized and even dangerous settings can pose problems for evaluation teams, particularly where evaluation findings may potentially be misused by different parties to a conflict or harm those involved. Second, the risk of harm may mean that the information obtained is biased, incomplete and/or (voluntarily or involuntarily) censored. Consequently, evaluations must address the operational and methodological consequences of the risk of violence. More specifically, in order to deal with this challenge, it is advisable that the evaluation itself include a conflict analysis in order to assess the intervention and to ensure that the evaluation process and product is conflict sensitive." (OECD/DAC, 2012, p. 28) Impact evaluations, to a greater extent than other evaluation methodologies, rely on the collection of primary data from a large number of units both in a treated and a comparison population. This means that evaluation teams may have increased exposure to the above-mentioned risks.

Carrying out evaluations in conflict affected settings can potentially cause harm to participants of the evaluation team (which under field visits may include implementation staff) and the local population interviewed. For example, in a community driven reconstruction evaluation in the DRC, "the harsh conditions produced great costs to enumerators with high incidence of sickness including malaria and cholera. Although safety regulations were in place in all areas, one of the teams was involved in a tragic accident in which a child died…. Despite the precautions undertaken we did encounter some security issues: 31 villages were not visited due to security risks; one team was ambushed and had to hand over their equipment; and one IRC staff member was abducted (and subsequently released unharmed)" (Humphreys et al., 2012, pgs. 34-35).

Carrying out evaluations in conflict affected settings could furthermore potentially adversely affect intergroup relations and the course of intergroup conflict. While examples were not found of this having happened, an ongoing study in Cote D'Ivoire (not included in the Samii et al. review) made evaluation design adjustments to minimize the risk of exacerbating existing conflict. The evaluation, which looks at the impact of couples discussion groups in addition to a savings intervention to combat intimate partner violence (Gupta and Annan, *ongoing*), *included* women who would not otherwise have been included in the sample. They were interviewing women in savings groups, but were only interested in women who had partners because the outcomes of interest were about partner relations and decision making. Given that the villages were in areas where there had been high ethnic tensions and conflict, the program team felt that if they separated the women and interviewed some and not others, there was the potential to create conflict and suspicion. They therefore decided also to administer a shorter survey to non-partnered women.

Broadly speaking, there are three reasons for evaluation teams to conduct conflict analyses: 1) to assess the relevance and impact of the program; 2) to assess the risks of negative effects of conflict on the evaluation design and process; and 3) to assess the risks of the evaluation exacerbating conflict (conflict sensitivity) (DFID, 2002). When reviewing the conflict assessments reported on in the summarized studies or commented upon in the survey responses, we found a varied approach. Some teams reported having relied on the assessment of the program implementing agency and partners in the country. In other cases, an assessment of risk to subjects was conducted as part of the institutional review board (IRB) approval process. A number of the studies derived insights from baseline studies that included questions about conflict experience, regular program reviews of the methods and measures with a 'do no harm' approach in mind, survey piloting and discussions with people working in similar communities, as well as behavioral monitoring. All of these approaches indicate an adapted use

of conflict assessment. A finding of some concern is that none of the studies or of the survey respondents indicated that they had given any thought to the potential of adverse effects from the publication and dissemination of results and findings.

Determining what is good enough in terms of conflict assessment is a difficult question and a balancing act between time, resources and a priori knowledge of risk. The instability of conflict-affected settings poses significant challenges to the rigor of evaluation design and the quality of data collected. Evaluations in these settings also introduce risks and potential harm to evaluators and the evaluated. Drawing from experiences of evaluators in these contexts, below are a set of questions about ethical and feasibility issues that research teams should consider:

(i) Does the evaluation factor in time for delays, which are more likely to occur in unstable conditions?

(ii) Does the sample size factor in the potential for higher attrition due to potential security, issues, migration or ethical concerns?

(iii) Have the potential ways that the evaluation may introduce risk and harm to participants, interviewers and implementing partners been adequately considered and have strategies been devised to mitigate these risks and harm?

(iv) Have interviewers been trained in ethical data collection and conflict sensitive approaches to study participants? Have the characteristics of the interview team been thought through in light of the conflict (i.e., ethnicity, age, gender, status).

(v) Is there a security protocol or guidelines for evaluation staff? Does evaluation staff fall under any organizational protection for security?

(vi) Who carries the legal responsibility for the risks taken? Have the researchers partnered with an organization able to bear the risks?

(vii) Have methods of monitoring the potential ethical and conflict-related issues throughout data collection process been considered and planned?

(viii) Does the evaluation team have strong key informants who can provide thoughtful analysis about the security situation and the research implications at the design phase and throughout the evaluation?

(ix) Is a flexible approach to the evaluation in place such that adjustments can be made throughout the process in light of potential harm, security or other programmatic issues?

(x) Is the responsibility of the dissemination and communication of the findings clarified, is there a communications plan in place and is it conflict sensitive?

We have seen that impact evaluations are feasible to carry out in diverse and challenging circumstances, but require precautionary measures, flexibility, and conflict sensitivity on the part of the evaluators. Having acknowledged the challenges and resources impact evaluations in

these contexts take, the next section attempts to address the core question of the value of impact evaluations.

5. Why are impact evaluations important?

5.1. Testing theories of change of conflict prevention and peacebuilding interventions

While large *n* impact evaluations of peacebuilding are possible, they are also difficult, so the question remains why do we need them? The answer is that only impact evaluations allow us to measure net impact and thus attribute the effects of the intervention. As a result, only impact evaluations allow us to test whether the intervention and its various inputs and outputs, lead to the hypothesized changes, outcomes and impacts in our theory of change (White, 2009). The simplest case for this claim is the before-after fallacy. Consider measuring an outcome both before the program and after the program. Typically if there is an improvement, the evaluator (and program manager) considers the intervention a success. But over the period of any program, many other factors come into play, not least of which, all the other programs that are being implemented in the same country. Without a valid counterfactual, there is no way of knowing whether the improvement can be attributed to the program's activities or may have happened in spite of these.

In conflict-affected settings, the before-after fallacy may be even more misleading, as the general situation may actually deteriorate over the period of the program. The before-after measurement would show the outcomes worsening, but a comparison to a counterfactual could very well reveal that the program prevented the outcomes from worsening to an even greater extent — a crucial result for a peacebuilding program. Similarly, a before-after measurement could show an improvement that is entirely due to other factors and may indeed mask unintended negative consequences of the program in question—again a crucial result for peacebuilding programs.

So, when returning to the question of the importance of impact evaluation, we suggest focusing on the two key tenets that tend to distinguish this type of evaluation from more traditional program evaluation, namely the need to account for other possible confounding factors and to focus on results rather than the intentions implicit in the process. While we stand to be corrected, our impression is that most disagreements and discussions about the importance of

impact evaluations, the way we have defined these in the paper, revolve around the need for a control or a comparison group to account for confounding factors. We will not deal with this larger debate here but refer to recent literature (Stern et al., 2012).

A limitation to quantitative impact evaluation often cited is the fact that large *n* impact evaluations can only be applied in large *n* situations, therefore significantly limiting the questions that can be addressed. While large *n* situations can be possible to implement even in what may seem like small *n* situations, such as when a nationwide policy is being implemented from which no one is or can be excluded (we have argued earlier in this paper that even in this case an encouragement design can help us give a dose-response perspective of the policy in question), quite often they are not. In these cases, rather than to move on and look for the next question that is evaluable by large *n* methods, we call for small *n* attribution analysis and will revert to these in the next sub-section. First, however, we present the type of learning and insights that can be gained by large *n* impact evaluations.

The results of several large *n* impact evaluations of peacebuilding interventions provide compelling evidence that many key assumptions and theories of change about conflict prevention and peacebuilding need to be tested. This section presents examples of impact evaluations whose findings challenge the theories that personal beliefs and prejudices need change in order to change behavior; that discussion and debate necessarily leads to improved tolerance; and that Community-Driven Development (CDD) or Community Driven Reconstruction (CDR)projects, at least in the way these have tended to be implemented, improve social cohesion.

Two studies by Elizabeth Levy Paluck (2012) test psychological theories of attitude and behavior change from media interventions designed to help rebuild communities following conflict. In Rwanda, she evaluated a reconciliation-themed radio soap opera (Paluck, 2009a) and in eastern Democratic Republic of Congo (DRC), she evaluated a radio talk show that was aired in conjunction with a talk show (Paluck, 2010).

The first evaluation tested conflicting psychological theories about the relationships between personal beliefs, societal norms, and behaviors, and how those can be influenced by media. In Rwanda, the NGO La Benevolencija produced a radio soap opera called *Musekewaya* ("New Dawn") that was designed to promote reconciliation by playing out a story that includes similar sources of tensions and violent outcomes as the 1994 Rwandan genocide[5], but that speaks out

[5] The Rwandan Genocide was the 1994 mass murder of an estimated 800,000 people in the East African state of Rwanda. It was the culmination of longstanding ethnic competition and tensions between the minority Tutsi, who

against violence and includes characters banding together across ethnic groups (which were proxied by "communities" as the government forbade the use of the word "ethnic"). Although the radio program was aired nationwide, Paluck created a pair-wise matched cluster randomized controlled trial using an "encouragement" design. She established listening groups to encourage the beneficiary, or treatment, group to listen to the "New Dawn" program and to concurrently encourage the control group to listen to an alternate radio program on health.

Since the ultimate goal of the program was to reduce intergroup conflict, the questions the experiment tried to answer were first, can such a radio program influence both personal beliefs and prejudices as well as perceived societal norm, and second, is a change in personal beliefs a necessary precondition to influence behavior. While psychological theories conflict, "theories of media persuasion claim that beliefs are influenced by media cultures and programs" (Paluck 2009a, p. 575). The findings were startling; the perceptions of social norms as well as behaviors changed significantly in the treatment group with respect to intermarriage, open dissent, trust, empathy, cooperation, and trauma healing, while the program did not significantly change listeners' personal beliefs.

The second evaluation tested the effectiveness of discussion to reduce conflict. In the DRC, a radio soap opera *Kumbuka Kesho* ('Think of tomorrow') emphasized conflict reduction through community cooperation. While the radio program was aired in all the experiment's regions, Paluck again used an encouragement design, this time by pair-wise matching regions and randomly choosing one broadcast region in each pair to air a talk-show directly following the soap opera, and the other the soap opera only. The talk show was designed to encourage listeners' reactions and discussions. While there is a resurgence in the use of discussion as a policy tool to reduce conflict (evidenced by the proliferation of terms such as "deliberation", "dialogue", "participatory" and "community driven" in the literature on interventions designed to promote peace) psychological research has also flagged potential hazards of discussions including opinion polarization, social pressure and cognitive errors (Paluck; 2010). Paluck carried out this research to learn more about the success of discussion-based conflict-reduction programs. The findings were sobering: those listeners who were encouraged to discuss through the additional talk show did indeed discuss more, but were also found to become more intolerant and less likely to aid disliked community members.

---

had controlled power for centuries, and the majority Hutu peoples, who had come to power in the rebellion of 1959–62 (Wikipedia; accessed 31/10/2012).

A third group of evaluations examined the effectiveness of Community-Driven Development projects to strengthen social cohesion. A commonly proposed theory is that of the importance of social cohesion, or the (re)building of interpersonal or intergroup networks, trust, and reciprocity, as a crucial factor for peacebuilding and conflict prevention. In a recent talk at the launching conference for the High Commissioner on National Minorities (HCNM) Guidelines on Integration in Diverse Societies, Stefan Wolff answered his rhetorical question about what it is about social cohesion that is so important for successful conflict prevention in the following way: "One of the fundamental ideas underlying the notion of conflict prevention in diverse societies is that different population segments can resolve any differences by recourse to institutional processes rather than violence. For such institutional processes to be effective, a viable and resilient state is required whose fundamental constitutional principles are broadly accepted and respected across all segments of society. If this is the case, societies may well be diverse across any number of indicators, including, ethnicity, language, and religion, but they will also be characterized by a sufficient level of social cohesion." (Wolff, 2012). Efforts to strengthen social cohesion have increased among development organizations, most often operationalized through Community-Driven Development (CDD) or Community Driven Reconstruction (CDR) projects. In a systematic review of interventions to promote social cohesion in Sub-Saharan Africa, including in several conflict-affected countries (King et al. 2010), the authors outlined the theory underlying CDD interventions: "projects promote social cohesion by supporting and building community capacity for decision-making and collective action through a process of participation. The hypothesis is that, by handing over control of decisions and resources to the community, the sub-projects will better meet communities' needs and enhance ownership, and that the experience of being involved in this participatory process will empower communities, improve capacity for local development and improve social cohesion." (King et al. 2010, p. 347) Drawing upon the available evidence from impact evaluations that fulfilled a set of quality criteria, the review finds that the evidence of pro-social effects from Community-Driven Development (CDD) type interventions is weak. More surprisingly, a negative effect on individuals' perceptions of inter-group relations is found across the three studies that measured this factor.[6]

The preceding examples of how impact evaluations have been used as tools to test and critically examine commonly-held assumptions about how development interventions affect change were all based on large *n* impact evaluations. But what happens when we have a question about

[6] The review indicates that this finding may be partly explained by the fact that broad and substantive participation, including in actual decision-making, was often lacking and suggests that the implementation of the CDD interventions may have been flawed.

results and impact of an intervention, be it a policy reform or a service delivered, on the ground (on the so-called 'beneficiaries') and do not have a large number of units of assignment? The next section discusses the use and commonalities of small *n* impact evaluations.

5.2. Small n impact evaluations

What distinguishes impact evaluation from other types of evaluation is that it relies on a counterfactual analysis to attribute an effect to a particular intervention or set of interventions, or said differently; to make causal inferences. We further distinguish between large *n* impact evaluations which involve tests of statistical significance between outcomes for treatment and comparison groups, with *n* referring to the unit of assignment, and small *n* impact evaluations carried out when a treatment and comparison group of sufficient size cannot be identified, be it individuals, communities or countries, and thus where tests of statistical significance are not possible.

While there exists considerable consensus among impact evaluators conducting large *n* impact evaluations as to what constitutes a high quality impact evaluation, no such consensus exists for small *n* impact evaluations. In a recent paper by White and Phillips (2012), they examine various small *n* evaluation approaches that have been used and find that a methodological core which could provide a basis for consensus exists: 'This common core involves the specification of a theory of change together with a number of further alternative causal hypotheses. Causation is established beyond reasonable doubt by collecting evidence to validate, invalidate, or revise the hypothesized explanations, with the goal of rigorously evidencing the links in the actual causal chain'[7]. This type of approaches they refer to as process- or mechanism-based approaches. They go on to summarizing the main difference between large and small *n* evaluations in the following manner: 'Whereas experimental approaches infer causality by identifying the outcomes resulting from manipulated causes, a mechanism-based approach searches for the causes of observed outcomes'[8]. The small *n* evaluations will typically gather information on both the 'what' and the 'why', but are at risk of suffering from substantial biases likely to arise from the collection, analysis and reporting of qualitative data.

Quite often however, when large *n* impact evaluation is not possible, evaluators revert to process evaluations[9] or impact assessments based on association[10] rather than to small *n*

[7] White and Phillips, 2012, p.3.

[8] White and Phillips, 2012, p.18.

[9] What distinguishes impact evaluations from process evaluations - evaluations of how the implementation was carried out – is that the benchmark against which we compare in process evaluations is not a counterfactual

attribution analysis, not out of methodological disagreement but rather due to a whole range of supply and demand limitations (related to time and resources, evaluation skills etc.) (see Grävingholt and Leininger, forthcoming 2013).

An illustration of an evaluation that used elements of the methodologies referred to as small *n* attribution analysis to critically assess important theories of change is the evaluation of Norwegian peace efforts in Sri Lanka (Goodhand et al., forthcoming 2013). Among the main objectives of the Sri Lanka evaluation was to assess results achieved through the Norwegian facilitation of the peace process. This is a case where the total population (N) is 1 (and small *n* can obviously not be larger than large N). In other words, there was only one peace negotiation process going on with Norwegian involvement in Sri Lanka, and that was what the researchers set out to evaluate. Clearly, no large *n* impact evaluation was feasible. What about small *n* attribution analysis? One of the main challenges to attributing results to the Norwegian facilitation efforts is that the 'treatment', Norwegian facilitation, cannot be assumed to be an independent variable – rather the Sri Lankan and international actors chose to contact Norway, or did not object to this, requesting it to play the role as facilitator (and Norway chose to accept). It is likely that assumptions about what role Norway could or would play will have influenced the decision of approaching Norwegian policymakers. Indeed, according to the report "Norway was chosen as a facilitator, not only for its expertise, but also because it was a small power without geo-strategic interests and colonial baggage. Being a less powerful player, Norway felt it had to consult the US and India, the former as the world's superpower and the latter as the regional hegemon" (Goodhand et al., 2011, p.73). Assuming the Norwegian treatment as exogenous would have led to overplaying the role of agent, as opposed to context and path-dependence being crucial factors. Indeed, the provocative title of the evaluation report, "pawns of peace", alludes directly to the endogeneity issue.

The methods chosen by the team include features designed to explicitly assess the plausibility of causal claims; the common feature of small *n* attribution analysis. In particular, the 'inside out' and 'outside in' approaches that they seek to combine allows them to critically assess whether it is realistic to believe that if Norway had acted differently different outcomes would have ensued

---

scenario but rather non-tested (or in the best-case scenario previously tested) assumptions of what underlies a 'good process'.

[10] Association claims are very widespread in the small *n* evaluation world, as is raised elsewhere in this book (Grävingholt and Leininger, forthcoming). These are claims of having contributed to an outcome (or sometimes even claiming attribution) by having contributed an input or claiming to have done so (e.g. by having been present at the same time). This approach does not explore alternative causal hypothesis, the minimum criteria for small *n* attribution analysis, and is clearly not good enough.

at various points in time, given the structural constraints in which key actors operated. The study is also very explicit about the many data collection constraints and biases they faced, including missing key informants, secrecy and safety issues, conflicting and unreliable accounts, and not being able to interview a number of key informants in person due to visa problems. The main strategy used to deal with these challenges was that of triangulation.

6. Ethical concerns about impact evaluations in conflict prevention and peacebuilding

The descriptors of large *n* impact evaluations in conflict-affected settings raise several ethical concerns. There is the concern that impact evaluation designs require that only some individuals[11] receive the intervention. This is considered an ethical problem and some claim that for certain peacebuilding interventions, it simply is not feasible to involve some individuals and not others. These objections are not unique to evaluations in conflict-affected settings although the risks in these settings may be heightened. We will see that just as possibilities for ethical evaluations abound in other types of development interventions, they also exist in conflict-affected settings. Randomization or quasi-experimental designs do not necessarily drive the fact that only some individuals receive the intervention; they are particularly well-suited when for financial or logistical reasons the implementation and roll-out is slow or staggered, or when comparable groups are left out for other reasons. This is the reality of most development interventions, as well as those in fragile settings.

Part of what underlies the ethical concern about impact evaluations is the premise that assignment to a comparison or control group implies 'not receiving a benefit'. This is not necessarily the case for two reasons. First, the comparison group can be receiving a treatment with which another competing intervention is being compared. For example, in the case of the impact evaluation of the agricultural training program for ex-combatants in Liberia (Blattman and Annan, 2011) questions about whether to invest in capital or skills in agricultural programming arose as some of the results suggested that the private returns to capital could be higher than those for skills. In a future impact evaluation one could compare the impact of providing capital versus skills without the necessity of a control group that does not receive any program interventions.

[11] For the purpose of discussion, we use the term "individuals" for the unit of analysis, although households or communities or other entities may also be the unit of analysis.

Second, it is important to examine the assumption that receiving a development intervention, or more of one, is always a benefit. The reality is that the effectiveness and impact of a large number of development interventions have yet to be proven (CGD, 2006). When a genuine state of uncertainty exists about the benefits of an intervention, so that in theory it could be harmful or ineffective, there is an urgent need for it to be critically examined. This state of uncertainty, known as *equipoise* in the medical literature, is considered a necessary ethical condition for the use of a control group which is reflected (with a couple of important caveats) in the Declaration of Helsinki on the use of placebo controls (World Medical Association, 2001; Lau et al. 2003). When there is no known effective medical treatment, a new drug might produce better, worse, or the same results as no treatment, and so there is no ethical conflict in trials where this equipoise is present.[12] The evaluation discussed in section 5.1 provides a case in point. Discussion groups as a tool to reduce ethnic conflict was tested in the DRC context, and found to increase rather than decrease intolerance (Paluck, 2010).

Another concern raised is about randomizing a program's activities across possible beneficiaries instead of selecting according to other criteria (e.g., those who first apply or those easiest to access). In a conflict-affected setting, prioritizing certain beneficiaries could be important for defusing volatile situations or prioritizing quick wins. On the other hand, in cases where a program cannot be implemented across all individuals immediately, randomization of eligible individuals can in fact be more ethical and politically feasible than determining who benefits first and who later, especially in a sensitive situation where particular choices can be construed as being politically motivated.

While the ethical concerns may sometimes be misplaced or exaggerated for the reasons just described, it is nevertheless critically important to always carefully consider the potential ethical issues that may arise when designing and conducting impact evaluations. Guidelines exist to help determine when not to do an impact evaluation for ethical reasons, and there exist a number of strategies to alleviate ethical concerns. Many agencies and universities have formal ethical clearance procedures, and the standards typically include (i) ensuring informed consent, (ii) guaranteeing the confidentiality of participant data; (iii) limiting the burden associated with study participation; and (iv) making sure that no one is denied essential services for the purpose of the evaluation (Friedman, 2011; USDA, 2005).

[12] It is worth noting that development interventions may differ slightly from the medical "equipoise" case of a zero probability event, in the sense that the development practitioners tend to hold priors of an intervention being beneficial but with some remaining doubts.

7. Conclusion: High risk, high return?

Carrying out impact evaluations in conflict-affected settings can be risky and methodologically challenging, though we have discussed ways in which the evaluation designs and data collection practices can be adapted and risks reduced to make their implementation feasible. Impact evaluations are also costly, due to the reliance on data from large samples to achieve statistical power, ranging from as little as US$50,000 for quasi-experimental impact evaluations with preexisting survey data to over US$1 million for large multi-year RCTs with several rounds of survey-data gathered. For both of these reasons, the returns to the studies in terms of learning and programmatic improvements should also be high for the effort to be worth it.

We have argued that if we are interested in the actual development effects of interventions and programs on the people they are supposed to benefit, rather than whether the program was implemented as planned, and if we want to know whether this effect was due to, or despite of, the intervention in question, then a well-designed and executed impact evaluation is the most reliable approach. The potential usefulness and importance of impact evaluation is well exemplified by the way impact evaluations have tested and challenged many of the key assumptions and theories of change that underpin conflict prevention and peacebuilding activities.

Important insights have therefore been gained and it is important that this knowledge feeds back into the way we design and implement conflict prevention and peacebuilding programs, as well as the way we carry out program-theory evaluations. To date, evidence of putting learning into practice from impact evaluations is limited. Of the 13 programs in which survey respondents had been involved, two were rated as not having led to any learning, three as having contributed to program improvements or general learning around a program type, and in all of eight cases the respondents said any learning impact was unclear or 'too early to tell'. It may be that learning happens without the knowledge of the researchers, and clearly learning takes time. Especially when trying to draw lessons that have validity beyond a single program, country, and point in time, it is necessary to build up a body of evidence and systematically review it. Nevertheless, the survey responses are a good reminder that dissemination and learning from evaluation work, the *raison d'être* of these risky and challenging endeavors, cannot be taken for granted. Whether the high risk leads to high returns remains an open question. The returns will to a large extent depend on the international development community's capacity to more strategically incorporate evidence-based learning into interventions operating in contexts of conflict and fragility.

References

Annan, Jeannie and Christopher Blattman, 2011, "Reintegrating and Employing High Risk Youth in Liberia: Lessons from a randomized evaluation of a Landmine Action agricultural training program for ex-combatants." Evidence from Randomized Evaluations of Peacebuilding in Liberia: Policy Report 2011. 1, Yale and IPA.

Baker, Judy. L, 2000, Evaluating the Impact of Development Projects on Poverty, A Handbook for Practitioners, The International Bank for Reconstruction and Development/THE WORLD BANK, 1818 H Street, N.W., Washington, D.C. 20433

Barron, Patrick, Macartan Humphreys, Laura Paler, and Jeremy Weinstein, 2009, Community-Based Reintegration in Aceh: Assessing the Impacts of BRA-KDP, *World Bank,* www.columbia.edu/~lbp2106/docs/arls/FINAL_BRA-KDP_WB.pdf.

Beath, Andrew, Fotini Christia, Ruben Enikolopov, and Shahim Ahmad Kabuli, 2010, Randomized Impact Evaluation of Phase II of Afghanistan's National Solidarity Programme (NSP): Estimates of Interim Program Impact from First Follow-up Survey, http://www.nsp-ie.org/reports/BCEK-Interim_Estimates_of_Program_Impact_2010_07_13.pdf.

Biton, Yifat and Gavriel Solomon, 2006, Peace in the Eyes of Israeli and Palestinian Youths: Effects of Collective Narratives and Peace Education Program, *Journal of Peace Research* 43, no. 2: 167-180, http://jpr.sagepub.com/cgi/doi/10.1177/0022343306061888.

Blattman, Christopher, 2011, Uganda: Enterprises for Ultra-poor Women after War, (*in progress*) http://chrisblattman.com/projects/wings.

---, 2011, Uganda: Post-war Youth Vocational Training.‖(*in progress*) www.chrisblattman.com/projects/nusaf_yo/.

---, 2011, Peace Education in Rural Liberia.‖ *Innovations for Poverty Action*. (*in progress*), www.poverty-action.org/project/0139.

Blattman, Christopher and Jeannie Annan, 2011, Reintegrating and Employing High Risk Youth in Liberia: Lessons from a randomized evaluation of a Landmine Action agricultural training program for ex-combatants, Evidence from Randomized Evaluations of Peacebuilding in Liberia: Policy Report 2011.1, Interventions for Policy Action, IPA, Yale University.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, 2011, Reshaping Institutions: Evidence on External Aid and Local Collective Action*, National Bureau of Economic Research Working Paper* no. 17012, http://www.nber.org/papers/w17012

Center for Global Development, 2006, When Will We Ever Learn? Improving Lives through Impact Evaluation, Report of the Evaluation Gap Working Group, May 2006.

Diamond, A. and J. Hainmueller, 2007, The Encouragement Design for Program Evaluation, IFC; http://www.ifc.org/ifcext/rmas.nsf/AttachmentsByTitle/Encouragement/$FILE/The+Encouragement+Design+for+Program+Evaluation.pdf

DiNardo, J. (2008). "Natural experiments and quasi-natural experiments". In Durlauf, Steven N.; Blume, Lawrence E. *The New Palgrave Dictionary of Economics* (Second ed.). Palgrave Macmillan.

Department For International Development (DFID), 2002, Conducting conflict assessments: guidance notes, issues, http://www.conflictsensitivity.org/sites/default/files/Conducting_Conflict_Assessment_Guidance.pdf (accessed 06/01/2013).

Fearon, James, Macartan Humphreys, and Jeremy M. Weinstein, 2008, *Communitty-Driven Reconstruction in Lofa County*.www.columbia.edu/~mh2245/FHW/FHW_final.pdf.

Friedman, Jed, 2011, Development Impact blog: The ethics of a control group in randomized impact evaluations – the start of an ongoing discussion, accessed 31/03/2013; http://blogs.worldbank.org/impactevaluations/node/598

Gilligan, Michael, Eric Mvukiyehe, and Cyrus Samii, 2010, Reintegrating Rebels Into Civilian Life: Quasi-experimental Evidence From Burundi, *United States Institute of Peace*, http://www.columbia.edu/~cds81/docs/bdi09_reintegration100701.pdf.

Glennerster, Rachel, and Edward Miguel, 2010, The role of information and radios on political knowledge and participation in Sierra Leone, *Poverty Action Lab* (*in progress*), http://www.povertyactionlab.org/evaluation/role-information-and-radios-political-knowledge-and-participation-sierra-leone.

Goodhand, Jonathan, Bart Klem and Gunnar Sørbø, 2013, Evaluating Norwegian peace efforts in Sri Lanka, in O. Winckler Andersen, B. Bull and M. Kennedy-Chouane, *Evaluation Methodologies for Aid in Conflict*, London: Routledge, Taylor and Francis Group.

Grävingholt, Jörn and Julia Leininger, (forthcoming, 2013), Evaluating Statebuilding Support in Fragile States: Learning from Experience or Judging from Assumptions?, in O. Winckler Andersen, B. Bull and M. Kennedy-Chouane, *Evaluation Methodologies for Aid in Conflict*, London: Routledge, Taylor and Francis Group.

Gupta, Jhumka and Jeannie Annan, ongoing IRC project, Evaluating an economic and empowerment intervention on the prevention of partner violence.

Hossain M, Zimmerman C, Kiss L, Watts C. (2010), *Violence against women and men in Côte d'Ivoire: A cluster randomized controlled trial to assess the impact of the 'Men & Women in Partnership' intervention on the reduction of violence against women and girls in rural Côte d'Ivoire - Results from a community survey,* London: London School of Hygiene & Tropical Medicine.

Humphreys, Macartan, 2008, Community-Driven Resonstruction in the Democratic Republic of Congo, Baseline Report, Columbia University and the International Rescue Committee.

Humphreys, Macartan, and Jeremy M. Weinstein, 2007, Demobilization and Reintegration, *Journal of Conflict Resolution* 51, no. 4 (August): 531-567, http://jcr.sagepub.com/cgi/doi/10.1177/0022002707302790.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt, 2012, *Social and Economic Impacts of Tuungane. Final Report on the Effects of a Community Driven Reconstruction Program in Eastern Democratic Republic of Congo*, Columbia University, June 2012, pgs 34-35, http://www.oecd.org/countries/democraticrepublicofthecongo/drc.pdf

King, Elisabeth, Cyrus Samii and Birte Snilstveit, 2010, Interventions to promote social cohesion in sub-Saharan Africa. Journal of Development Effectiveness, 2(3), pp. 336–370.

Kondylis, Florence., 2007, Agricultural Outputs and Conflict Displacement: Evidence from a Policy Intervention in Rwanda, *Households in Conflict Network Working Paper* 28, http://www.csae.ox.ac.uk/conferences/2007-edia-lawbidc/papers/046-kondylis.pdf.

Lau, J.T.F., J. Mao, and J. Woo, 2003, "Ethical Issues Related to the Use of Placebo in Clinical Trials." *Hong Kong Medical Journal* 9.3 (2003): 192-98.

Lee, David and Thomas Lemieux (2010), Regression Discontinuity Designs in Economics, *Journal of Economic Literature 48 (June 2010): 281–355*

Levely, Ian, 2012, Measuring Intermediate Outcomes of Liberia's DDRR Program, Institute of Economic Studies, Faculty of Social Sciences Charles University in Prague, IES Working Paper 2/2012.

Malhotra, D., 2005, Long-Term Effects of Peace Workshops in Protracted Conflicts, *Journal of Conflict Resolution* 49, no. 6 (December): 908-924, http://jcr.sagepub.com/cgi/doi/10.1177/0022002705281153.

Mvukiyehe, Eric and Cyrus Samii, 2011, Peace from the Bottom Up: A Randomized Trial with UN Peacekeepers, Paper presented at the FBA Peacekeeping Working Group, Stockholm, February 11-12, 2011.

Mvukiyehe, Eric and Cyrus Samii, 2010, Quantitative Impact Evaluation of the United Nations Mission in Liberia: Final Report, Typescript, Columbia University, www.columbia.edu/~cds81/docs/lib/unmil_final100209.pdf.

Mvukiyehe, Eric and Cyrus Samii, 2009, Laying a Foundation for Peace? Micro-Effects of Peacekeeping in Cote d'Ivoire, Paper prepared for the 2009 American Political Science Association Conference, Toronto, http://www.columbia.edu/~cds81/docs/unoci/mvukiyehe_samii_unoci090801.pdf.

OECD (2012), *Evaluating Peacebuilding Activities in Settings of Conflict and Fragility: Improving Learning for Results*, DAC Guidelines and Reference Series, OECD Publishing. doi: 10.1787/9789264106802-en

Paluck, Elizabeth Levy, 2009a, Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda, *Journal of Personality and Social Psychology* 96, no. 3 (March): 574-587, http://www.ncbi.nlm.nih.gov/pubmed/19254104.

Paluck, Elizabeth Levy, 2009b, "Entertainment, Information, and Discussion: Experimenting with media techniques for civic education and engagement in Southern Sudan." Memo presented at the Experiments on Government and Politics (EGAP) Conference at the Institution for Social and Policy Studies, Yale, April 24-25, 2009. http://isps.research.yale.edu/conferences/EGAP/egap/download/Paluck_4.25.09_MEMO.pdf.

Paluck, Elizabeth Levy, 2010, Is It Better Not to Talk? Group Polarization, Extended Contact, and Perspectives Taking in Eastern Republic of Congo, Personality and Social Psychology Bulletin 36 no. 9: 1170-1185.

Paluck, Elizabeth Levy, and Donald P. Green, 2009, Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda, *American Political Science Review* 103, no. 04 (October): 622, http://www.journals.cambridge.org/abstract_S0003055409990128.

Pugel, James, 2007, What the Fighters Say: A Survey of Ex-combatants in Liberia, *United Nations Development Programme – Liberia*, www.lr.undp.org/UNDPwhatFightersSayLiberia-2006.pdf.Samii, Cyrus, Annette N. Brown, and Monika Kulma, 2012, Evaluating Stabilization Interventions, *Working draft 2.0*, August 16, 2012

Samii, Cyrus; Annette Brown and Monika Kulma (2012), Evaluating Stabilization Interventions, International Initative for Impact Evaluation (3ie) White Paper, https://files.nyu.edu/cds2083/public/docs/evaluating_stabilization_interventions_120816shortenedb.pdf (accessed 06/20/2013).

Stern, Elliot, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, and Barbara Befani, 2012, Broadening the range of designs and methods for impact evaluations, *report of a study commissioned by the Department for International Development,* Working Paper 38.

USDA, Food and Nutrition Service, 2005, Nutrition Education: Principles of Sound Impact Evaluation, http://www.fns.usda.gov/Ora/menu/Published/NutritionEducation/Files/EvaluationPrinciples.pdf, accessed 31/03/2013.

White, Howard and Daniel Phillips, 2012, Addressing attribution of cause and effect in small n impact evaluations: towards and integrated framework, International Initiative for Impact Evaluation, 3ie; Working Paper 15.

White, Howard, 2009, Theory-Based Impact Evaluation: Principles and Practice, 3ie Working paper 3.

White, Howard, 2011, An introduction to the use of randomized control trails to evaluate development interventions, the International Initiative for Impact Evaluation, Working Paper 9.

Winckler Andersen, Ole, Bull, Beate and Megan Kennedy-Chouane (forthcoming, 2013), Evaluation Methodologies for Aid in Conflict, London: Routledge, Taylor and Francis Group.

Wolff, Stefan, 2012, Integration and Conflict Prevention in Diverse Societies, The Ljubljana Recommendations of the OSCE High Commissioner on National Minorities in the Post-Soviet Context, Launching Conference HCNM Guidelines on Integration in Diverse Societies, November 7, 2012, http://www.stefanwolff.com/talks/integration-and-conflict-prevention-in-diverse-societies, accessed 19/11/12.

World Bank, 2012, Impact Evaluation of the Burkina Faso Community Monitoring for Better Health and Education Service Delivery Project, ongoing evaluation presented at a World Bank seminar July 10 2012; http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDEVIMPEVAINI/0,,contentMDK:23238146~menuPK:7637304~pagePK:64168445~piPK:64168309~theSitePK:3998212,00.html (accessed 06/01/2013).

World Bank, 2013, Evaluation Designs (webpage), http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20188242~menuPK:415130~pagePK:148956~piPK:216618~theSitePK:384329,00.html, accessed 30/03/2013.

World Development Report, 2011: Conflict, Security and Development, the World Bank Group.

World Medical Association, 2000, Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA* 284:3043–45.

World Medical Association, 2001, Note of clarification on paragraph 29 of the WMA Declaration of Helsinki. Geneva: World Medical Association. Available from: http://www.wma.net/e/home.html.

Annex A: Studies reviewed in the Samii, Brown, and Kulma (2012) paper

	Article	Category	Country	Status	IE Type	Counterfactual
1	Annan, J. and C. Blattman (2011)	Ex-Combatant Reintegration	Liberia	Ongoing	RCT	Randomized Control Group
2	Beath, A. et al. (2010)	Peace Dividends	Afghanistan	Completed	RCT	Randomized Control Group
3	Blattman, C. (2011a)	Peace Structures	Liberia	Ongoing	RCT	Randomized Control Group
4	Blattman, C. (2011b)	Victims of War	Uganda	Ongoing	RCT	Delayed Treatment Control Group
5	Blattman, C. (2011c)	Victims of War	Uganda	Ongoing	RCT	Randomized Control Group
6	Casey, K. (2011)	Peace Dividends	Sierra Leone	Completed	RCT	Randomized Control Group
7	Fearon, J. et al. (2009)	Peace Dividends	Liberia	Completed	RCT	Randomized group assignment of villages
8	Fearon, J. et al. (2008)	Peace Dividends	Liberia	Completed	RCT	Randomized Control Group
9	Glennerster, R. and E. Miguel (2010)	Peace Messaging	Sierra Leone	Ongoing	RCT	Randomized Control Group
10	Paluck, E. and D. Green (2009)	Peace Messaging	Rwanda	Completed	RCT	Clustered random assignment
11	Paluck, E. (2009a)	Peace Messaging	Sudan	Ongoing	RCT	Clustered random assignment with factorial model
12	Paluck, E. (2009b)	Peace Messaging	Rwanda	Completed	RCT	Randomized assignment of clusters with matching
13	Pugel, J. (2007)	Ex-Combatant Reintegration	Liberia	Completed	RCT	Randomized selection of 20 person clusters
14	Paluck, E. (2010)	Peace Messaging	DRC	Completed	RCT	Randomized assignment of clusters with matching
15	Barron, P. et al.(2009)	Peace Dividends	Indonesia	Completed	Quasi Experimental	Matched Control Group
16	Biton, Y. and G. Solomon (2006)	Consensus & Dialogue	Israel	Completed	Quasi Experimental	Matched-pair randomization of classes in selected schools/ natural
17	Gilligan, M. et al. (2010)	Ex-Combatant Reintegration	Burundi	Completed	Quasi Experimental	Natural control group with matching
18	Humphreys, M. and J. Weinstein (2007)	Ex-Combatant Reintegration	Sierra Leone	Completed	Quasi Experimental	Matched control group
19	Kondylis, F. (2007)	Victims of War	Rwanda	Preliminary	Quasi Experimental	Natural control group
20	Levely, I. (2010)	Ex-Combatant Reintegration	Liberia	Completed	Quasi Experimental	Matched Control Group
21	Malhotra, D. and S. Liyanage (2005)	Consensus & Dialogue	Sri Lanka	Completed	Quasi Experimental	Natural control group

22	Mvukiyehe, E. and C. Samii (2009)	Peace Dividends	Cote d'Ivoire	Preliminary	Quasi Experimental	Natural control group
23	Mvukiyehe, E. and C. Samii (2011)	Community Security Initiatives	Liberia	Preliminary	Quasi Experimental	Matched Clusters (communities)
24	Mvukiyehe, E. and C. Samii (2010)	Ex-Combatant Reintegration, Peace Dividends	Liberia	Completed	Quasi Experimental	Cluster matched sampling