Evaluating Stabilization Interventions

Cyrus Samii, Annette N. Brown, and Monika Kulma[1]

Working draft 2.0

August 16, 2012

1. Introduction

In the worlds of public health, education, and poverty relief, a disposition toward "evidence-based" policy-making has redefined the way that many policymakers assess policy options.[2] The evidence-based approach emphasizes the use of rigorous impact evaluations of pilot programs and previous applications of a given intervention. The goal is to determine whether a particular policy is worth pursuing, and to do so above and beyond theoretical arguments for why the policy is worthy. The question arises whether this approach is relevant to other policy areas—in particular, to areas of foreign assistance and conflict management that have traditionally been viewed as too sensitive or fluid for such demanding methods of evaluation.

We argue that the scope for applying evidence-based approaches to policy-making in the stabilization sector is much greater than current practice, especially US government practice, reflects. We focus here on impact evaluation for stabilization interventions, by which we mean discrete projects that have a clear beginning and end, take place in post-conflict or other "fragile state" contexts, and that aim to directly or indirectly strengthen the durability of peace and prevent individual level "human security" from being compromised by violence or intimidation by state or non-state actors.[3] By impact evaluations, we mean studies that measure impacts attributed to an intervention using experimental or quasi-experimental methods to either compare "treated" and "control" units or compare different varieties of an intervention. To make our case, we present examples of state-of-the-art impact evaluations of interventions undertaken in the politically and logistically challenging settings that characterize stabilization environments. We focus on the lessons they offer for credibly establishing causal attribution. We also review initiatives recently undertaken

[1] Cyrus Samii is Assistant Professor, Wilf Family Department of Politics, New York University, Email: cds2083@nyu.edu. Annette Brown is Chief of Advancement and Impact Evaluation Services as well as Chief Evaluation Officer, International Initiative for Impact Evaluation (3ie), Email: abrown@3ieimpact.org. Monika Kulma is researcher for 3ie (2011) and a recent Master of Public Policy graduate at Georgetown University. This research has received support from 3ie. We thank participants at the 2011 U.S. State Department Conference on Program Evaluation, 2011 3ie Impact Evaluation Conference, and 2011 3ie Delhi Seminar Series for helpful feedback. This document does not reflect the positions of any of the above named organizations, and any errors or omissions are the sole responsibility of the authors.

[2] See "Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide," U.S. Department of Education; Banerjee and Duflo (2011); and Karlan and Appel (2011) for some examples and discussion.

[3] Our definition rests on the notion of human security as developed in King and Murray (2001) and Sen (2000).

to provide guidance on measuring stabilization outcomes. Our review of state-of-the-art evaluations and measurement initiatives leads us to conclude that available models and frameworks are more than adequate for dramatically expanding the scope for evidence-based policy-making in the stabilization sector.

Our analysis is motivated by recent advances in the agenda to apply principles of evidence-based policymaking to U.S. government assistance overseas. In February 2011, the U.S. Agency for International Development (USAID) released a new evaluation policy that called for substantial improvements in the rigor of evaluations of development assistance. The new policy reflected the growth of a "scientific" evaluation consensus in Washington, a consensus that has been driven in part by parallel advances in the social sciences over the past decade. In a January 2011 guidance statement on stabilization programming, USAID Administrator Rajiv Shah indicated:

> Every activity is an opportunity to learn what works, what does not, and why. Finding innovative solutions that can be scaled for impact requires methodical experimentation. Create mechanisms to self-critique, continuously adapt, and share lessons learned, especially with our interagency, international, and cross-border colleagues. Build a culture that rewards adaptation, innovation, and problem solving. (USAID 2011)

The quote from Shah echoes almost verbatim what the International Initiative for Impact Evaluation (3ie) proposes as the objective of impact evaluation, namely to "measure the net change in outcomes that can be attributed to a specific program. Impact studies help inform policy as to what works, what does not, and why" (International Initiative for Impact Evaluation,2011).

Up until now, impact evaluations have tended *not* to be applied to U.S. government-sponsored stabilization programs. Thus, a move toward doing so would represent a marked shift in standards and procedures for those working with the U.S. government in this sector. The first section below explores the U.S. government experience in more detail. Nonetheless, for all of the reasons that evidence-based policy making has made advances in other sectors including public health, education, and economic development, there is potential to learn many lessons about how to make stabilization policies more effective and efficient. The experience of other agencies suggests that with some creativity and attention to contextual details, rigorous impact evaluation is indeed feasible in this sector. Thus, the next two sections below examine strategies employed for (1) designing studies that allow one to attribute impacts to the stabilization intervention in question and (2) measuring the complex outcomes that are the targets of stabilization policies. Because there are few examples of U.S. stabilization program impact evaluations, we draw our examples from stabilization programs implemented by other agencies, including the World Bank and various international non-governmental organizations (NGOs). The hope is that this discussion will demonstrate that a stronger orientation toward evidenced-based policy could be a reality in U.S. sponsored stabilization programming.

2. U.S. experience in evaluating stabilization programs

In this section, we review evaluation practices for stabilization interventions over the last two decades by USAID, in particular those associated with the Office for Transition Initiatives (OTI), the 1207 Funding mechanism, and by the United States Department of State Bureau of International Narcotics and Law Enforcement (INL), which has become especially active in rule of law

programming in fragile states.[4] We discuss some of the shortcomings of the evaluation practices to date for these types of interventions. By focusing on stabilization interventions, we are looking at a subset, albeit a large subset, of interventions in post-conflict or fragile state environments. As noted above, we consider those interventions that are designed primarily to strengthen the durability of peace. Many of these interventions include other elements or state additional objectives, for example reconstruction or civil society development. We distinguish stabilization interventions from programs that have another primary development objective, for example, public health, but are implemented in unstable environments. Based on the published program descriptions of those interventions we identified, we categorize the interventions into ten sub-elements of the Foreign Assistance Standardized Program Structure and Definitions (U.S. Department of State, 2010). See Appendix 7.1 for the full definitions. Understanding the key elements of these categories of interventions is the first step to considering impact evaluation designs.

We conducted an extensive search for publicly available evaluations of U.S. stabilization programs conducted over the last two decades, and found over 50 evaluations. While most evaluations include a discussion of what the evaluators observed as impacts, the larger part of the evaluations are focused on the process of the projects and the performance of the implementers. One typical example is "Final Evaluation: The OTI Program in East Timor" (Clark 2003). OTI's program included peace structures, peace messaging, and consensus-building activities. The evaluation methodology involves three evaluators traveling to East Timor for three weeks at the end of the three-year program. They collected data through key informant interviews, grantee focus groups, comprehensive document review, analysis of project databases, and focused field trips outside of the capital. The evaluation's conclusions and lessons learned include: "the psychological and political impact of OTI programming, especially in the early days of the transition, should be acknowledged as making a substantial contribution to calming turbulent seas"; "the procurement mechanisms and small grants authorities exercised by OTI were key to its success"; "OTI succeeded in earning USAID and the U.S. Government considerable credit in the eyes of the Government and of donor officials, as well as of the general public"; and "USAID was well-served by the personnel choices made by OTI and DAI in East Timor".

We did find some "impact evaluations" that focus on program outcomes and collect data with the intent to measure the changes brought about by the program. One example is "Impact Evaluation: Youth Reintegration Training and Education for Peace Program" by Fauth and Daniels (2001). In this study, the authors randomly sampled the program participants and surveyed them at the end of the program. They asked questions to directly assess whether the different training modules met their objectives; for example, they asked participants whether they were better able to manage conflict after receiving training on conflict management. Not surprisingly, 99 per cent reported they were. 98 per cent reported being better able to solve problems after receiving problem solving training; 99 per cent reported being more self-aware after the self-awareness training; and so on. In an attempt to examine what might have happened absent the intervention, the authors asked participants whether they have engaged in certain activities since the training and then asked them if they would have done so without the training. Again not surprisingly, the majority of participants responded that without the intervention they would not have engaged in these activities. The authors conclude, "The program has made a significant difference in the lives of the participants, and they have done things they would not have done if they had not participated." (p. 13)

While this evaluation provides some interesting and useful qualitative information, particularly in the form of case studies and respondent comments, it cannot attribute impact to the

[4] We searched public sources for all descriptions of current and planned projects or programs. We added to this list the past programs for which we found evaluations.

intervention as it does not measure the non-intervention, or counter-factual, state. In addition, it relies solely on self-reported outcome data. These shortcomings severely limit our ability to learn what works, when, and why based on the evaluation. Of all the evaluations that we found of U.S. government-funded stabilization interventions, we only identified one that carefully examines the counter-factual in order to attribute impact (Paluck 2009a).

The U.S. government is well aware of the limitations of existing evaluation efforts. U.S. Government Accountability Office (GAO) reports routinely note the lack of a thorough and effective evaluation strategy in international development programs and missions. In fact, our survey of GAO reports on stabilization interventions shows that evaluations are typically thwarted by a lack of evaluation design, resources, logistics, or security issues. For example, a 2005 GAO report on Afghanistan security programs notes that the German, U.S., and Afghan governments providing police training fail to evaluate the performance of police trainees after graduation in the more remote areas of Afghanistan (GAO 2005). A 2010 GAO report on the management of USAID programs in Afghanistan concludes that USAID has not consistently followed its own guidelines for monitoring and evaluation and recommends that USAID, "fully assess and use program data and evaluations to shape current programs and inform future programs." (GAO 2010, p. 3) These limitations are not unique to international development programs. A 2009 GAO report on Program Evaluation states that while federal agencies track progress towards goals through performance measures, "few seem to regularly conduct in-depth program evaluations to assess their programs' impact or learn how to improve results." The GAO goes on to say that "randomized experiments are considered best suited for assessing intervention effectiveness where multiple causal influences lead to uncertainty about program effects and it is possible, ethical, and practical to conduct and maintain random assignment to minimize the effect of those influences" (GAO 2009).

3. International experience in impact evaluations of stabilization programs

To explore the scope for impact evaluations of stabilization interventions, we conducted a comprehensive search for impact evaluations of stabilization interventions implemented by any donor or government.[5] We identified over two dozen studies, some still ongoing, but unfortunately only covering seven of the ten program categories. See Appendix 7.3 for the full list. In this section, after a brief overview of experimental and quasi-experimental methodology, we examine the general characteristics of this group of impact evaluations, and then we elaborate specific strategies for estimating the impacts of stabilization interventions by focusing on four examples.

Numerous discussions are available on general strategies for estimating the impacts of interventions.[6] The basic premise is that estimating the impact of an intervention versus no intervention requires that one establish a valid "counter-factual." Thus, in estimating the impact of an intervention, we mean to estimate, on average, how outcomes among the group of beneficiaries who received the intervention differ from what would have happened had they instead *not* received the intervention. That latter condition is contrary to fact, and therefore represents the "counter-factual" condition.

By statistical reasoning, the most reliable method for establishing such a counter-factual is to take a large pool of potential beneficiaries and randomly assign some to receive the intervention and

[5] The search was not systematic according to the standards of systematic reviews (Higgins and Green 2008). It was, however, extensive. Please see Appendix 7.2 for a description of the search methodology, including the key words used and the databases searched. We identified ongoing studies primarily through direct contact with the researchers.

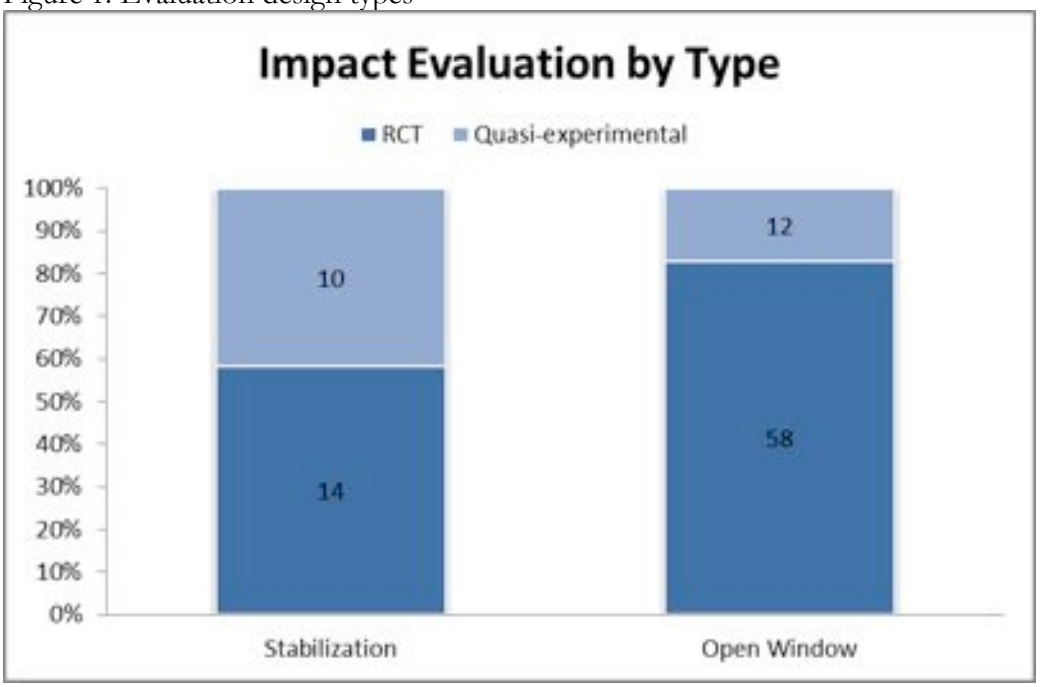
[6] See Samii (2010) for a general discussion that is specifically relevant for post-conflict and humanitarian programming.

others to not receive the intervention. Then, both those who receive the intervention (the "treated" group) and those who do not (the "control" group) are examined over time. The combination of randomization over a large pool of potential beneficiaries ensures that those who receive the intervention will, on average, resemble those who do not. The latter group thus provides a valid counter-factual for the group that receives the intervention, and vice versa. Measuring outcomes over time on both the intervention and non-intervention group allows one to avoid the "before-after fallacy" (see Appendix 7.4). The very same logic holds when one is interested in studying the effectiveness of one type of intervention relative to another. Sometimes, randomization is either infeasible or unethical, in which case non-randomized, "quasi-experimental" methods may be used to establish a counter-factual. All quasi-experimental methods require that the analyst impose additional assumptions and therefore make larger leaps of faith in order to justify the counter-factual. This is why researchers sometimes refer to randomized evaluations as the "gold standard" and consider quasi-experimental evaluations as imperfect approximations to this gold standard.

3.1. General features of stabilization impact evaluations

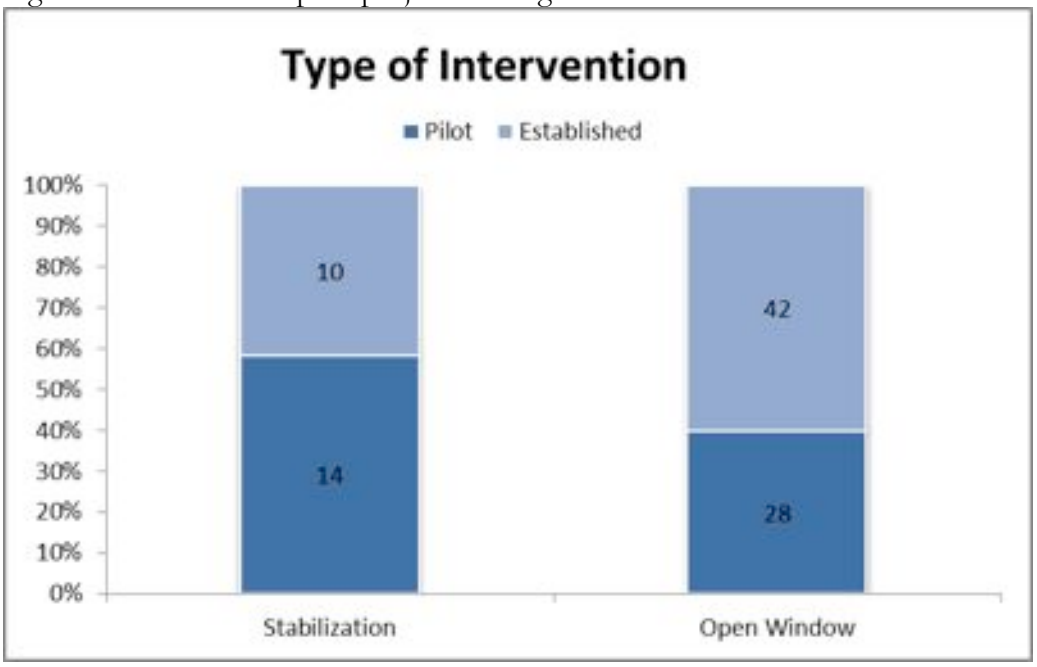
In the set of identified impact evaluations of stabilization interventions, we find that researchers are able to implement randomized or cluster randomized treatments in almost 60 per cent of the studies, while quasi-experimental approaches account for 40 per cent of the studies. To get a sense of how this breakdown compares to impact evaluations of development programs in general, we can consider the set of 70 grants awarded by 3ie in its three Open Window competitions. This sample is certainly not a random sample of the population of relevant impact evaluations, but it is a sample of fairly recently funded evaluations where the grant program uses the same definition of "impact evaluation" that we do in this paper and states no preference for sector of intervention or for experimental versus quasi-experimental designs. Figure 1 presents the breakdown by type of evaluation design for the two sets of impact evaluations. The figure shows that the evaluation designs from the Open Window applications employ experimental methods in much greater proportion than for the stabilization evaluations. This casual evidence of higher prevalence of quasi-experimental designs is not surprising given that one of the objections cited to conducting impact evaluations of stabilization programs is that randomized treatment is often not feasible or ethical.

Figure 1. Evaluation design types



Our sample of impact evaluations reveals another prevalent strategy for using rigorous methods in these environments—conducting evaluations of pilot programs rather than established program. In fact, the majority of the studies we found examine pilot programs. Figure 2 shows the comparison with the Open Window applications.

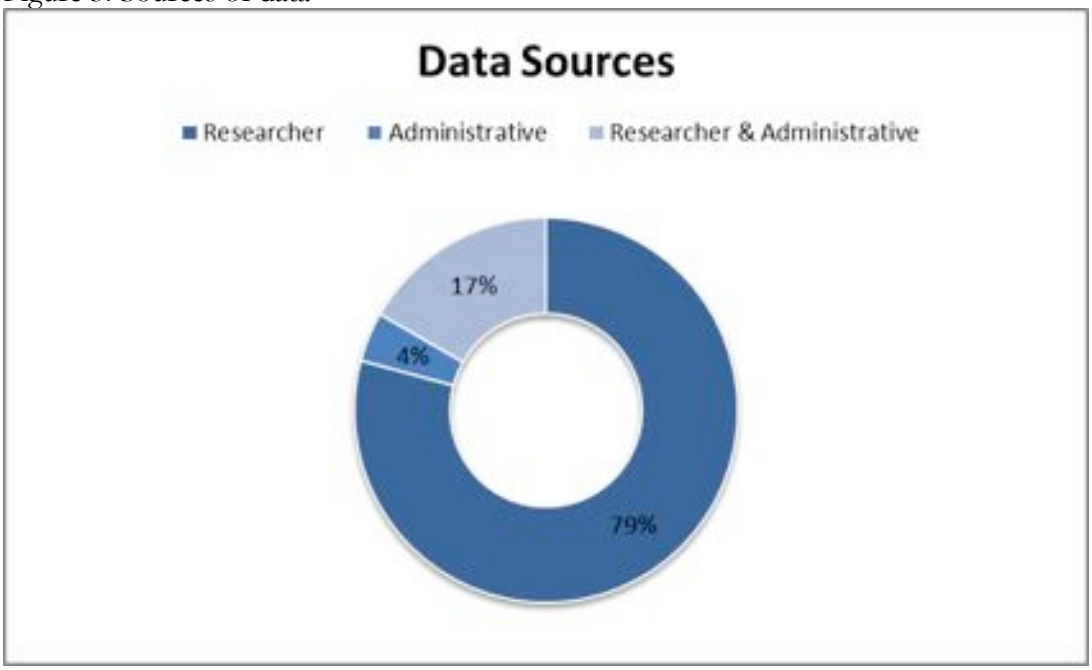
Figure 2. Prevalence of pilot projects among interventions evaluated



Two other tendencies among these studies are consistent with the challenges of conducting impact evaluations in these contexts. Only one of the stabilization studies identified relies on

administrative data. The vast majority rely exclusively on primary data collected by the researcher, particularly for the measurement of outcomes. A few use both primary and secondary sources. Figure 3 shows this breakdown.[7]

Figure 3. Sources of data



Finally, we see striking differences in the units of analysis used between the stabilization and the Open Window studies. Here the differences reflect the nature of the interventions, which are often directed at individuals with a view to changing individual behaviors (such as ex-combatants) or beliefs and perceptions (such as animosity to another group). Many stabilization interventions are delivered at the community level, such as peace dividends, and 13 of the 24 studies report community-level outcome measures in addition to household or individual measures. The two pie charts in figures 4 and 5 also present the average sample sizes for the stabilization evaluations and the proposed designs in the Open Window applications. Not surprisingly, the averages are smaller for the stabilization studies where we would expect data collection to be more difficult.[8]

[7] We are still coding the data from the Open Window applications for this variable, but based on our general knowledge of these studies, we believe the prevalence of primary-data-only studies to be greater among the stabilization set.

[8] The sample sizes for the Open Window studies will likely change during the research. If we believe that sample sizes are more likely to decrease during study implementation, then the difference between the averages for the stabilization evaluations and the Open Window evaluations is biased up.

Figure 4. Units of analysis and average sample size, stabilization studies

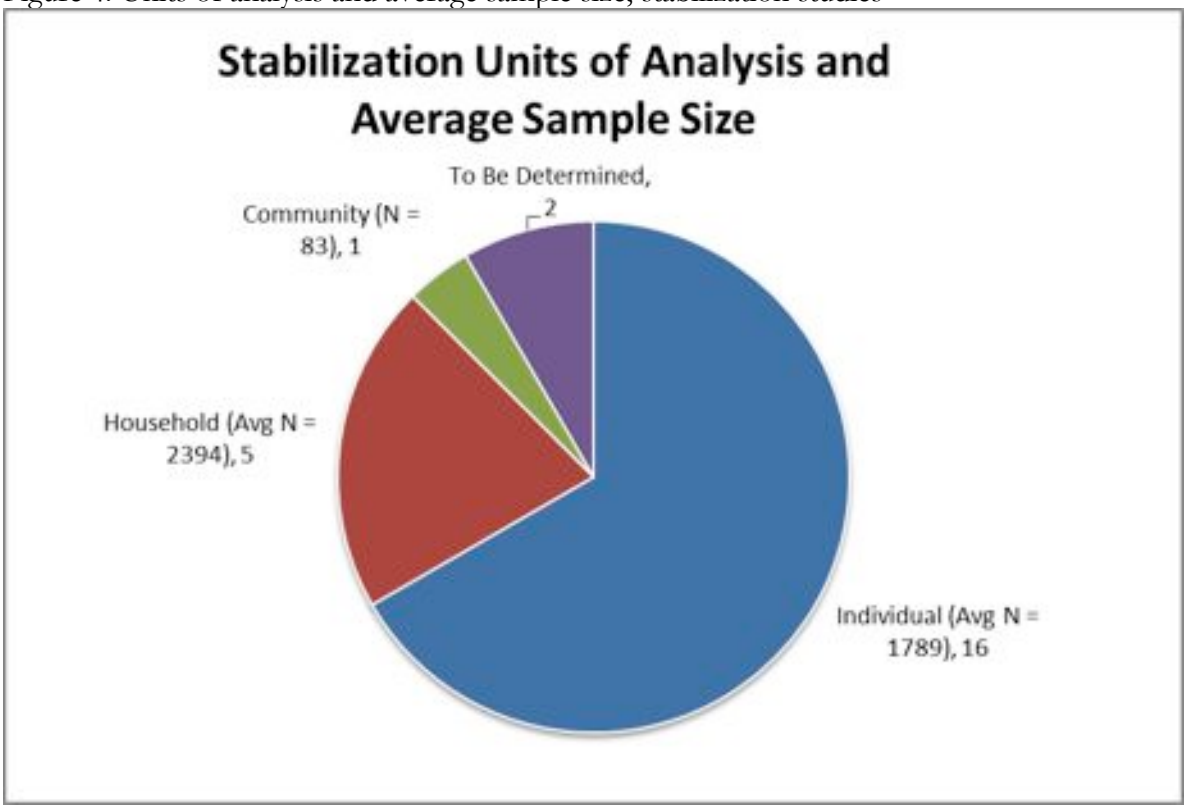
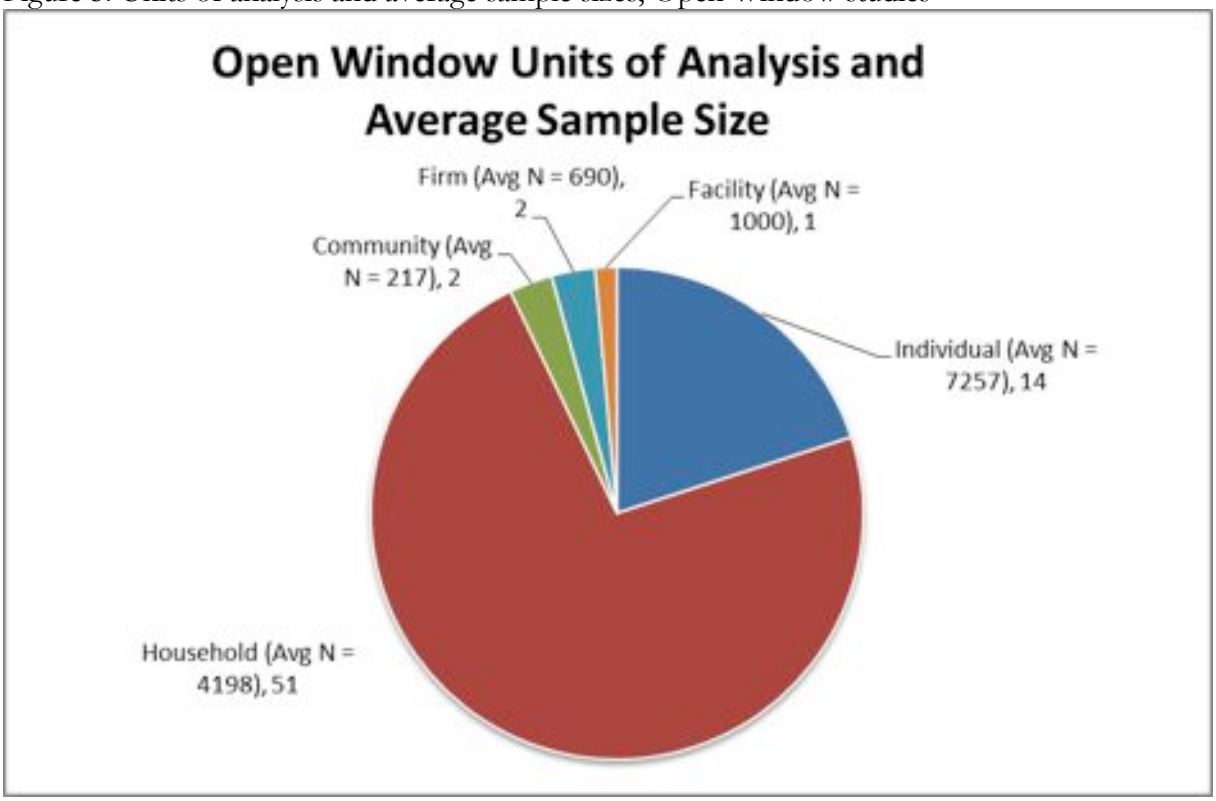


Figure 5. Units of analysis and average sample sizes, Open Window studies



Considering the existing impact evaluations of stabilization interventions as a group reveals some prevalent features. These features are not surprising given the nature of stabilization interventions and the challenges of conducting impact evaluations in these unstable environments. Anticipating these features can help commissioners of impact evaluations to select the programs to be evaluated and select the researchers to evaluate them, as well as manage their expectations about the results. Experimental designs, although still in the majority, appear to be feasible less often for stabilization interventions. The careful and creative use of quasi-experimental methodologies will therefore be important for increasing the use of impact evaluations for learning about stabilization. Researchers in the field of stabilization impact evaluations appear to rely more often on pilot projects in order to conduct an evaluation. Impact evaluations of pilot programs, particularly where untested and innovative approaches are being studied, can be very useful, as recognized in the USAID Evaluation Policy (2011). Pilot projects may not always fully mimic the conditions of a scaled-up program, however, so the results should be interpreted cautiously.

Impact evaluations of stabilization interventions typically require the collection of primary data. There are examples where pre-conflict data can be used for tasks like sampling and matching. Given the difficulty of data collection, it is not surprising that we find that these impact evaluations also appear to be based on smaller average sample sizes, which may limit the analysis or require more sophisticated techniques.

3.2. Examples of strategies for identifying impacts

We illustrate the scope for evaluating stabilization intervention by taking a detailed look at the strategies employed in four selected impact evaluations of stabilization interventions. For this analysis, we have selected examples that illustrate a variety of techniques, contexts, programs, and practical challenges. The first two examples use randomization to establish causal attribution. We present these first as they demonstrate that near-ideal strategies can be applied in the stabilization context, given sufficient advanced planning and cooperation among implementing partners. The next two examples use "quasi-experimental," non-randomized approaches to establish causal attribution. While quasi-experimental approaches are appealing in many ways, these examples usefully demonstrate that quasi-experimental designs require considerable technical sophistication if they are to produce reliable estimates of program impacts. Our discussions of each example follows a common structure that characterizes the program and intended impacts, key features of the strategy for establishing causal attribution, the strategy for measuring outcomes, and the substantive findings in terms of estimated impacts.

3.2.1. Peace dividends and community directed reconstruction in Sierra Leone

Casey, Glennerster, and Miguel (2011) study the impact of "GoBifo," a 2006-2009 community-directed post-conflict reconstruction intervention in rural Sierra Leone, also known as community-driven development (CDD). The World Bank was the primary funder for the intervention, although the Government of Sierra Leone managed its implementation directly. The intervention made block grants of $4,667 available to each of the 118 villages in the treatment group. The "community-directed" aspect of the intervention was based on the fact that development committees in beneficiary communities were assigned to draft village development plans. These were then submitted to regional government offices for approval, upon which the block grant would be disbursed to implement the plan. The intended impacts of the program were to deliver material benefits to beneficiary villages, to ensure that such benefits reached historically marginalized groups (here, women and youth), and also to enhance the collective action capacity of the villages for

solving future village-level problems. As Casey et al. write, "once an organizing body is in place and residents have reached consensus on local priorities, the next collective project should be less costly to identify and execute" (p. 10).

The GoBifo study established causal attribution based on the fact that out of a pool of 236 eligible villages, 118 were randomly assigned to be beneficiary communities and the other 118 were assigned to receive no support during the four year program cycle, and thus serve as a control group. The pool included areas with limited NGO presence. The pool of candidate villages was also restricted to villages of appropriate size for a community-directed program. As such, *all* of the villages in the pool could be considered as being worthy of assistance. But resources were limited to providing benefits to only *half* of them. We imagine that this precise circumstance is a common one that arises in stabilization programming. By using randomization as the mechanism to determine which of the eligible villages would be selected as beneficiary villages, the program allowed for a rigorous assessment of impact. In addition, random selection of beneficiaries from a pool of eligible candidate villages was ,arguably, a fair way for scarce program resources to be allocated. The random assignment was done within wards to ensure balance between treated and control villages in terms of region, politics, and ethnicity.

Casey et al. designed an outcome measurement strategy that consisted of a combination of household surveys, focus group discussions, and "structured community activities" undertaken in both control and treated communities. The structured community activities were quite innovative, consisting of a matching grant program, a meeting to decide on how a free resource would be allocated, and the provision of free tarpaulin for villages to use in whatever manner they liked. Each of these activities provided an opportunity for the villages to reveal, behaviorally, their collective action capacity and whether marginalized groups would receive access to village benefits. Another exceptional aspect of the design of this study was that the researchers pre-registered their outcome analysis plan at the start of actual programming. In the section below on measurement, we discuss the benefits of establishing an early outcome measurement strategy.

The study found that the intervention had mixed impacts. While the intervention succeeded in delivering material benefits to the beneficiary communities, impacts on collective action and inclusion of marginalized groups were not significant. That is, from the surveys, focus groups, and activities, the treated communities did not exhibit significantly better collective action capacity or tendency to include marginalized groups in decision making or sharing of benefits. The results raise important questions about the assumptions guiding community-directed reconstruction and development programs, and pave the way for new thinking about how social outcomes such as inclusion and collective action capacity may be targeted more directly.

The Casey et al. evaluation was exceptional in its rigor. Some readers may wonder whether the Sierra Leone setting was unusually conducive to such a rigorous evaluation. Our sense is not: similar randomized evaluations have been carried out to evaluate community-directed reconstruction programs in Afghanistan (Beath et al., 2010), Democratic Republic of Congo (Humphreys, 2008), and Liberia (Fearon et al., 2008; 2009). Without question these are among the most challenging settings for either programming or research. Thus, the factors that limit the application of rigorous methods appear to be less the difficulties associated with the context, and more the goals, planning, and cooperation among those involved in the programming.

3.2.2. Peace messaging in Rwanda

Paluck (2009b) studies the impact of a reconciliation-messaging soap opera, "New Dawn," broadcast nationally over radio in post-war Rwanda over the course of 2004. New Dawn was produced by the international NGO, La Benevolencija, whose goal was to create a program to

promote inter-ethnic reconciliation after genocide and war. The intervention targeted individuals in post-conflict regions and intended to change their beliefs about ethnic out-groups, perceptions about prejudicial behavior and ethnic hostility, and willingness to be more assertive in voicing individual opinions rather than deferring to authority.

At first glance, it would appear difficult to establish causal attribution between a nationally broadcast radio program and the outcomes of interest. Everyone in the country had, in principle, access to the program and a comparison between those that actually listened and those that did not would likely result in spurious findings, as listeners and non-listeners probably differ in ways that did not necessarily have anything to do with the radio program. Paluck's strategy for overcoming this hurdle was to establish listening groups, and then to randomly assign half of those groups to listen every week to New Dawn and the other half to listen every week to an alternative, non-reconciliation-based radio program at the same time that New Dawn aired. (The alternative program was a health program.) In this way, Paluck was able to construct a control group that allowed her to isolate the effects of the content of the New Dawn soap opera. The randomization plan matched 12 communities into pairs, and then a coin flip was used to determine which half of a pair would listen to New Dawn and which would listen to the alternative program. This created a matched-pair randomized evaluation design.

In a manner similar to Casey et al., Paluck's measurement strategy included a combination of survey-based measures, focus groups, and observed behavioral measures. The survey measures included questions that asked subjects about the appropriateness of various types of behavior involving co-ethnics and non-co-ethnics – for example, whether it would be okay for one's daughter to marry a man from the other ethnic group. Transcripts from the focus groups were coded in order to assess degrees to which participants were willing to express dissenting opinions. Finally, a group-sharing task was constructed to provide a behavioral measure of the degree to which members of treated and control listening groups were willing to voice individual opinions. In the task, groups were provided with a cassette player and a set of audio cassettes; and enumerator observed the manner in which the groups decided on access to the player and cassettes. The results of the study showed that the program produced strong effects on subjects' perceptions of what types of behavior others would deem acceptable, but no discernible impact was recorded on subjects' own beliefs about what was acceptable. Strong effects were also estimated on subjects' willingness to dissent or voice individual opinions during group deliberations.

What the New Dawn study demonstrated was that, with some creativity, a randomized trial could be fashioned from a national level program. One can imagine other possibilities for such an approach. For example, sometimes, national institutions are put in place but awareness or understanding of these institutions may be weak. An assessment of the impact of such institutions might be constructed by randomly selecting communities to receive intensive public awareness campaigns.[9]

3.2.3. Ex-combatant reintegration in Burundi

Gilligan, Mvukiyehe, and Samii (2011) study the impact of a national ex-combatant reintegration program in Burundi in 2006-2007. The program was financed by the World Bank administered

[9] A study in progress by Chong et al. (2010) uses a strategy along these lines. In the Chong et. al. study, citizens are randomly assigned to receive intensive public information on municipal spending and corruption in order to assess how spending and corruption affect citizens' political participation and willingness to hold politicians accountable. Interestingly, the study finds that increasing awareness about corruption causes voters to abstain from voting rather than motivating them to use their vote to hold officials accountable, whereas information about spending increases participation and support for incumbents.

Multi-Country Demobilization and Reintegration Program (MDRP), and consisted of a caseload of 23,000 ex-combatants, including 14,000 ex-rebels. The program benefits included 18 months of a reinsertion allowance, livelihood and psychological counseling, and a socio-economic reintegration package. The evaluation was limited to studying the impact of the socio-economic reintegration packages on ex-rebels' reintegration prospects. Program beneficiaries could choose between a package that included support for continued formal education and a package that included training and in-kind material assistance to start a livelihood. Over 90% of beneficiaries elected to receive the latter, and so the evaluation focused on the impact of such livelihood assistance both on establishing a productive, legal livelihood ("economic reintegration") and on adopting a social and political disposition that accepted civilian status and the legitimate authority of the country's democratically elected leaders ("social and political reintegration"). The assumption behind the program was that both types of reintegration were important in removing motivations to return to arms. The intended impact of the program was to boost economic reintegration and, as a consequence, induce political reintegration.

The program was intended to provide benefits to all ex-combatants, and so there was no way that the program could deny benefits altogether to some ex-combatants through random assignment. That being the case, Gilligan et al. exploited the fact that the roll out of the program occurred according to different timelines in different regions of the country. In two out of three programming regions, programming was initiated in the autumn of 2006, whereas in the third, it did not begin until about a year later.[10] The would-be beneficiary ex-combatants in the delayed region thus provided a pseudo-control group relative to the ex-combatants in the regions that began to receive benefits in 2006.

There are some important limitations to this evaluation design, however. First, while the reasons for the delay in the one region had nothing to do with types of ex-combatants that resided there, there are incidental differences that distinguish the economic and political context in this region relative to the other two regions, and there are also incidental differences in the demographics of the ex-combatants in this region relative to the other two regions. Thus, Gilligan et al. had to use various statistical adjustments to account for these incidental differences. This requires that we accept the validity of the statistical adjustments, which may require a leap of faith for some. Second, the staggering in the roll-out of the program was of relatively short duration, and as a result only relatively immediate impacts could be assessed. Ideally, one would want to examine the long-term impacts of the program, but this evaluation design does not permit such possibilities. Third, it is possible that a staggered rollout design can be tainted by the fact participants with delayed treatment nonetheless anticipate that they will eventually receive benefits and as a result change their behavior.[11] Thus, the pseudo control group acts differently than if it had no expectation of eventually receiving the program and does not serve as a true counter-factual. Certain assumptions have to be believed in order for one to believe that this design provides a proper counter-factual to the program. Nonetheless, as Gilligan et al. argue that even with these limitations, the evaluation design is very useful because it allows us to assess possible program impacts and avoid bias that is present in comparisons of those who voluntary chose to participate in a reintegration program.

---

[10] The staggered roll-out was actually unintended in this case, occurring because of a contract dispute between the national program and the implementing organization in the delayed region. However, in principle, it is often necessary to stagger roll-out intentionally, and in these cases, one can use such staggering in construct a quasi-experiment along the lines of this one.

[11] Of course this anticipation may occur in many randomized treatment designs as well. For ethical reasons, non-treated groups often do receive the treatment after the end of the study. Also, when programs are pilots, those in the control may realize that the program is one that may be implemented in full.

The measurement strategy for the evaluation was based solely on individual-level surveys. Thus, ex-combatants in the treated and control groups were asked questions about their income and livelihoods, as well as their attitudes toward civilian versus combatant life, satisfaction with the peace accords, and support for current government and institutions. The results of the analysis showed a large reduction in poverty incidence and a moderate increase in the attainment of semi-skilled or skilled occupations over unskilled, suggesting that the program did cause substantial economic reintegration gains. However, there were no perceptible effects on the measures of social or political integration.

Opportunities to use phased roll-out in this manner likely arise often in programming contexts, and with due care they can be used to test out a program concept before full-scale implementation. A slightly different approach would be to experiment with various alternative program concepts at a pilot level and use the results to determine which of the program concepts under consideration is most suitable for full-scale programming. While ex-combatant reintegration programs are often considered as "emergency" programs that suffer from exigent circumstances and require haste, the experience relayed by the Gilligan et al. study suggests that there can be sufficient flexibility to pilot ideas and use phased roll-out to refine the program concept.

3.2.4. Peace dividends and community directed reconstruction in Aceh

Barron, Humphreys, Paler, and Weinstein (2009) study the impacts of a 2006-2007 World Bank-sponsored program in Aceh, Indonesia to reconstruct communities and reintegrate displaced households and ex-combatants after 30 years of conflict As in the GoBifo intervention studied by Casey et al. (above), the Aceh reconstruction and reintegration program used a community-directed reconstruction mechanism to determine how resources would be allocated locally. The intended impacts of the program were to improve average material well-being in beneficiary communities; to improve social cohesion across various lines of cleavage, including ex-combatants and non-combatants, as well as non-displaced and resettled displaced households; and to improve trust in Aceh's government.

The Bank did not have resources to provide assistance to all communities in Aceh, and so some beneficiary selection process was required. Above, we have seen how randomization and staggered roll-out have been used. In this case, the Bank used another type of beneficiary selection process that still allows for rigorous impact evaluation so long as adequate records are kept. The Bank used a targeted selection process: they targeted the most conflict-affected areas, but restricted access to communities that had local administrative capacity sufficient to manage the block grants. This targeting meant that, by construction, beneficiary and non-beneficiary communities differed in important ways—namely, in terms of their conflict exposure and local administrative capacity. Thus, the beneficiary and non-beneficiary communities could not be treated as if they were randomly assigned. However, the Bank's selection process offered an advantage in that detailed records were kept on exactly how and with what data the beneficiary communities were selected from the overall pool of communities. Using this information, Barron et al. were able to identify exactly which non-beneficiary communities had a very strong chance of being selected had more resources been available, and which of the beneficiary communities would have been excluded were fewer resources available. These "marginal" non-beneficiary and beneficiary communities can thus be treated as if they had been randomly assigned to beneficiary and non-beneficiary status.

Barron et al. used a combination of surveys and behavioral games to measure outcomes. In the surveys they recorded subjective measures, such as village leaders' assessments of poverty rates and social conflicts in their village, and household members' perceptions of their material and social well-being and satisfaction with government. They also used the surveys to collect more objective

measures, including household assets, land use, wages, employment, education, and health, as well as public goods-related outcomes reported by village heads. A behavioral game measure asked household survey respondents to decide how to split a payment that they received for participating in the survey between themselves and a fund for local government development projects. This was intended as a measure of trust in local government. The evaluation found numerous positive welfare impacts—for example, an 11 percentage point decrease in perceived poverty as well as asset and land use improvements. However, they found no discernible impact on trust toward government and little impact on social cohesion. In fact, in the latter case, the evaluation found that beneficiary communities tended to be less accepting of ex-combatants.

The key take-away from this case study is that, even with a targeted selection strategy, rigorous impact evaluation is possible so long as beneficiary selection is based on (i) clearly recorded steps and (ii) background data that are accessible to evaluators outside the program. With these two elements in place, those evaluating the program can retrace steps and replicate the entire beneficiary selection process. Not only is this a transparent and fair way for beneficiaries to be selected, but it provides a firm foundation for constructing a counterfactual. Our experience suggests that this kind of approach may be applicable in many settings where randomization is not possible. Indeed, some kind of targeting is routine in many international development interventions. The problem has been that the beneficiary selection processes are often imprecise, based on incomplete or low-quality data if any data at all, and have little or no systematic record keeping of the steps involving in narrowing the pool of potential beneficiaries to those actually selected. Thus, the opportunity to apply this sort of strategy does not require so much of a change in the general approach to beneficiary selection, but rather a systematization of the general approaches already used.

4. Principles for measuring stabilization outcomes

In this section, we address principles for constructing valid measures of stabilization outcomes and using such measures to estimate impacts. As demonstrated by the descriptions in section 2.1, stabilization interventions target a wide range of complex, multi-dimensional outcomes and sectors. Furthermore, "stability" can be assessed at the individual, household, and national levels of analysis. While employment or consumption-based poverty rates might suffice as primary measures for evaluating poverty eradication programs, in many cases stabilization programs intend to bring about changes that are less clearly defined at the individual or household levels (while simultaneously supporting employment and poverty eradication). Moreover, while agencies may commission an evaluation of the contributions of an intervention to national-level stability, national outcomes depend on a wide range of social, economic, and political factors among which a given program is typically a miniscule component. Considerable thought should go into ensuring that the right metrics are chosen and that they are being applied at the right level of analysis. The appropriate focus will depend on the nature of the intervention.

A few initiatives have been undertaken to catalogue outcome measures for stabilization interventions. Perhaps the most comprehensive to date is the MPICE framework, outlined in Dziedzic, Sotirin, and Agoglia (2008; 2010). MPICE consists of a large catalog of indicators grouped into the themes of security, politics/governance, rule of law, economic conditions, and social well-being. The indicators are based mostly on a "macro-stability" perspective (that is, at the country or region level), but for impact evaluations we typically require measurements that can be assessed at the "micro" level (that is, at the level of individuals, households, or communities). Nonetheless, many of the indicators could be suitably redefined at the micro level for use in impact evaluation. The MPICE framework is quite thorough on social and political indicators, but less so on economic indicators. The economic indicators presume a level of development that is beyond many

stabilization contexts (e.g., it focuses on industrial activity and formal employment, whereas non-mechanized agriculture predominates in most societies that host stabilization interventions).

The World Bank's Living Standard Measurement Survey modules for poor countries may be more appropriate as measures of economic vitality. Other proposals include a useful contribution by Christia (2010) on a community-level stabilization assessment framework based on *routine* micro-level activities that could be interrupted by insecurity. Geared to the Afghanistan context, Christia's suggestions included measures of proportion of local officials who live and work in their home communities, rates of school attendance, costs of transporting goods into the community, levels of activity in local markets, and willingness of locals to share information on sources of insecurity. The set of measures may vary from context to context, but the general principle is to measure recovery in the rates of routine activities that may be sensitive to insecurity. A working paper by Brück et al. (2010) reviews household- and individual-level survey-based indicators of violence and insecurity. Our review of U.S. government F-indicators suggests that they would not provide an adequate framework for impact evaluation, as they focus only on outputs and programmatic outcomes that may give little indication of genuine impact.

The set of evaluations that we have studied in the previous section leads us develop a typology of outcomes and associated measures. We distinguish between outcomes that are *subjective* attitudes and perceptions versus outcomes that are *objective* behavior. Subjective attitudes and perceptions include individuals' grievances ("are you getting what you deserve?"), normative beliefs ("is it okay for your kids to marry members of the out-group?"[12]), and other hard-to-observe conditions ("how much do you worry about theft in the night?"). These are typically measured with survey interviews. In contrast, objective behavior may be self-reported ("did you vote in the last election?") or behavior that is directly observed by members of an evaluation team (actually consulting voter turnout records, if they are available).

While attitudes and perceptions may be precise and easy to collect, they suffer from many known problems. Survey interviews are "obtrusive" in that they require evaluators to intrude into the natural environment of subjects. Subjects are thus very aware that they are being assessed. In addition, incentives for respondents to reply honestly in a survey may not be great. This combination introduces the possibility of "social desirability" bias, if subjects respond in a manner that they think the interviewer will receive favorably, rather than providing honest answers. These biases may be especially severe in the stabilization context because subjects often have reason to think that the result of the evaluation will influence the likelihood of continued assistance. Attitudes and perceptions can be "noisy" measures in that people change their minds. In addition, the questions may be abstract, and so evidence of impacts on an attitudinal variable may not imply any material changes. Finally, it is difficult to judge the substantive importance of attitudinal impacts, given that the scales on which they are reported are often arbitrary. Statistically significant changes on a Likert, Guttman, or binary scale may not provide a clear indication of substantive significance. Self-reported behavior measures share many similarities with attitude- and perceptions-based measures. They too are easy to collect. But again, they are vulnerable to social desirability bias. In addition, self-reported behavioral measures rely on subject memory, which is prone to error.

The drawbacks of attitudinal measures and self-reported behavior have led many social scientists to place a premium on using observed, preferably unobtrusive, behavior to measure impacts (Webb, Campbell, Schwartz, and Sechrest 2000). Observed behavioral measures include those that are produced in "real world" settings, those that are produced in "artificial" settings, and those that are somewhere in between—what we will refer to as "artifactual" settings. The studies reviewed above contain a mixture of each of these types of measures. "Real world" observed

[12] Such a normative belief is used in Paluck (2009b).

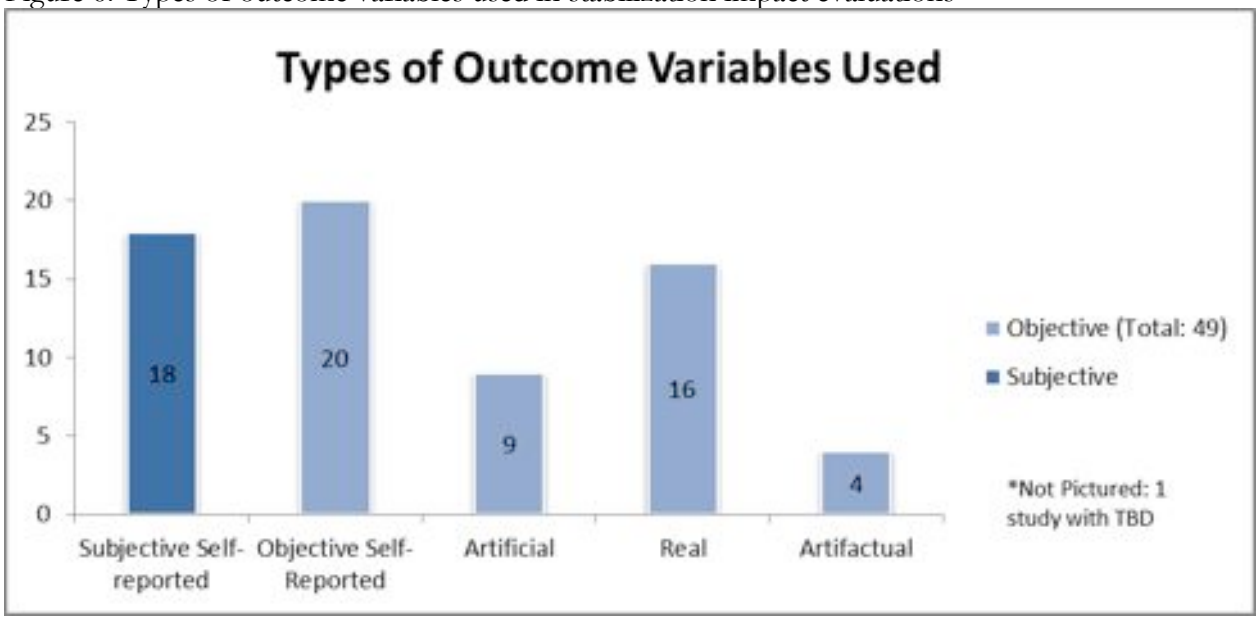
behavior might be measured via, for example, satellite readings of agricultural activity, crime statistics, or mobile phone company records of credit purchases. Such measures tap directly into material conditions and, if systems are in place to obtain such measures reliably, then they are often readily accessible. However, such systems may be expensive to set up. In addition, measures such as these may suffer from the problem of "overdetermination"—that is, they may be affected by many forces of which a given intervention is only a small contributor, or there may be many conceivable causal pathways linking an intervention to such outcomes.

For this reason, social scientists also rely on "artificial" and "artifactual" behavioral measures to assess the relative importance of different potential causal pathways. Artificial behavioral measures include economic games, which try to overcome the obtrusiveness problem by providing incentives for subjects to act sincerely. Economic games can help to isolate impacts on hard-to-observe traits. In this way, games can be used in combination with real world measures to try to tease out whether one or another causal pathway is responsible for changes that we observe on the real world outcome. For example, Fearon, Humphreys, Weinstein (2009) use economic games to study the impact of the a CDD program in eastern Liberia to measure impacts on community-members' sense of obligation to each other. These researchers also measured income, consumption, and other aspects of material welfare (as reported in Fearon, Humphreys, and Weinstein (2008)). They find that feelings of obligation, as measured in the games, increased but there was no substantial increase in material welfare. This combined measurement strategy has the virtue of allowing us to draw a nuanced conclusion: while the program produced gains in terms of social relations, this did not translate into meaningful effects on material welfare. The drawbacks of artificial behavioral measures are that they are difficult to collect and they are vulnerable to subjects' misinterpreting incentives as intended by the researcher.

A middle ground between measures of "real world" and "artificial" behavioral outcomes comes in the form of "artifactual" measures, which involve researchers designing a task to be performed by subjects in the evaluation, but the task is designed in such a way so that the subjects are not aware that it is being used as a measurement tool. An excellent example of this strategy was given in the study discussed above by Casey et al. (2010) undertaken in Sierra Leone. In that study, both treated and control communities were given the opportunity to submit a proposal for a second round of assistance. The assistance opportunity was a real one—there would indeed be aid delivered to those communities who succeeded in submitting a proposal—but the opportunity itself was consciously constructed by the evaluation team and implementing agency to assess the capacity-building effects of the intervention. In their study, Casey et al. found that CDD programming did not increase a community's ability to succeed in submitting a proposal, leading them to conclude that whatever social gains there may have been from the program were insignificant when it came to trying to solve a real, high-stakes problem.

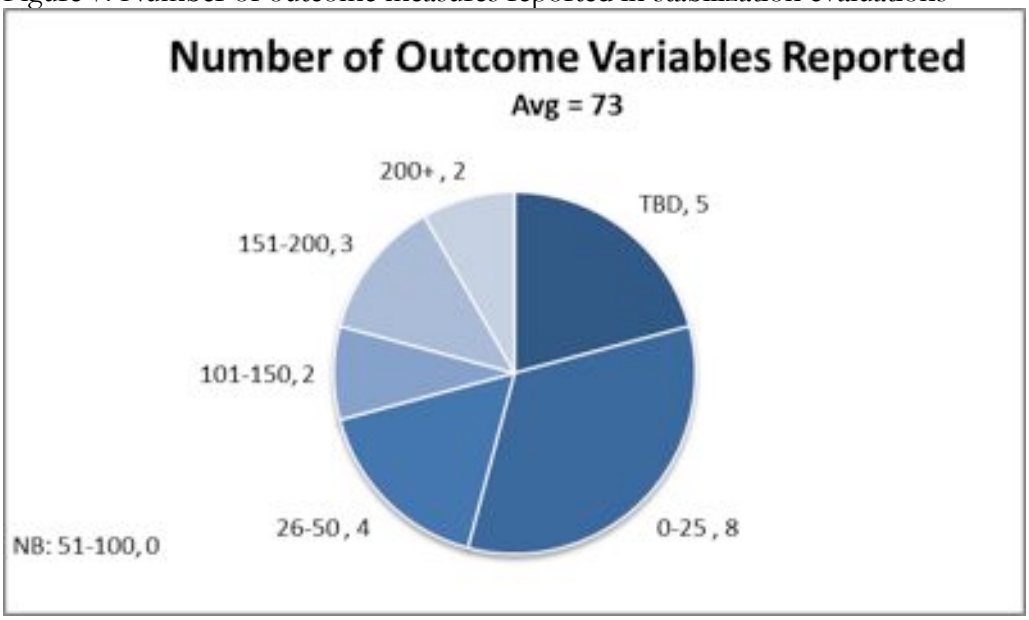
Artifactual measures such as this combine the structure of games, allowing for assessment of difficult-to-observe traits, with the unobtrusiveness and clear interpretability of real-world behavioral outcome measurement. The only drawbacks are their (typically high) cost and the considerable planning that they require. In general, it is our view that the strongest evaluations combine multiple types of measurement in order to obtain meaningful and easy-to-interpret outcome measures (e.g., using "real world" or "artifactual" behavioral outcomes) while also being able to provide insights on causal pathways that are at work (e.g., using "artificial" behavioral measures as well as survey questions getting at attitudes and perceptions). Figure 6 shows the mix of the different outcome measurement types in the stabilization evaluations. On average, these studies use at least two types of outcome measures. 20 of 24 studies use a self-reported objective measure of some sort. Nine use an artificial measure, and four use an artifactual measure.

Figure 6. Types of outcome variables used in stabilization impact evaluations



Given that there is not always a single- individual- or household-level outcome that serves as a satisfactory indicator, the potential exists for evaluators to get lost in a soup of dozens, if not hundreds, of indicators and associated impact measurements. Figure 7 below shows the distribution of the number of different outcome measures reported in the stabilization evaluations. Eleven of the 24 studies report more than 25 different outcome measures. Accordingly, some structure is required to combine a multitude of outcome measures and associated analyses into useful conclusions. In a recent review of impact evaluations of "social cohesion" interventions, King, Samii, and Snilsveit (2010) proposed two general strategies for structuring analyses that involve multiple complex multidimensional outcome measures. The first is to construct a directional map of the theory of change, tying together the intervention, potential causal paths, and end states, and then position each of the potential metrics in this directional map, as is done in a results framework or evaluation logic model. A basic example of this strategy is described above in the discussion of Fearon et al.'s (2008; 2009) evaluation of the CDD intervention in Sierra Leone. The games allowed one to assess whether the causal pathway of "improved social relations" was active when trying to make sense of the estimated impact on the end state of "material welfare."

Figure 7. Number of outcome measures reported in stabilization evaluations



The second strategy is to nest individual indicators into sets that relate to "core hypotheses", and then to perform tests in a manner that combines the information provided by each of the individual indicators. In doing so, one obtains a single test of a core hypothesis that takes all of the available evidence into account simultaneously. For example, in the study by Casey et al. (2010), the authors sought to test a hypothesis that the CDD intervention in Sierra Leone would increase beneficiary villages' social capital. In the evaluation, social capital was measured with a set of 146 indicators, including a combination of attitudinal and behavioral measures. In this case, if there were *no effects at all* but each of these measures was treated as an independent test, statistical reasoning would have us expect about 7 or 8 of these tests to show an effect that is significant at the 5% confidence level. A mistake (and one that is seen often) would be to then interpret these 7 or 8 tests as *demonstrative* of an effect on social capital. This is known in the technical literature as the "multiple comparisons" problem.[13] In order to avoid such pitfalls, Casey et al. (2010) use statistical methods that estimate effects for these 146 outcomes jointly and come up with an overall estimate for the aggregate impact on "institutions" in the abstract along with estimates of impacts on each of the 146 component measures which were adjusted to take into account dependencies across the measures. This highly technical exercise results in a very interpretable and policy-friendly way to present conclusions.

As emphasized by King, Samii, and Snilsveit, it is crucial that the measurement and aggregation strategies be written into a protocol prior to start of the evaluation—and, ideally, prior to the start of the intervention itself. There are three main reasons to do this. First, it provides an opportunity for all those involved in the intervention and evaluation to check assumptions and make revisions to either the intervention or evaluation before resources are committed. Second, it provides clear expectations against which we might judge "failure" or "success," and, in the process, learn how to design more effective interventions. Contrary to the claims of Stave (2011), establishing *a priori* expectations does not close off opportunities for learning. Rather, by clearly laying out a theory of change that links assumptions to expectations, one is well positioned to use deviations

[13] See Ioannidis, John P.A. (2005) "Why Most Published Research Findings Are False." PLoS Med 2(8).

from expected outcomes as opportunities to revisit assumptions and, ultimately, to learn. Third, by establishing analytical protocols for interpreting a large number of indicators (using the two strategies outlined above) one ensures that the evaluation will produce coherent answers to focused questions and guards against the multiple comparisons problem. For those commissioning or using the results of evaluation studies, the point is that *there are* reasonable ways to measure stabilization impacts, despite their complexity.

5. Conclusion

What we have attempted to show is the ideals of evidence-based policy-making do indeed have a place in informing stabilization programming. That being the case, the evaluation record of U.S. stabilization programs has been dismal. Through an exhaustive search of over 50 evaluations of U.S. stabilization programs over the past decade, we found only *one* example of a study that applied state of the art methods to establish a counter-factual for attributing impact. In a time of budget austerity combined with ongoing international commitments, the U.S. stabilization policy community cannot afford to fall behind in adopting state of the art evidence-based practices. Fortunately, this policy community can look to international experience with impact evaluation in stabilization programming for insights. International NGOs, other donor countries such as the United Kingdom, and multi-lateral institutions such as the World Bank have been much more progressive in applying the evidence-based policy approach to stabilization programming. It is in their work that the U.S. policy community can draw lessons, catch up, and make good on the goals outlined in the current USAID evaluation policy. Our four case studies and the subsequent discussion illustrate the diversity of approaches available for organizing productive research collaborations, establishing causal attribution, and measuring complex stabilization outcomes.

6. References

Banerjee, Abhijit V. and Esther Duflo (2011) *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty* Public Affairs: New York.

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.

Biddle, Stephen, Fotini Christia, and Alexander Their. (2010) "Defining Success in Afghansitan: What can the United States Accept?" *Foreign Affairs* (July/August), pp. 48-60.

Brück, Tilman, Patricia Justino, Philip Verwimp, and Alexandra Avdeenko. (2010) "Identifying Conflict and Violence in Micro-Level Surveys," IZA Discussion Papers 5067, Institute for the Study of Labor (IZA).

Casey, Katherine, Rachel Glennerster, and Edward Miguel. (2010) "Community-Driven Development in Sierra Leone."

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.

Chong, Alberto, Ana L. De La O, Dean Karlan, and Leonard Wantchekron. (2010) "Information Dissemination and Local Governments' Electoral Returns: Evidence from a Field Experiment in Mexico." Working Paper, Yale University.

Christia, Fontini. (2010) "Viewpoint: Measuring success in Afghanistan." BBC News online: http://news.bbc.co.uk/2/hi/8524137.stm, February 22.

Clark, Jeffrey. (2003) "Final Evaluation of the OTI Program in East Timor." USAID.

Dziedzic, Michael, Barbara Sotirin, and John Agoglia. (2008; Updated 2010). *Measuring Progress in Conflict Environments (MPICE): A Metrics Framework for Assessing Conflict Transformation and Stabilization.* Washington, D.C.

Fauth, Gloria and Bonnie Daniels. (2001) "Impact Evaluation: Youth Reintegration Training and Education for Peace (YRTEP) Program." USAID.

Fearon, James D., Macartan Humphreys, and Jeremy M. Weinstein. (2009) "Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-Conflict Liberia." *American Economic Review* 99(2): 287-291.

Fearon, James, Macartan Humphreys, and Jeremy M. Weinstein. (2008) "Community-Driven Reconstruction in Lofa County."

Gilligan, Michael, Eric Mvukiyehe, and Cyrus Samii. (2010) "Reintegrating Rebels into Civilian Life: Quasi-Experimental Evidence from Burundi." United States Institute of Peace.

Higgins, J.P. and S. Green (eds) (2008) *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*. John Wiley & Sons, Ltd: Chichester, UK.

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

Ioannidis, John P.A. (2005) "Why Most Published Research Findings Are False." PLoS Med 2(8): 124. doi:10.1371/journal.pmed.0020124.
http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124

International Initiative for Impact Evaluation. (2011) www.3ieimpact.org.

Karlan, Dean and Jacob Appel. (2011) *More Than Good Intentions: How New Economics is Helping to Solve Global Poverty* Dutton Adult.

King, Elisabeth, Cyrus Samii, and Birte Snilstveit. (2011) "Interventions to Promote Social Cohesion in Sub-Saharan Africa." *Journal of Development Effectiveness*. 2(3):336-370, 2010 (link), and *International Initiative for Impact Evaluation* Synthetic Reviews Series

King, Gary and Christopher Murray. (2001) "Rethinking Human Security." *Political Science Quarterly* 116(4):584-610.

Paluck, Elizabeth Levy. (2009a) "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. (2009b) "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda." *Journal of Personality and Social Psychology* 96(3):574-87.

Samii, Cyrus. (2010) "Evaluation." Forthcoming, *The Management Handbook: A Practical Guide for Managers in UN Field Missions*, International Peace Institute and UK Department for International Development.

Sen, Amartya. (2000) "A Decade of Human Development." *Journal of Human Development* 1(1): 17-23.

Stave, Svein Erik. (2011) "Measuring peacebuilding: challenges, tool, actions." NOREF Policy Brief, No. 2, May 2011.

Social Impact. (2006) "The USAID/OTI Community-focused Reintegration Programs in the Democratic Republic of Congo and Burundi: Final Evaluation," USAID/Office of Transition Initiatives.

United States Department of Education. (2003) "Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide" http://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

United States Department of State Bureau for International Narcotics and Law Enforcement Affairs. (2011) "Program and Budget Guide: Fiscal Year 2011 Budget".

United States Agency for International Development. (2011) "Administrator's Stabilization Guidance" http://pdf.usaid.gov/pdf_docs/PDACQ822.pdf.

United States Agency for International Development. (2010) "Standardized Program Structures and Definitions." http://www.state.gov/f/c24132.htm

United States Government Accountability Office. (2005) "Afghanistan Security: Efforts to Establish Army and Police Have Made Progress, but Future Plans Need to Be Better Defined" GAO-05-575

United States Government Accountability Office. (2009) "Program Evaluation: A Variety of Rigorous Methds Can Help Identify Effective Interventions" GAO 10-30.

United States Government Accountability Office. (2010) "AFGHANISTAN DEVELOPMENT: USAID Continues to Face Challenges in Managing and Overseeing U.S. Development Assistance Programs" GAO 10-932T.

Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest. (2000) *Unobtrusive Measures, Revised Edition.* Thousand Oaks: Sage Publications.

7. Appendix

7.1. Stabilization categories and definitions used in this paper from the "F" Structure

Foreign Assistance Standardized Program Structure and Definitions (4/8/2010)

Peace and Security

Sub-Element 1.3.2.3: Reintegration
Sub-Element 1.3.7.1: Civilian Police Reform
Sub-Element 1.3.7.3: Community Security Initiatives/Community Policing
Sub-Element 1.6.1.2: Peace Dividends
Sub-Element 1.6.2.1: Peace Structures
Sub-Element 1.6.2.2: Peace Messaging

Governing Justly and Democratically

Sub-Element 2.1.1.3: Transitional Justice
Sub-Element 2.3.1.1: Consensus-Building and Dialogue Processes
Sub-Element 2.3.1.2: Civil Society Advocacy and Oversight of Consensus Building Processes

Investing in People

Sub-Element 3.3.2.2: Victims of War

Sub-Element 1.3.2.3: Reintegration
Definition: Reintegrate belligerents back into their communities, and women and children associated with armed groups, by: supporting infrastructure, quartering and civic training for ex-combatants; providing temporary jobs; funding education and vocational training for ex-combatants and their families; providing funding for income-generation; and offering remedial schooling, trauma counseling and family reintegration. It may also include family tracing and reunification of former child combatants, both boys and girls. Such programs are normally community based, and activities conducted herein should be integrated whenever possible with their sectoral counterparts, e.g. "Investing in People/Education" or "Economic Growth/Economic Opportunity." They should also be connected to the Durable Solutions activities undertaken as part of Humanitarian Assistance.

Sub-Element 1.3.7.1: Civilian Police Reform
Definition: Develop police forces through capacity-building (training and education both in the classroom and in the field); organizational development; civil service reform (pay and rank reform); management and leadership; equipping, infrastructure, aviation, gender sensitivity, and public affairs, among other activities. As the foundation for such a service is fundamentally rooted in the rule of law and respect for human rights, activities conducted in support of this sub-element should be coordinated with programs under the Rule of Law elements in GJD.

Sub-Element 1.3.7.3: Community Security Initiatives/Community Policing

Definition: Mobilize communities and individuals (women as well as men) to prevent or reduce crime, violence or insecurity singly or through community-police coordination. Provide safety and security services – via both state and non-state providers – to outlying areas. Assist communities and individuals to work with governments and police to reassert control over ungoverned spaces, militia strongholds, and/or ganglands in urban, peri-urban and rural environments and may include the expansion of essential services.

Sub-Element 1.6.1.2: Peace Dividends
Definition: Support quick-impact, results-based activities required to demonstrate the positive impact of a peace process, operation or other event, such as mobilizing small grants for communities and local/national governments; ensuring delivery of services (e.g. "ministry in a box"); bringing local and national government authorities closer to their constituencies (e.g. town hall meetings, consultations, production and dissemination of information); generating employment for potential spoilers; and managing expectations.

Sub-Element 1.6.2.1: Peace Structures
Definition: Create substitute mechanisms in the absence of formal peace and in the midst of peace processes in order to meet the needs of people affected by conflict. Provide capacity building and training support to the parties in conflict; develop knowledge, attitudes and practices surveys; engage local communities in the peace process; and serve as a catalyst and connector between the national process and people. Support informal - most often civil society driven - multi-stakeholder confidential or public dialogues that facilitate dialogue and stimulate the exchange of ideas between the nation's political stakeholder groups including civil society and community actors in a constructive, inclusive forum.

Sub-Element 1.6.2.2: Peace Messaging
Definition: Support the media during peace processes in order to ensure balance, transparency, and accountability. Includes: supporting innovative media programs that inform and prepare people to accept the outcome of credible negotiations; creating better understanding between parties and their followers – including providing forums for dialogue; educating the public about the process and issues involved in the negotiations; and facilitating important attitude and behavioral changes towards a more just and peaceable society.

Sub-Element 2.1.1.3: Transitional Justice
Definition: Address past war crimes and human rights violations through retributive or restorative justice mechanisms, including vetting, truth and reconciliation commissions; international, local or hybrid tribunals; community-based approaches, and customary/traditional practices.

Sub-Element 2.3.1.1: Consensus-Building and Dialogue Processes
Definition: Support consensus building political processes at national, sub-national and/or local levels that incorporate views of all stakeholders including political parties and groupings, citizens, and formerly warring factions to establish a national consensus on the political structures of the state. These processes can be directly related to broader peace agreements or may occur in the narrower context of a political transition. Support citizen knowledge of and participation in consensus building forums, including marginalized groups and vulnerable populations.

Sub-Element 2.3.1.2: Civil Society

Definition: Support civil society oversight of the consensus-building processes, as well as advocacy into the process. Support citizen knowledge and civic education related to consensus building processes.

Sub-Element 3.3.2.2: Victims of War

Definition: Remove barriers to enable the full participation of victims of war in supportive communities. Help people obtain prosthetics and rehabilitation as well as training to return as functioning members of society and to be able to provide for themselves and their families after suffering injuries caused by conflict or the remnants of conflict, including landmines and other unexploded ordinance (which may be linked to the Peace and Security Objective, specifically the Explosive Remnants of War Element).

7.2. Search methodology

Databases used include:
- JSTOR, EBSCOhost, National Bureau of Economic Research, World Bank Policy Research Working Paper, World Bank Development Impact Evaluation Initiative, IdeasRepec, JPAL

Journals used include:
- International Security, American Economic Review, Review of Economics and Statistics, Econometrica, Journal of Conflict Resolution, American Political Science Review, Quarterly Journal of Economics, American Journal of Political Science, Military Review,

Keywords used include:
- Security Intervention, Stabilization and Reconstruction, Peace Operations, Disarmament, Demobilization, Reintegration, Peacebuilding, Peacekeeping, Security Sector Reform, post-peace, Post-conflict, Impact evaluation, outcomes

7.3. Bibliography of identified impact evaluations of stabilization interventions

Annan, Jeannie, and Christopher Blattman. (2010) "Why Men Don't Rebel: Experimental Results From an Ex-combatant Reintegration Program." *Typescript, Yale University and IRC.*

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, Y 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.

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

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

---. (2009) "Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-Conflict Liberia."*American Economic Review* 99, no. 2 (April): 287-291. http://pubs.aeaweb.org/doi/abs/10.1257/aer.99.2.287.

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. "The Role Of Information And Radios On Political Knowledge And Participation In Sierra Leone." *Poverty Action Lab*(2010).(*in progress*)http://www.povertyactionlab.org/evaluation/role-information-and-radios-political-knowledge-and-participation-sierra-leone.

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.

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.

Levely, Ian. (2010) "Reintegration in Post-War Liberia: A Failed Approach or Simply a Failed Program?" (*unpublished*)

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.

Paluck, Elizabeth Levy. (2009a) "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.

---. (2009b) "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda." *Journal of Personality and Social Psychology* 96, no. 3 (March): 574-87. http://www.ncbi.nlm.nih.gov/pubmed/19254104.

---. (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.

7.4. Before and after fallacy

Impact is a subtle concept and it is often misunderstood. People sometimes define impact as the "difference between beneficiaries' well-being before and after a program." This is generally incorrect. We can call it the "before-after fallacy." It is a fallacy because *many things* affect how beneficiaries' well-being changes over time. There is no reason to *attribute* such change (whether positive or negative) to the program. We need to have a comparison group. Figure 1 illustrates this. It shows a case where the "before-after fallacy" would result in an unfairly negative judgment on the program. The well-being of beneficiaries (the "treatment group") goes down over time. A naïve interpretation would be that the program caused harm. We avoid this fallacious conclusion by looking at a "control" group. What we see is that both groups experienced a decline in their well-being. However, the decline is less severe for the program beneficiaries than for the control group. Therefore, the program had a *positive* impact.

Figure 1: Illustrating the Before-After Fallacy