META-REVIEW OF INTER-RELIGIOUS PEACEBUILDING PROGRAM EVALUATIONS

Conducted by Jennie Vader
For CDA Collaborative Learning Projects
June 2015

This report was made possible by the Carnegie Corporation of New York.

Foreword

By Melanie Kawano-Chiu, Director of Learning and Evaluation, Alliance for Peacebuilding

Religious communities have powerful potential to contribute to sustainable and peaceful societies – and their contribution and inclusion to peacebuilding has never been more critical. In the past two decades there has been a plethora of academic research, by scholars and practitioners like Mohammed Abu-Nimer, Scott Appleby, Marc Gopin, Ayse Kadayifci-Orellana, and Katherine Marshall, and initiatives, like the US State Department formed the Religion and Foreign Policy Working Group and the US Institute of Peace Center for Religion and Peacebuilding, illustrate the belief that religious actors must be a part of the larger diplomatic, development and peacebuilding agendas. Despite all this interest and research however, engagement with religious communities still has not recognized its full potential as a key factor in development and peacebuilding. Part of this is based on a lack of understanding about how to effectively integrate this work into broader efforts. While we know, for instance, open-ended dialogue over the religious divide aimed at bridging various rival groups can be an effective tool for peacebuilding, we still do not know much about the optimal timing, sequencing, and method of engaging religious actors.

In 2014, the Peacebuilding Evaluation Consortium (PEC) began a partnership with the GHR Foundation to address the knowledge gap, particularly around the contributions of inter-religious action to peacebuilding through a new program: the Effective Inter-religious Action in Peacebuilding. The goals of the program are two-fold: 1) generate guidance on how to evaluate inter-religious action, and 2) develop a framework for ongoing learning regarding what constitutes effective inter-religious action. While this report is made possible by the Carnegie Corporation of New York, it is a foundational first step in understanding how the field currently measures inter-religious peacebuilding. As a baseline of sorts for our inter-religious evaluation practices, this report will also provide insights on how to improve inter-religious evaluation.

The Alliance for Peacebuilding and CDA Collaborative Learning Projects would like to acknowledge and thank the peacebuilding organizations that contributed their program evaluations for this effort - and actually practicing the principal of transparency for the sake of better peacebuilding practice: Catholic Relief Services, the Center for Interfaith Action on Global Poverty, Karuna Center for Peacebuilding, Nansen Dialogue Center, Nigerian Inter-Faith Action Association, and Search for Common Ground.

* The PEC is a long-term partnership of the Alliance for Peacebuilding (AfP), CDA Collaborative Learning Projects, Mercy Corps, Search for Common Ground (SFCG), BESA: Catalyzing Strategic Change and the US Institute of Peace. Through whole-of-community efforts, the PEC aims to: 1) enhance the methods by which evidence on peacebuilding programming is gathered and assessed, through greater methodological rigor in peacebuilding evaluation; 2) promote transparency, shared learning and collaboration among peacebuilding practitioners, policy makers and evaluators for more whole-of-field learning and evidence-generation; and 3) build better evidence-based policy and practice.

Table of Contents

Acronyms

AfP	Alliance for Peacebuilding
CDA	CDA Collaborative Learning Projects
CMM	USAID's Office of Conflict Management and Mitigation
EIAP	Effective Inter-religious Action in Peacebuilding Program
FGD	Focus Group Discussion
MSC	Most Significant Change
OECD DAC	Development Assistance Committee of the Organization for Economic Cooperation and Development
PEC	Peacebuilding Evaluation Consortium
TOR	Terms of Reference

Executive Summary

This meta-review on inter-religious peacebuilding has been commissioned by CDA Collaborative Learning Projects as part of a grant from the Carnegie Corporation of New York for the Peacebuilding Evaluation Consortium. This review seeks to understand what the current trends are in the evaluation of inter-religious peacebuilding programs and to assess the quality of evaluations.[1] It is also meant to enhance the evidence base for inter-religious action in peacebuilding by emphasizing the need for robust independent evaluations and enhanced evaluative thinking in order to increase the use of evaluation for both accountability and learning.

The meta-review included seven evaluations that assessed programs in six different countries, conducted by a total of 15 different organizations. Two of the evaluated programs utilized inter-faith action towards development aims, while the other five focused on increasing religious tolerance or decreasing conflicts between different religious communities. Four of the evaluated programs used training as a significant activity intended to achieve results. Two utilized mass media, and three involved dialogue groups.

The overall strengths and weaknesses of the evaluations in relation to both evaluation foundations and evaluation quality can be summarized as follows:

Strengths

- **The vast majority of reviewed evaluations provide evaluation purposes and evaluation questions** that were (for the most part) relevant to the evaluation purpose and/or criteria. The most common objective was to assess program effectiveness.

- **Most of the evaluations utilized mixed methods**, some of which included desk reviews, focus group discussions, key informant interviews, and surveys. While the implementation of mixed methods is an overall strength of the evaluations, the manner in which results from the various methods were presented was less than effective.

- **All evaluations list limitations to the evaluation designs**. The most common were constrained sample size and composition, lack of baseline data, and exclusive use of one data collection approach (quantitative or qualitative). While presentation of limitations is

[1] This report's focus on evaluation of inter-religious peacebuilding is also designed to contribute to the PEC's three-year initiative, the *Effective Inter-Religious Action Program* (EIAP). The EIAP was launched in November 2015 in partnership with the GHR Foundation to: 1) generate guidance on how to evaluate inter-religious action and 2) develop a framework for ongoing learning regarding what constitutes effective inter-religious action. This meta-review is part of an initial assessment of the 'state of play' in evaluation of inter-religious action for peacebuilding, including strengths, gaps and challenges of evaluation in the field.

important for the evaluation consumer to be able to assess the quality and credibility of findings, it is possible that many of the limitations could have been avoided with proper planning and also oversight by the evaluand.[2]

Weaknesses

- **Very few evaluations provided well-defined evaluation criteria for assessing the programming, and the vast majority of evaluations did not specify intended users or evaluation approaches.** Without the tools to produce specific, actionable results – such as targeted users, criteria, and an approach that guides the process – the vast majority of the reviewed evaluations lack a clear, objective judgment of the program's merit.

- **Almost no evaluations consistently include conclusions supported by strong data or evidence**. A primary issue was relying almost entirely on self-reported participant or implementing staff perspectives instead of gathering outside perspectives, collecting independently verifiable data, and triangulating data points. The use of evidence could have improved significantly if the appropriate methodologies were utilized to gather and analyze the data necessary to respond to the respective evaluation purposes and objectives.

- **None of the evaluations appeared to implement conflict- or gender-sensitive evaluation processes.** While it is possible that evaluation teams did consider gender-and conflict-sensitive measures when planning and implementing the evaluations, no evaluation report documents those, which is important for evaluation consumers to better assess the quality of information provided by the evaluation and for potentially improving the quality of future evaluations.

Recommendations

- **The evaluand and/or intended users should be specific about the evaluation foundations to ensure that the evaluator(s) produce high quality findings that serve the intended purpose.** Evaluation foundations are elements decided during the planning stages that lay the foundations for high quality evaluations. Specifying intended users, evaluation criteria, and an approach helps determine the direction of the evaluation and will assist the evaluators in designing the evaluation so that findings and recommendations are practical and useful for the intended users.

- **Evaluators, working with evaluands, should increase the robustness of evaluation design for more valid and reliable data and, therefore, more credible evidence.** This can be done through use of baseline data, gathering outside perspectives or utilizing comparison groups, and fully integrating mixed methods to increase validity and reliability of evaluation information.

[2] Evaluand means the organization or specific project being evaluated.

- **Methodologies that go beyond self-reported data and actually independently measure changes in attitude and behavior should be implemented.** If the evaluation truly seeks to learn whether a program has had effects on the population outside of the participants, the evaluation should be designed (and also appropriately resourced!) to measure those effects.

- **Include conflict- and gender-sensitive evaluation designs and processes and clearly describe them in the evaluation report.** This should be set as a field-wide expectation and both evaluation commissioners and evaluands should take responsibility for stipulating that specific conflict- and gender-sensitive processes be implemented and documented.

- **Build the evaluation capacity of relevant parties such as evaluators and implementing organizations.** Opportunities to enhance expertise include creating curricula and conducting trainings, connecting professionals in various learning fora; for organizations, deliberately utilizing internal evaluations, empowerment evaluation approaches, learning facilitators or other methods can help build internal capacity to plan for, manage, and utilize evaluations.

1. Introduction

In November 2014, the Peacebuilding Evaluation Consortium (PEC), led by the Alliance for Peacebuilding (AfP), began a new project called the Effective Inter-religious Action in Peacebuilding Program (EIAP). This meta-review was conducted to contribute to understanding the 'state of play' of evidence of effectiveness of inter-religious action in peacebuilding. It was conducted as part of a grant funded by the Carnegie Corporation of New York for PEC activities under its International Peace and Security Program. The grant promotes innovative and improved practice in peace and security by integrating cutting-edge evaluation thinking and practice, including in inter-religious peacebuilding.

This meta-review of seven evaluations on inter-religious action was commissioned by CDA Collaborative Learning Projects as part of the PEC. It aims to understand what the current trends are in the evaluation of inter-religious action and to assess the quality of evaluations. It is designed to contribute to a larger effort undertaken by the AfP to 1) improve the evaluation of inter-religious action in support of peacebuilding; 2) understand what evidence exists on what is effective in inter-religious peacebuilding; and 3) build better evidence-based policy and practice. This review, in support of EIAP, is meant to enhance the evidence base for inter-religious action by emphasizing the need for robust independent evaluations and enhanced evaluative thinking in order to increase the use of evaluation for both accountability and learning.

The evaluations assessed programs conducted by a total of 15 different organizations. Six of the organizations are international organizations, while the rest are local organizations, although at least one international organization was connected to each evaluation. The programs took place in six different countries: Nigeria, Kenya, Indonesia, Israel, Sri Lanka and Bosnia-Herzegovina. Two of the evaluated programs utilized inter-faith action towards development aims. The other five focused on increasing religious tolerance or decreasing conflicts between different religious communities. Four of the evaluated programs used training as a significant activity intended to achieve results, two utilized mass media, and three involved dialogue groups.

All of the programs were longer than one year but no longer than three years; five of the seven programs were two years in duration. The scope of the programs was not easily comparable given a lack of information regarding program budget (only one evaluation listed the program budget) and due to the varying theories of change and intervention strategies. The chart in Annex 1 gives some background information on the programs evaluated to provide a frame of reference for this review's findings.

Following the Methodology section, the report is organized into two main sections: Section 3, Evaluation Foundations, analyzes elements of the evaluation that are decided in its planning stages, most likely by the evaluation commissioner or evaluand. Section 4, Quality of Evaluations, explores the processes and results of the evaluation reports. These sections are organized according to their respective evaluation questions. The final section of Conclusions and Recommendations provides some concrete ideas about how these and future evaluations can be improved based on the analysis conducted in this meta-review. Throughout this report, "Supporting Evidence" boxes present information from similar meta-evaluations and other evaluative bodies of work in the peacebuilding field that are not directly related to inter-religious peacebuilding. While not representing evidence from a full literature review, the information given in Supporting Evidence boxes comes from similar meta-evaluations or meta-reviews; in this sense, those analyses have been used to bolster the generalizability of this review's findings.[3]

2. Methodology

2.1. Data Sources

This meta-review consisted solely of a desk review, with materials obtained through AfP. In order to obtain a diverse set of evaluations for the meta-review, AfP reached out to 14 member organizations engaged in inter-religious action and introduced them to EIAP. In addition to general support for EIAP, including recommendations for the involvement of additional organizations working in this field, AfP asked members to share their experiences with current (and/or recently concluded) inter-religious programming, including best practices, lessons learned, and challenges. They were asked to provide concrete evidence of such programming, including reports and evaluations (both internal and external). Out of the 14 organizations contacted, a total of three evaluations were collected to assist in the meta-review. One organization, a PEC principal partner, actively worked with AfP to identify evaluations that were available on their website and AfP also conducted a broader Internet search for evaluations. Other member organizations acknowledged that they had evaluations on their inter-religious projects but, for a number of reasons, were unable to share them.

All of the evaluations utilized in this report were conducted between 2011 and 2014. One evaluation was a mid-term evaluation,[4] while the others were final evaluations. The evaluations used have been listed in Annex 6 in alphabetical order, but no specific quotes or other

[3] The information provided in the Supportive Evidence box does not represent the entirety of the analysis provided in those supporting documents. More details on the resources used for the Supportive Evidence is provided in the Methodology section.

[4] The final evaluations are classified here as such because they attempted to provide an assessment of the program's value and occurred in the final months of program implementation or even slightly after the program ended. The documents themselves, however, were not necessarily titled final evaluations; terminology used by the documents' authors included "final qualitative assessment," "qualitative impact evaluation," and simply "project review report." The mid-term evaluation was labeled such by the document's author, but it is unclear at which point the evaluation fell in the life of the program as the program's duration was not indicated. It is clear, however, that the evaluation occurred approximately two years after the program started.

characteristics have been used in this report in order to avoid linking an evaluation to a particular finding of this meta-review. This has been done to mitigate any undue harm to the evaluation commissioner, evaluator(s), or implementing organizations.

The Supporting Evidence boxes draw on a number of documents, which are specifically referenced. One document in particular is featured heavily: the Evaluative Learning Review Synthesis Report on USAID/Conflict Management and Mitigation's (CMM) People-to-People Reconciliation Fund.[5] The two-year Evaluative Learning Review sought to identify lessons from the reconciliation projects funded by the Fund, while also building CMM's technical leadership in the evaluation of complex programs.[6] The review consisted of three phases: 1) Knowledge Management and Study of the Reconciliation Projects, 2) Field Evaluation of Selected Programs and 3) Reflective Learning. CMM's meta-evaluation and meta-analysis of 10 project evaluations were included in the first phase, the findings of which make up the bulk of the information in this evaluation's Supporting Evidence boxes. As the programming and evaluations from the USAID/CMM Learning Review closely mirror those featured in this meta-review, and the synthesis is a significant body of evaluative work in the conflict resolution field, it is drawn upon extensively in this meta-review.

2.2. Analysis

The questions driving the meta-review analysis, mutually agreed upon by both the commissioning organization and evaluator, were used directly as criteria upon which to compare and assess the evaluations. These have also been divided into two sections which are reflected in the organization of this report. Many of the questions were derived from the OECD Development Assistance Committee's Quality Standards for Evaluation.[7]

Comparative information (found in the section 3, Evaluation Foundations)

1. What was the evaluation purpose? (Section 3.1)
2. What evaluation criteria (if any) were used? (Section 3.2)
3. What were the evaluation questions? (Section 3.3)
4. Who were the users and who was/were the evaluator(s)? (Section 3.4)
5. What were the evaluation approaches and methodologies? (Section 3.5)
6. What were the data collection methods? (Section 3.6)

[5] Allen, S. et al. "Evaluative Learning Synthesis: USAID/CMM's People-to-People Reconciliation Fund, Annual Program Statement (APS)." Washington, DC: Social Impact and USAID, 2014. The People-to People Reconciliation Fund seeks to create a "safe space where representatives from conflicting groups can interact, prejudices and perceived differences of 'others' can be confronted, challenged, and hopefully ultimately replaced by 'mutual understanding, trust, empathy, and resilient social ties.'"

[6] Allen, S. *et al*, 1.

[7] OECD DAC, "Quality Standards for Development Evaluation." Paris: OECD, 2010. Available at http://www.oecd.org/development/evaluation/qualitystandards.pdf.

Comparative analysis of strengths/weaknesses/quality of evaluations[8] (found in section 4, Quality of Evaluations)

1. To what extent are conclusions supported by data and evidence? Are the findings specific and supported with strong quantitative and/or qualitative evidence? Have plausible alternative explanations of the evidence been explored and explained? (Section 4.1)
2. Was the information gathered valid and reliable? Are reliability problems and limitations reported, analyzed, and acknowledged? (Section 4.2)
3. Are sources of information properly identified and listed? Is there a variety of sources? (Section 4.3)
4. What are the limits and shortcomings of the evaluation approach and methodology? Are they identified within the evaluation? (Section 4.4)
5. Is the evaluation design appropriate for the questions? What are other limits that an informed observer can identify? (Section 4.5)
6. Was the evaluation process gender- and conflict-sensitive? (Section 4.6)

2.3. Limitations

The evaluations collected for the meta-review hold a potential bias since AfP's membership network is largely U.S.-based international NGOs, and as a result largely represent a western perspective of inter-religious action in the peacebuilding field. It is important to note the difficulty in obtaining comprehensive evaluations on inter-religious projects. It is likely that there is simply a lack of rigorous evaluations on inter-religious projects, as evidenced by the fact that several people from one of the leading peacebuilding networks spent multiple months actively searching for evaluations. Still, there is a possibility that the sample of evaluations presented in this report represent only a fraction of the available evaluations on inter-religious work. Regardless, the findings presented in this report are only valid for the sample of seven evaluations and cannot be generalized beyond that.

Utilizing only a desk review for the analysis of these evaluations is also a significant limitation as there was very little information about the context in which the evaluations were conducted. A more detailed picture of the state of inter-religious action and evaluations could be provided by use of a more diverse set of quantitative and qualitative methods to learn about the planning and implementation processes (what went well, what were the key challenges) and whether and/or how the evaluation findings were actually utilized.

3. Evaluation Foundations

[8] Questions 4 and 5 as stated here are an adaption from the TOR; this was done to improve the organizational flow of this report.

This section discusses key elements necessary, but not sufficient, to produce high quality evaluations: the evaluation purpose, criteria, questions, and approaches, in addition to key people involved (like the users and evaluator(s)), and data collection methodologies used. Many of these elements should be decided – or at least discussed – by the evaluand even before an evaluator is chosen. They lay the foundation for the evaluation and help guide the process so that the results are useful and of high quality. The section will be organized by this review's evaluation questions, listed in section 2.2 above.

3.1. What was the evaluation purpose?

It is important that the evaluation's audience and evaluator(s) are clear about the purpose of the evaluation so that the findings can be used effectively by the intended audience(s). The purpose will also affect how the rest of the evaluation is conducted, including what program elements are assessed and how; for example, an evaluation intended primarily for organizational needs (versus in response to donor requirements) may necessitate different types of data and analysis. Without a clearly understood purpose, the evaluation may be too broad or miss the mark, diminishing the overall usefulness of the findings.

Evaluation Purpose

The evaluation purpose is a statement about why the assessment is being conducted. Evaluations may serve the purpose of learning or accountability, although those two concepts need not be mutually exclusive.

All evaluations reviewed made an explicit statement about the purpose or aim of the evaluation. Five of the seven evaluation reports contain a concise statement with an overarching goal of the evaluation. Examples include:

> *"...to assess the effectiveness of this new institution and its centerpiece program."*
>
> *"This evaluation is aimed at assessing the degree to which the objectives and activities were met in accordance with the specific targets developed for each, providing a better understanding of the impact of the interventions."*
>
> *"...aims to capture changes in attitudes and behaviors as well as evidence that these changes are related to the project."*

The other two evaluation reports contain a list of objectives. For example:

> *"This final evaluation...focused on assessing: 1. The appropriateness and relevance of the project; 2. The validity of the three-tiered theory of change; 3. The most significant results; and 4. The sustainability of the project impact."*

All of the evaluations' purpose statements, although they mostly outline the evaluation objectives (see the next section), provide a sufficient amount of guidance about what the audience hopes to learn from the assessment. The most common objective, either stated or implied by half of the evaluation reports in this review, was to assess program effectiveness, meaning to what extent the program achieved the intended changes.

3.2. What evaluation criteria (if any) were used?

Evaluation criteria should form the main content of an evaluation report and increase the likelihood that findings can actually be used by ensuring that the intended audience and evaluator(s) share a common understanding of what information is needed.[9] Building on the OECD DAC Criteria for Evaluating Development Assistance,[10] there are now seven criteria meant specifically for evaluating conflict prevention and peacebuilding activities.[11] Five of the criteria are those used in evaluating development assistance (relevance, effectiveness, efficiency, sustainability, and impact), with the two additional criteria of coherence and coordination which are particularly pertinent to situations of conflict and fragility. These are not obligatory criteria, but they are widely accepted. The DAC guidance on *Evaluating Peacebuilding Activities in Settings of Conflict and Fragility* provides explanations of how they can be used in the context of conflict and fragility and also offers corresponding sample evaluation questions.[12]

Evaluation Criteria
Evaluation criteria, also known as evaluation objectives, are principles by which a project could be evaluated. These criteria – usually incorporated into the evaluation questions – help focus an evaluation on certain areas to assess and should contribute directly to the evaluation's purpose.

[9] OECD DAC. "Evaluating Peacebuilding Activities in Settings of Conflict and Fragility: Improving Learning for Results." Paris: OECD, 2012. p. 65. Available at http://www.oecdilibrary.org/docserver/download/4312151e.pdf?expires=1421264186&id=id&accname=guest&checksum=35DA4FB719CB604C22A67CE9B512C707.

[10] These were first laid out in the DAC Principles for Evaluation of Development Assistance. Paris: OECD, 1991. *See* http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm.

[11] *See* OECD DAC. "Evaluating Peacebuilding Activities in Settings of Conflict and Fragility", 46.

[12] For another related guidance document, *see* van Brabant, K. "Peacebuilding How? Criteria to Assess and Evaluate Peacebuilding." Geneva: Interpeace, 2010.

Three of the evaluations specifically state criteria they used, and two of those specifically reference the OECD DAC criteria. Only the two evaluations that refer to the DAC criteria provide a specific definition of exactly what the criteria mean; the other evaluation cites relevance and appropriateness, significant results, and sustainability of impact as criteria, but never provide definitions or any indication of what "relevance," "appropriateness," "significant results," or "sustainability" may look like. Although many of those are in fact criteria also recommended by the DAC, the evaluators never specifically reference the OECD DAC criteria, nor provide any further definition.

Overall, the evaluations were weak in the area of evaluation criteria, with only two evaluations providing clarity on how the programs would be assessed. With no established standards by which the evaluator sets out to assess the program, the likelihood of concrete and reliable evaluation findings diminishes. This could be a contributing factor to the overall lack of well-supported conclusions within the evaluations reviewed.

3.3. What were the evaluation questions?

Documentation of evaluation questions is considered a quality standard by the OECD DAC because it allows "readers to be able to assess whether the evaluation team has sufficiently addressed the questions, including those related to cross-cutting issues, and met the evaluation objectives."[13]

Evaluation Questions

Evaluation questions, also known as lines of inquiry, often accompany evaluation criteria and provide greater direction on what the audience wants to find out. The evaluation questions as well as the purpose and criteria can be included in the Terms of Reference (TOR) to give potential evaluators an idea of what needs to be assessed and whether their skillset is applicable to the job.

Five of the seven evaluations included specific evaluation questions. The vast majority of the evaluation questions appears relevant to the evaluation purpose and/or criteria. This area was an overall strength for the evaluations reviewed. The table in Annex 2 provides the stated purposes and questions from these five evaluations. The average number of evaluation questions asked was 9; the highest number was 15 which, at face value, seems quite high, as the amount of time and resources logically increases with the amount of questions asked. Still, for a more accurate conclusion on evaluation questions, more information is needed about the scope of the evaluation, including the budget, duration, and team size, many of which were not provided in any evaluation report.

[13] OECD DAC. "Quality Standards for Development Evaluation." para. 3.12.

Four out of the five evaluation reports with specific evaluation questions divide the questions into groups based on the criterion or area of inquiry to which the questions belonged. This is helpful to the evaluation consumer because it illustrates how the question contributed to the overall analysis and to which criteria the results of that question can be applied.

Two of the five evaluations consistently asked “why” questions, which can positively contribute to the usefulness of an evaluation. Whereas routine monitoring provides information to inform decisions while a program is being implemented, evaluation offers the opportunity for a more in-depth analysis by, for example, asking why something happened.

3.4. Who were the users and who was/were the evaluator(s)?

Establishing a clear audience is important as it helps the evaluator(s) tailor the recommendations and focus on the areas that are important for that audience’s purpose. Similar to the evaluation purpose, targeting an evaluation at an audience that is too broad can result in findings that are too general to learn from and act upon.

Evaluation Users and Evaluators

The primary or intended audience for an evaluation, also known as users, are those who will apply the findings and recommendations. Evaluators are the people who actually conduct the evaluation – they may be internal to the organization or external consultants.

Only one evaluation explicitly states the document's intended audience: *"The goal of the evaluation is to help [the organization] and partners see how the program affected and made significant changes…"*

For five of the evaluations, it can be inferred that at least one intended audience is implementing staff. Three of those evaluations made explicit statements such as: *"Key findings were summarized and shared with the [implementing organization's] team and the external facilitators"*; the remaining two made recommendations specifically to the implementing organization. Still, only one evaluation specified who in the implementing organization was best fit to use the evaluation recommendations (program staff), while the other evaluations' findings and recommendations appeared to be directed more generally to "implementing partners" or "stakeholders."

The remaining evaluation does not specify intended users, nor could an audience be readily inferred. This evaluation's stated purpose was to capture attitudinal and behavior changes, but it is not clear why or who would utilize the information. There is no indication in any evaluation who commissioned the evaluation or to whom it was ultimately submitted. A lack of clear audience is a significant weakness of the evaluations reviewed that puts the evaluations at risk of never being used for accountability or learning.

In terms of selecting evaluators, there are many things that evaluation commissioners or evaluands should consider: whether they will be internal to the organization or external; how many evaluators are needed; whether they should be hired locally or from abroad. Other important characteristics of a potential evaluator or team are their experience in evaluation (and with a specific approach) as well as in the topical area of the programming; their expertise including education and ongoing professional development; their experience in the country or region; their oral, written, and facilitation skills; their language capabilities; and the gender/nationality/ethnic composition of the team. The composition of the team in terms of gender, nationality, ethnicity, or other identity groups is important to consider as the context may be such that members of certain identity groups have easier access to information on the ground and may be perceived by program staff and participants as having less innate bias.[14]

Supporting Evidence:

The USAID/CMM Evaluative Learning Review of people-to-people reconciliation programming also found that reports referenced generic users such as the implementing organization or USAID, but none clearly indicated who should use the evaluation.

Four of the seven evaluation reports provide some information on the individuals who made up the evaluation team. One simply lists the contracted firm, and two had no indication of who conducted the assessment. Of the four evaluations that list the evaluators, two give enough information that the reader can get a sense of their qualifications. Both of these evaluations were for programs conducted by the same organization, indicating that there may be some organizational guidelines about what should be included in an evaluation; this is supported

[14] Church, C. and Rogers, M. *Designing for Results: Integrating Monitoring and Evaluation in Conflict Transformation Programs*. Washington, DC: Search for Common Ground, 2006: pp. 126, 151; OECD DAC, "Quality Standards for Development Evaluation," p. 11.

by the fact that both of those evaluations also had clearly stated purposes, included evaluation questions, and were the only two evaluations which utilized OECD DAC criteria.[15]

The positive aspects of the qualifications described in the two evaluations with detailed information are that the team included at least one lead evaluator who had experience in M&E and had team members with regional experience. One lead evaluator had experience in "*participatory evaluation of education and peacebuilding programs…teaching graduate research methods…results-oriented planning and evaluation"* and specialized in *"the design and implementation of collaborative evaluation models, in multicultural settings, intended to measure results and lead to greater stakeholder involvement."* That evaluator also had experience in the topical area of programming – conflict resolution and peacebuilding.

These were good examples of the kind of qualifications necessary in an evaluator and of the presentation of information in the evaluation report. Even though the evaluator biography only provides a small snapshot of the abilities of the evaluator(s), it is still useful to enhance the credibility of the evaluation report because the reader has a better sense of the evaluator(s) experience and expertise and how those may have contributed to the quality of the evaluation.

3.5. What were the evaluation approaches and methodologies?

Evaluation Approach
An evaluation approach is the philosophy with which the evaluation is conducted; it is not a specific method or technique, but rather a way of structuring and undertaking the analysis. Some examples include developmental evaluation, empowerment evaluation, self-evaluation, utilization-focused evaluation, and theory-based evaluation. ~ Church and Rogers, *Designing for Results*, p. 118; Treasury Board of Canada Secretariat, "Theory-Based Approaches to Evaluation: Concepts and Practices", p. 2

No evaluation specified an evaluation approach, and in five of the seven evaluation reports, no specific approach is discernible. Given the overall lack of evaluation approach, evaluation criteria, and stated intended users, the vast majority of the assessments are more like research projects than evaluations. Research and evaluation share some similar qualities, but there are

[15] Interestingly, this did not seem to have a bearing on the overall quality of those evaluations as they did not have valid nor reliable data, generally over-relied on the use of anecdotes instead of using evidence to support findings, and the evaluation design was not well-suited to the questions asked. This could reflect a gap between the evaluation Terms of Reference (TOR) and the final product, either because the evaluator had his/her own idea about how to conduct the evaluation and tried to fit the results into the TOR specifications or because the person(s) writing the TOR had greater capacity than the chosen evaluators. It could also mean that much more guidance is needed throughout the evaluation process to ensure high quality results – simply stating high expectations through evaluation purposes, criteria, questions, etc. does not necessarily lead to high quality results.

important differences; while research seeks to produce information for the purpose of building a knowledge base or advancing a theory, evaluation judges the merit or worth of a policy or program and provides information for decision-making to a specific audience. Thus, without the tools to produce specific, actionable results – such as targeted users, criteria, and an approach which guides the process – the vast majority of the reviewed evaluations lack a clear, objective judgment the merit of the programming at hand. Importantly, evaluation rather than research is necessary to learn what types of programs work in which contexts and to hold stakeholders accountable for using good practices and avoiding bad ones.

Supporting Evidence:

None of the evaluations from USAID/CMM's Evaluative Learning Review were informed by an evaluation approach. The review states, "…instead the evaluations were designed as mini-research efforts." The lack of evaluation approaches means that organizations miss out on the potential utility and/or rigor of an intentional evaluative process.

It is possible that two evaluations used, at least in part, a theory-based evaluation approach, although neither evaluation report states that a specific approach was implemented. These two evaluations are the only ones to state a program theory of change explicitly; and in one case, an evaluation objective was indeed to validate the theory of change. In theory-based approaches, each specific step in a causal chain is tested and if they can be validated by empirical evidence, then there is a basis for making a causal inference. The evaluations mentioned above do indeed list and attempt to test each step in the causal chain that illustrates the theory of change. Still, the evaluations cannot be considered to entirely embody the theory-based approach because the theories of change are not necessarily the center of the evaluation design, there is no examination of the assumptions on which the theories were based, nor does it provide enough rigor to determine whether the successes or failures of the program were due to implementation or the theory.[16]

[16] This finding is based on descriptions of theory-based evaluation approaches in Centre of Excellence for Evaluation. "Theory-Based Approaches to Evaluation: Concepts and Practices." Ottawa: Treasury Board of Canada Secretariat, 2012 and Church and Rogers, *Designing for Results*, 119.

3.6. What were the data collection methods?

Data Collection Methods
Data collection methods are the ways in which information (data) will be collected. The approach and scope of the evaluation help to determine which methods to use and it is often the case that multiple methods need to be utilized in order to address the evaluation criteria or questions.

Quantitative methods explain the who, what, when, where, how much, and how many – generally through surveys and questionnaires – and are often designed to produce statistically reliable data. Qualitative methods give more in-depth understanding of why; this can complement quantitative methods and together provide a full picture of the situation.[17] Mixed methods are commonly seen as a best practice because they allow evaluators to draw on the strengths of both quantitative and qualitative approaches and to integrate them to overcome each one's weaknesses.[18] Mixed methods are an intentional or planned use of diverse and integrated social science methodologies, or, more specifically, is a combination of quantitative and qualitative approaches to theory, data collection, and data analysis and interpretation.

Four of the seven evaluations used mixed methods; the other three used qualitative methods only. The most common methods were focus group discussions (FGD), interviews (including key informant interviews) and desk reviews. Of the four that conducted some quantitative analysis, three of those were in the form of surveys, and one was an analysis of the program data collected. In the event it is not possible or appropriate to use quantitative methods, qualitative methods alone can be used in ways that provide credible evidence; however, attention should be paid to ensure rigor so that the qualitative methods provide valid and reliable data. Some methods that may be used to increase the rigor of qualitative methods are purposive sampling to help mitigate selection bias, triangulation, multiple coding (checking of coding strategies and interpretation of data by different people), and respondent validation (checking interim findings with respondents).[19] For further discussion on mixed methods approaches, see Section 4.4.

[17] Church and Rogers *Designing for Results*, pp. 203-204.

[18] Bamberger, M. "Introduction to Mixed Methods in Impact Evaluation." *Impact Evaluation Notes* No. 3, in Impact Evaluation Guidance Note and Webinar Series. Washington, DC: Interaction, August 2012. p. 3. http://www.interaction.org/document/guidance-note-2-linking-monitoring-and-evaluation-impact-evaluation.

[19] Barbour, RS. "Checklists for improving rigor in qualitative research: a case of the tail wagging the dog?" *BMJ: British Medical Journal* 322 (7924), p. 1115.

4. Quality of Evaluations

This section examines the processes undertaken in the evaluations and the resulting analysis and findings. The elements assessed in this section are largely the responsibility of the evaluator(s) and include the use of evidence to support conclusions, data quality (reliability and validity), information on data sources, the appropriateness of the chosen methodologies, and the use of conflict and gender sensitive procedures.

4.1. To what extent are conclusions supported by data and evidence?[20] Are the findings specific and supported with strong quantitative and/or qualitative evidence? Have plausible alternative explanations of the evidence been explored and explained?

The consistent use of quantitative and qualitative data and evidence to support evaluation conclusions can significantly enhance the credibility of the evaluation and also gives more weight to programmatic successes. On the other hand, if conclusions are not based in evidence and happen to be wrong, organizations looking to implement similar programming (or the implementing organization which seeks to continue its programming) may erroneously utilize evaluation conclusions, and undertake programming that may not produce results or could be doing harm. Without data accompanying the analysis, there is no way to know whether evaluation findings are valid and reliable sources of evidence and not simply the presumptions of program and evaluation staff. Evidence is needed to provide accountability by allowing accurate assessment of whether the resources provided are beneficial, and also to contribute to the field's learning about what is going well, what is not, and how to adjust.

The vast majority of evaluation reports do not consistently include conclusions supported by strong data or evidence. While one evaluation stands out by effectively utilizing quantitative and qualitative data to demonstrate attitude shifts as a result of inter-religious trainings and joint activities, most of the evaluations could have significantly improved in their use and presentation of evidence to support conclusions.

Four of the seven evaluations somewhat use data or evidence to support their conclusions; they either utilize questionable evidence or only use evidence for a fraction of the stated conclusions. Two of these evaluations use evidence, but it was not necessarily strong or

[20] For the purposes of this analysis, conclusions were considered to be supported by data and evidence if the explicitly stated evaluation findings were accompanied by quantified results from cited data collection instruments. Note that 'quantified results' need not to have been collected by quantitative methods – even "*the majority of* interview participants said…" was considered enough to be a presentation of evidence. A list of quotations and anecdotes with no indication of the prevalence of those views, on the other hand, was not considered evidence. This definition of evidence is not scientific nor perfect, but is considered sufficient given the overall lack of actual data provided in the evaluations.

convincing. For example, one evaluation team found that the program increased understanding of the conflict for the general participants and for youth in particular. The data used to substantiate the finding was this: *"When asked to describe the conflict and what they had learned about conflict most key informants stated that they had a better understanding of the conflict (25) including the political aspects (8), the role of religious intolerance (6) and an increased understanding of the role of youth in the conflict (4)."* This statement demonstrates an attempt at utilizing survey results as evidence, but is lacking in that the methodology does not actually attempt to measure participants' understanding, and provides no indication whether the answers are representative of the general/youth participant pool.

Another evaluation's findings are mostly bolstered by a sense of how many respondents provided a given answer using a qualification such as "a large majority" or "a few voices"; for example, the report states, *"a remarkable number of participants gave examples of the kinds of personal initiatives they have already carried out and/or intend to continue…"* This and other similar statements are followed with corroborating participant quotations; however, the findings would have been much more powerful if the actual numbers or percentages of respondents were given. It is important to note that the numbers may not have been given in this case due to the small sample size used, although that was not acknowledged by the evaluators.

Three out of seven evaluations regularly make strong statements about program successes with no supporting evidence. For example, one evaluation provides a quote from a program implementer: *"I think the whole…program is sufficient and in accordance with the aim of peace…The [program] has introduced tolerance…Formerly, a teenager would be confused if s/he was asked about tolerance. Currently, s/he will understand because of [the programming], or because it is often discussed in this place."*[21] The evaluation team goes on to conclude: "*The above statement indicates that the programs are sufficient to promote tolerance and prevent religious extremism. It is also evidence that the notion of religious freedom has been successfully cultivated in the minds of the [participants], particularly the youth, without realizing the process or even feeling indoctrinated."* While the implementer's quote is illustrative and potentially true, it does not constitute evidence 1) that the programming was sufficient to promote tolerance and prevent religious extremism; 2) that the notion of religious freedom was cultivated in participants; or 3) that participants did not realize they were learning and did not feel indoctrinated. This particular evaluation design included key informant interviews, focus group discussions, and a questionnaire distributed to targeted program participants, so it can reasonably be assumed that information from those three methodologies – which included a sufficient variety of sources – could have been used to provide actual evidence supporting the evaluation team's conclusion. In fact, perhaps the evaluators even based their conclusion on the data

[21] Specifics of the programming were omitted from both quotations by the author in order to preserve anonymity of the examples being used.

collected, but if that was the case, they did not present the relevant data in a manner which strongly supported their finding.

Many of the issues regarding a lack of robust evidence to support conclusions likely stem from an overall lack of methodological rigor[22] (though in some cases it may reflect the evaluators' failure to present the supporting evidence for the conclusions in the report); this will be discussed further in this report. Most of the evaluations could have improved significantly if the appropriate methodologies were utilized to gather and analyze the information necessary to respond to the respective evaluation purpose and objectives. Recurring weaknesses in the reviewed evaluations included a lack of valid and reliable data (demonstrated by a lack of non-participant perspectives and other data points necessary for triangulation), selection and response biases (recognized or unrecognized), and analytical limitations, particularly in fully leveraging mixed methods or sufficiently rigorous qualitative methods. These, in turn, may have diminished the amount and certainly decreased the quality of robust evidence used for interpretation. A thorough understanding and deliberate consideration of which methods to employ (i.e. increased rigor) could have helped avoid many of the specific issues mentioned and therefore, would have enabled the evaluator(s) to support conclusions with more robust evidence.

Just over half (4/7) of the evaluations provide plausible alternative explanations for their findings, usually in the form of multiple possible interpretations of the available data. One illustrative example is: *"It is interesting to note that religious leaders cited ethnic or religious tensions as a risk more than twice as often at the end of the program as they had at the beginning…Since there was no follow-up question, we cannot draw any decisive conclusions from their responses. It may well reflect concerns over the new inter-faith/ethnic tensions that have surfaced…in recent months. Given the overwhelming positive feedback we received from our participants, it is also likely that the data reflects a heightened state of awareness of the tensions that still simmer in the aftermath of years of war and the obstacles these pose to reconciliation and sustainable peace."* This explanation, and others from different evaluations, demonstrate both that the evaluators recognize the limitation of certain methodologies and are analyzing the data within the given context. This information could certainly be useful for organizations to consider when designing or implementing programs aimed at decreasing ethnic or religious tensions and also when planning an evaluation.

[22] Rigor in this sense means that the evaluation applies the appropriate tools to answer the evaluation questions and respond to the evaluation purpose; this includes whether data collection tools produce valid and reliable data and that the data collection tools and analytic techniques "maximize the chance of identifying the full range" of relationships, patterns, and interpretations thereof. This definition was influenced by: Ryan, G. "What Are Standards of Rigor for Qualitative Research?" Paper presented at the National Science Foundation's Workshop on Interdisciplinary Standards for Qualitative Research, p. 2005. Available at www.wjh.harvard.edu/nsfqual/Ryan%20Paper.pdf.

While many of the evaluations left room for improvement with regard to effectively demonstrating the use of strong evidence to support conclusions, it is nonetheless encouraging that five out of seven evaluations did use some evidence and that four out of seven evaluations included alternative explanations.

Validity and Reliability

Validity means that information serves the intended purposes and supports well-founded interpretations. Reliability means that the information produces sufficiently dependable and consistent information. Evaluation information is reliable when repeated observations using similar instruments under similar conditions produce similar results. (OECD DAC Glossary of Key Terms)

4.2. Was the information gathered valid and reliable? Are reliability problems and limitations reported, analyzed, and acknowledged?

The validity and reliability of data is important to increase the credibility of the evaluation and, therefore, its conclusions. As Church and Rogers state in *Designing for Results*, “Since the conclusions of an evaluation are what inform program decision-making, the consequences of using flawed instruments can have significant negative effects on the project and the people it is meant to assist.”[23] Using already verified instruments and testing new instruments increases the likelihood that the data produced will suffer from less bias, will be more reliable, and will be less likely to lead to false conclusions. Some examples of highly reliable methods include document and secondary data review and also surveys and questionnaires. Qualitative methods such as focus group discussions may produce less reliable data due to potential biases in selection, how the facilitator directs the conversation, and the veracity of the participants’ responses.[24] Still, triangulation – using different methods to collect and analyze data on the same issue – can improve data quality, particularly if the evaluator(s) is uncertain whether a data source is able or willing to provide the full story.[25] In addition, documenting the evaluation methodologies is an important way to increase reliability as it could inform future evaluators and researchers into how the methods were designed and implemented.

While the lack of published data makes it difficult to definitively determine validity and reliability of the evaluations’ information, **four of the seven evaluations do not present valid data.** Instead they rely on the opinions and beliefs of program staff and participants, make broad

[23] Church and Rogers, *Designing for Results*, p. 215.
[24] *Id*, pp. 214-215.
[25] *Id*, p. 213.

statements without supporting evidence, or simply state conclusions that do not answer the evaluation questions.

For example, this evaluation finding is stated without any evidence or indication as to how the information was generated: *"The interfaith approach is effective because [interfaith council[26]] members are united around a common problem - getting girls to school and avoiding early marriage. As they began to address these issues, knowledge about gender and the complex social problems associated with early marriage and pregnancy probably evolved."* The validity of this statement is questionable, as no verifiable data is presented 1) to corroborate that the approach is "effective" (nor is there any indication of what effectiveness may look like in that context), 2) to confirm that members are indeed united by the conviction that girls should avoid early marriage, and 3) to support the hypothesis that their knowledge evolved throughout the programming. With the proper data collection methods, each of these three components could have been explored and, if found to be true, could have been substantiated with data, thus improving the validity of the statement. It is possible that the evaluator(s) did collect such data and that their analysis supports their finding; however, that is not articulated in the report. It is important to present the data along with the evaluation findings so that the user can assess whether the conclusion is truly supported by evidence or rather represents an opinion or biased assumptions.

Most of the evaluations suffer from poorly-formed conclusions like the one highlighted above. However, two of the evaluations stand out from the rest by providing, for the most part, concrete findings with seemingly valid and reliable evidence. For example, one evaluation stated: *"In conversations with representatives of the attacked religious communities we discovered that they are actually very satisfied with police work, and praised them for doing their best with such limited resources. This could be corroborated by evidence we received during a field visit to previously attacked . . . churches. This visit coincided with a religious holiday. In both cases, there were police patrols in vicinity of the Churches."* The validity of this finding is stronger than that of the previous example as it draws on information from a variety of sources, including from an eyewitness account by the evaluators themselves.

Three out of seven evaluations present reliable information that would likely be replicable. These evaluations indicate specific methodologies used, how individuals were chosen to participate, exactly the questions asked, etc. and are the same three evaluations that provided sufficiently valid information. Again, since almost all of the evaluations provide insufficient information about methods used or data collected, it is challenging to determine with certainty the reliability of the information. Yet, an overreliance on anecdotes and quotations from program staff and participants, with no other data points presented, makes it appear that the information –

[26] Full name omitted by the author for the sake of anonymity.

at least in the way it was presented – is unreliable, as it could likely vary if the evaluation was to be conducted again.

Only two of the seven evaluations recognize potential biases in the reliability of the information provided. One evaluation that demonstrates reliable and valid information as described above also identifies what was termed social desirability bias, *"wherein research participants may know and want to provide what they perceive to be the “right” response to the researchers.”* Those evaluators wrote, *“The [participants] may, therefore, tend to over-emphasize positive feelings and perceptions, especially if they wish to see project activities continue."* This evaluation team made sure no one on the team was an implementer and took precautions to ensure confidentiality so that participants could answer freely. Another evaluation cites selection bias that could have impaired the information’s reliability. This will be discussed at greater length in Section 4.5.

Overall, the vast majority of the reviewed evaluations could have improved in terms of validity and reliability by ensuring that the data collection methodologies they used would provide information directly responding to the respective evaluation question or objective, by substantiating each finding with the data used to derive that finding, and by providing detailed information on how the data was collected and analyzed.

Data Sources

Data sources are where the information will be accessed. In many cases, these are program participants or implementing staff; other sources include media program transcripts, newspaper articles, police records, among others.

4.3. Are sources of information properly identified and listed? Is there a variety of sources?

Data sources should be varied so as to gather the most comprehensive assessment of a program, who it did or did not affect, and how. Furthermore, these sources and their potential biases should be identified and listed so that the evaluation consumer can better understand where the evaluation information originated and if there is any reason to believe that the information is skewed toward any one perspective. Listing identifiable characteristics of data sources, however, should be treated with caution in particular contexts if being identified as a source will cause harm to that respondent.

All of the evaluations identify data sources, although many of them provide minimal information such as how many sources were utilized, and what position participants had in society. Two of the evaluation reports do an excellent job of listing and identifying sources, with important information like sex disaggregation, while also preserving the anonymity of those sources. For the most part, the evaluations offered a sufficient variety of sources, often utilizing secondary data in the form of a desk review while also gathering the perspectives of both program staff, participants, and other relevant stakeholders like leaders from religious institutions or civil society organizations. In some cases, increased variety of data sources, both in terms of utilizing non-participant perspectives and independently objective data points, could have improved the evaluation design. This is discussed in further detail in Section 4.5.

4.4. What are the limits and shortcomings of the evaluation approach and methodology? Are they identified within the evaluation?

Every evaluation or piece of research has limitations in method or implementation. Documenting those limitations and attempts to mitigate their effects demonstrates that the evaluation team adjusted their analysis and interpretation to fit the context. This allows the consumer to assess the accuracy and credibility of the findings better.

All of the evaluations describe some limits of the evaluation methodology. Recognizing and documenting at least some limitations is a strength of the reviewed evaluations.

Five evaluation reports cite a limited sample size and composition as the evaluation's main shortcoming. One example is: *"The limited sample size means that the results are not representative of the [entire] population and the results at the community level cannot be generalized. Instead, the results can only serve as a snapshot of the views of individuals living near and around the targeted [communities]."*[27] Another evaluation, although it claimed to have a representative sample, still examined specific limitations to the generalizability of its findings: "*Few of the following analyses proved statistically significant, so we cannot extrapolate these results to the larger participant list...*" The analysis goes on to state, however, that *"...when combined with the more extensive qualitative analysis found later on in this report, the quantitative data still stands to contribute much to the 'big picture.'"* This is a good example of complementarity in mixed methods, as discussed on page 27.

Some limitations regarding sample size and composition are closely related to the variety of data sources. One evaluation recognizes that the methodology was *"limited to a small subset and selected by the implementing organization based on convenience and availability; therefore, we*

[27] Contextual references omitted for the sake of anonymity.

cannot assume that they are representative of the broader experience of project participants." Unfortunately, the authors of that report did not point out that choosing participants based on convenience and availability may introduce significant bias into the results. In fact, most of the evaluations reported that the program staff helped choose the participants of interviews and FGDs; this opens space for biased answers, which, if unavoidable, should at least be cited as a limitation.

Significantly, no evaluation utilized the perspectives of people outside of the program, such as non-participants, to inform evaluation objectives. For example, one program sought to "promote a culture of tolerance and activism" in an entire community through organizing dialogue groups among university students from opposing sides of a conflict. In order to explore whether the program achieved change among the group members and then among the wider community, the evaluators held focus group discussions with the participants and key informant interviews with a couple of university administrators. A quote from one of the dialogue participants stated, *"[The program] influenced all the close people to those who participated in the project. Whoever took part also took it out of the group, forwarded the message to the outside world. When I learned new things about our reality and our conflict that surprised me I shared it with everyone I knew,"* and a facilitator stated, *"Through the participation, it reached the community, especially the families heard about the program."* While this is important information, getting the perspectives of other university students or the families themselves would have been a stronger method to obtain the true effects on non-participants.

Supporting Evidence:

Most of the evaluations from USAID/CMM's Evaluative Learning Review also used a "post-test project group only" design which, as the report states, "is a limited design that cannot ascertain the degree of change because data is gathered only once and only from people involved with the project."

Two evaluations did acknowledge this limitation; one stated, "*Using a control group of only non-listeners could have strengthened the survey, but the evaluation period did not allow time for this step,"* and another wrote, *"Interviewing some respondents who were more peripheral to the project could have strengthened the evaluation, but, again, limited time did not allow for this reach."* While these two evaluations demonstrate that the evaluators were aware of how to strengthen the evaluation but that insufficient resources prevented a more robust design, the fact that the vast majority of the evaluations do not even recognize this limitation may indicate that many evaluators are not aware of or do not deem important the value of gaining outside perspectives.

Three evaluations lament the fact that there were no baseline data with which to compare endline data; two of these three evaluations are those that also identify the limitations of not having a control group, as discussed in the section immediately above. One evaluation report states, "*...without baseline data and lack of a comparison group of non-participants against which to compare the responses of the project participants, it is impossible to attribute perceived positive changes to the influence of the project activities alone."* Only one evaluation team attempted to compensate for this type of limitation: "*Respondents were asked to recall information they knew about malaria from before the program as an attempt to compensate for lack of baseline."*

Three evaluations highlight the limitations of only using one method instead of mixed methods. Interestingly, taken together these evaluations present the importance of mixed methods well as one used quantitative methods only and the other two used exclusively qualitative methods. One report states, *"It is apparent that a simple analysis of quantitative data cannot establish a meaningful pattern for correlating a proactive approach by condemning the attacks to making no action, in reducing the total number of attacks. Deeper contextual analysis of each incident may give light to this correlation, but it is not in the scope of the project."* On the other hand, another evaluation states, *"Qualitative research is useful for identifying dominant trends and themes based on the perspectives of individuals. However, conclusions and findings are non-quantifiable and they are not representative of the broader population."* The third evaluation reports, "*…the qualitative nature of this evaluation limits the potential for generalization of findings and recommendations to other contexts."*

However, it should be noted that these last two statements are an inaccurate analysis of the advantages and disadvantages of qualitative versus quantitative data. Quantitative data is not necessarily representative of the broader population and qualitative information can in fact be quantified through the use of coding and can also produce representative findings with the proper sampling. Mixed methods enhance the likelihood of reliable information, but qualitative methods, applied with the appropriate amount of rigor, are still valid tools for evaluating conflict resolution and peacebuilding programs; however, misperceptions about the validity of qualitative methods like the one described here may result in evaluation consumers not taking findings based on qualitative data seriously.

Despite the potential for mixed methods to enhance an evaluation's assessment of inter-religious action programs, **the vast majority of evaluations did not leverage mixed methods effectively.** While slightly over half of the reviewed evaluations' designs incorporated both quantitative and qualitative methods, the reports do not present the results in a way that enhances the benefits of a mixed methods approach. According to Bamberger, the specific benefits of mixed method designs are 1) triangulation of evaluation findings, 2) development of instruments, 3) complementarity which extends the comprehensiveness of evaluation findings, 4) generation of new insights, and 5) incorporating a wider diversity of values.[28] The reviewed evaluations could have capitalized on the benefits of mixed methods by intentionally utilizing them to compare information obtained from different methods (triangulation) and by presenting the results in a way that deepened the analysis to provide richer findings (complementarity). The evaluation reports tend to present the quantitative findings in a completely different section than the qualitative findings, which diminishes their potential to support key findings. This makes the evaluation findings difficult to synthesize as the information provided to answer a certain evaluation question or support a conclusion is presented in multiple locations instead of in one coherent line of reasoning. This may also imply that the quantitative and qualitative data was analyzed separately and that the evaluation team did not wholly integrate all of the available information.

Three evaluations present challenges related to how their methodology was implemented; examples include a lack of clear instructions to surveyors, long surveys which made it difficult for surveyors to reach target totals, a tightly-packed data collection schedule and delayed receipt of translations which limited the opportunity to adjust the methodology as needed, and missing data due to incomplete survey forms.

Presenting these limitations is a positive feature of the respective evaluation reports, but many of them represent challenges that could have potentially been prevented with a more careful consideration of the methods and processes used. For example, thorough surveyor or enumerator training, the provision of written instructions, piloting surveys and other instruments, and routine data quality checks should be required steps of the evaluation team and should also be monitored by the evaluand. All of these activities and careful planning about how much time each activity will take should be carefully considered when planning the evaluation and both the evaluator(s) and evaluand should agree on the implementation and timeline.

Other limitations in methodology and approach that are mentioned once each include a steep learning curve for external consultants to understand the nuances of how the project works and also travel restrictions due to conflicts which limited data collection. It is likely that these challenges cannot be entirely avoided, but utilizing local consultants or organization staff that are not directly involved in the programming being evaluated can facilitate better and faster understanding of the context, and different methodologies may be developed for collecting data even if not in person. Again, careful planning and communication when designing the evaluation may help mitigate many challenges.

[28] Bamberger, *Introduction to Mixed Methods*, pp. 4-5.

4.5 Is the evaluation design appropriate for the questions? What are other limits (besides those identified in the evaluation report) that an informed observer can identify?

Most of the evaluation designs could have been improved to more adequately respond to the evaluation questions (or purpose). As was stated in Section 3.1, most of the evaluations sought to measure "effectiveness" or "results" or "impact," but their ability to truly conduct an informed analysis of those concepts was significantly hampered by the evaluation design.

The vast majority of the reviewed evaluation methodologies did not allow the evaluators to measure objectively whether participants had experienced change from when the respective programs began to when they ended, or whether they would have experienced change had the programming not occurred. These evaluations primarily relied on self-reported data from program participants or implementing staff. Even more unfortunate is that at least half of the evaluator(s) did not state these as limitations to their methodology. For example, one evaluation sought evidence that attitudes changed and, if they had, whether the change was linked to the program. The evaluators used both quantitative and qualitative methods at baseline and endline that convincingly demonstrated that attitudes had in fact changed. However, the only "evidence" in the report that the change occurred due to the programming came from actually asking the participants whether the programming effected the change. This evaluation was one of the strongest in terms of providing valid and reliable data, utilizing mixed methodology, and drawing from baseline and endline data; yet, the methods in this case were not rigorous enough to demonstrate even contribution.

The least biased method of determining whether a program catalyzed change are impact evaluations where attribution can be determined through use of counterfactual analysis. In some cases, the evaluation commissioner and intended users do not seek to prove attribution, or the methods commonly used to establish attribution may not appropriate for the context. Popular ways of measuring contribution rather than attribution include theory-based, case-based and participatory approaches, including contribution analysis, case studies, process tracing, Most Significant Change (MSC) studies, outcome mapping, and others.[29] Perhaps the most important lesson is that the evaluand and evaluator(s) planning an evaluation should carefully consider the purpose, objectives, and evaluation questions and their utility to the intended users. Is attribution necessary? Does the organization want to understand its contribution to participants' behavior change? What about behavior change in indirect beneficiaries? Upon understanding the

[29] *See* Stern, E. *et al. Broadening the Range of Designs and Methods for Impact Evaluations*. London: DFID, 2012; White, H. & Phillips, D. "Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework." International Initiative for Impact Evaluation Working Paper 3. New Delhi: 3ie, 2012.

necessary level of analysis, methodologies can be chosen that will be able to address the purpose of the evaluation.

Only two evaluations state potential biases in how data was collected. Through only a document review, it was not possible to determine all of the possible biases evident in the evaluation methodologies or how much of an effect they may have had on the findings of individual evaluations. However, selection bias and response bias are some common biases that should have at least been mentioned in nearly all of the evaluations, and, as also noted in Sections 4.2 and 4.4, failure to address these potential problems can significantly affect the validity and reliability of the evaluation's conclusions.

Given that six of the seven reviewed evaluations employed methods that were not designed to generalize beyond the sample selected, there should have been a discussion about the potential for selection bias and measures taken to mitigate its effects on the findings. *Selection bias* arises when evaluation participants are systematically different from non-participants; in the reviewed evaluations, the factor that posed the most risk of introducing selection bias was that the implementing organizations often chose which program participants or stakeholders to consult in focus group discussions and interviews. Only one evaluation report describes the potential for selection bias: "*despite efforts made to apply the sampling criteria when organizing the FGDs, ultimately the participant mix was determined by who showed up. This meant that respondents were skewed towards those with stronger connections to the program and organization."* The evaluators of that program did not take actions to mitigate the bias, but felt that there was a sufficient amount of perspectives to generate an accurate analysis.

Random sampling is one method to avoid selection bias, although that may not be ideal or feasible in a conflict context. Another sampling method is purposive sampling where the evaluator recognizes potential important differences in the population and deliberately chooses people to get a variety of perspectives. In the evaluation described immediately above, purposive sampling was used in an attempt to mitigate selection bias, using critical case and best case-worse case sampling methods, but the participants who chose to respond were still very few, and therefore, selection bias could have skewed findings. It is likely that selection bias may not be entirely avoidable, but the potential for systematic differences should be recognized, documented, and factored into the analysis and interpretation of data.

Response bias is another common problem that arises from self-reported data as well as leading or confusing/hard to answer questions. One example of self-reported data that is subject to bias emerged from interviews conducted by one evaluation team; the conclusion states, *"Both Palestinian and Israeli students indicated significant increases in their willingness to engage with the 'other'."* These answers may be biased by the fact that participants knew that willingness to engage with the 'other' was the aim of the program and that the evaluation sought to assess that aim; they may also be biased as the participants likely wanted to believe that they experienced these changes. An alternative way to accurately measure whether such an increase indeed occurred would have been to objectively measure the before and after engagement with the 'other.'

Another evaluation's conclusions could be subject to response bias due to the way that survey questions were asked. A series of questions about religious disputes included: *"To your knowledge, have there been any religiously driven incidents of violence in your village in the past 12 months?"; "During the past year, how often were religious disputes resolved peacefully in your community?"; "Were women involved in resolving any of the religious disputes?"* These questions may produce unreliable information as responses may differ widely depending on how well informed the participant is about incidents of violence and how they are resolved; it could well be that incidents occurred of which the participant was not aware or about which the participant had incorrect information. Not only should this information have been collected from – or at least triangulated with – more objective sources, but the potential bias involved should have been documented in the evaluation findings so that the evaluation users could make informed decisions about the extent to which answers can be considered factual and widespread.

At least **two of the evaluations could have further strengthened their analysis by utilizing more and more varied data sources** (aside from non-participants as mentioned in Section 4.4). For example, one evaluation sought to explore what outputs were produced by the program and if they were of appropriate quality. However, instead of independently evaluating the tangible outputs listed, such as manuals and discussion programs, the evaluation simply asked participants if they found the outputs to be of high quality. This is problematic because participants may not have any frame of reference for what a quality manual looks like, they may bias their answers if they believe that positive responses will continue programming, etc. While asking participants their opinions about the quality of outputs may satisfy one standard of quality, it should be triangulated with other objective standards for a comprehensive assessment and robust finding. That same evaluation sought to find out whether the program activities prevented further escalation of the conflict, again, primarily by asking the opinions of participants. However, other data points such as reported incidents of violence in the region during the implementation time might have been used for triangulation, adding validity to the findings. Another evaluation's design mostly used a desk review of program documents and routine monitoring data which, in the case of this program's activities, was appropriate for assessing the processes and expected results of the program, but was not enough to assess the overall goal which was "enhanced trust and improved relations between religious and ethnic communities." More qualitative information from a greater variety of sources would need to be collected in order to determine that magnitude of change.

Another evaluation sought to assess the program's relevance by asking program participants and staff if they felt the program was relevant; although that report does not define relevance, the OECD DAC definition is: "the extent to which the objectives and activities of the intervention(s) respond to the needs of beneficiaries and the peacebuilding process – i.e., whether they address the key driving factors of conflict revealed through a conflict analysis. Relevance links the outcomes of the conflict analysis with the intervention's objectives, although the relevance of the intervention might change over time as circumstances change."[30] The evaluation, then, only asked a small group of people what they deemed relevant and consulted no external sources to determine whether and to what extent the intervention actually responded to the needs of the peacebuilding process.

An additional point of interest for the PEC and peacebuilding evaluation enthusiasts may be whether the evaluation designs were unique in any way as evaluations of *inter-religious* peacebuilding programs. **Approximately half of the evaluations included questions or data sources specifically related to the inter-religious aspect of the peacebuilding program.** Four of the evaluations featured questions focused specifically on the inter-religious aspect of the program. Examples include, *"Did encouraging interfaith training help Muslim and Christian faith leaders establish greater trust and understanding of each other? Is there evidence that this contributed to any positive transformation of Muslim-Christian relations at the local level? What are the strengths/weaknesses of the interfaith element of the program?"* and *"How has the program contributed to thinking and the dialogue process between community leaders?"* Other evaluation questions focused on the perceptions of one's own religious identity and of the 'other's' religious identity and whether programming fostered greater interaction with participants of other faiths.

Besides the evaluation questions, three evaluations used religious leaders as main data sources, although those leaders were also the direct beneficiaries of the programming. Unfortunately, the evaluations that did ask questions about changes in attitude or behavior through inter-religious programming still suffer from many of the problems discussed in this report; data points were mostly self-reported through interviews and in two cases a survey, and little effort was put into triangulating that data with other methods. Future areas of interest in this area would be to explore theories of change in inter-religious programming, such as: what types of change results from inter-religious dialogue versus joint activities such as the people-to-people initiatives assessed in the Evaluative Learning Review of USAID/CMM's People-to-People Reconciliation Fund (and cited in this report)? What are the differences between inter-religious action for the sake of conflict resolution versus a separate development outcome? These and others are explored more in the Conclusion and Recommendations section.

[30] "Evaluating Peacebuilding Activities in Settings of Conflict and Fragility," p. 65.

Overall, the evaluation methodologies used were not entirely inappropriate and did yield some interesting and potentially useful findings; however, there were ways that each evaluation could have been strengthened by using more rigorous methods, recognizing and correcting for biases, and also by expanding the variety of data sources utilized.

4.6 Was the evaluation process gender- and conflict-sensitive? Was conflict analysis appropriately incorporated into the evaluation?

Gender and Conflict Sensitivity

Gender and Conflict Sensitivity was defined rather broadly; to be either gender or conflict sensitive, an evaluation had to include an explanation of gender-/conflict-sensitive measures taken, gathered data from appropriate variety of audiences, ensured anonymity/confidentiality to the maximum extent possible, incorporate data collection methods that did not ask potentially harmful questions or at least had procedures in place to mitigate harm.

Evaluations – just like interventions – should consider how different social dynamics can affect the implementation process and participants. Thus, with the expectation that all evaluations, simply by being conducted, may affect or be affected by both gender and conflict issues in a community, measures should be taken to understand these dynamics and diminish, to the extent possible, negative effects on both the evaluation and the community. Furthermore, these should be documented so that the evaluation consumer can assess to what extent gender and/or conflict factors may have affected the findings and whether or not actions were taken to improve the findings' validity. Documenting best practices in gender- and conflict-sensitive evaluation approaches can also further the field's learning and improve future evaluation methodologies.

No evaluation reports explain measures taken specifically for the purpose of being either gender-or conflict-sensitive. Only one evaluation discusses keeping the appropriate level of anonymity and confidentiality: *"Identifying characteristics are not assigned to university administrators and project team members in order to maintain confidentiality given the small number of respondents in these categories."* Although one might infer that this was done due to safety or other potential negative consequences of being associated with the program or the evaluation, this reason was not framed in a way that indicated anonymity was kept primarily for gender- or conflict-sensitive purposes.

In terms of gender-sensitivity, all evaluations gathered information from a variety of sources in terms of demographics, including sex, and two evaluations included the instruments they used (e.g., FGD and survey questions) in the report, and these, at face value, did not ask any obviously offensive or harmful questions. This is the extent of gender-sensitive measures which could be inferred from the reviewed evaluation reports.

Despite the lack of explicit measures taken to be gender-sensitive, four out of seven evaluations provide findings specifically regarding the treatment of women and the perceptions of gender relations within the programming. The two concepts should not be conflated; making statements about how gender relations were affected throughout the programming is not the same as a gender-sensitive evaluation process. However, if an evaluation directly seeks information about gender dynamics or the treatment of certain identity groups (like participating women), the evaluation designers should certainly consider the processes by which they will obtain the necessary information. The OECD DAC Guidance on Evaluating Conflict Prevention and Peacebuilding Activities states:

> *Those planning an evaluation will need to determine how it will cover gender issues. Field experience and extensive research show that women and men and boys and girls*
> *experience, engage in, and are affected by violent conflict in different ways. . . . A clear, critical understanding of gender equality within a particular conflict context is, therefore, important for policy makers and practitioners, as well as for evaluators.*
> . . .
> *Encouraging participation of both women and men and knowing the informal rules of communication between men and women, is central to selecting a gender-sensitive approach. The incorporation of both women and men in the sample or study population should be ensured and potential obstacles to women's participation in the evaluation addressed. For instance, it could be difficult for evaluators to speak directly with women and women may not express themselves freely in the presence of men. The methodological implications of these gender dynamics should be considered.*[31]

Since over half of the evaluations directly referenced gender dynamics, the fact that no evaluations documented a gender-sensitive process is a significant weakness.

Similarly, no evaluation report specifies any conflict-sensitive processes. Since all of the reviewed evaluations assessed programs that either sought to directly affect an existing conflict or sought to otherwise engage parties who were in conflict, all of the evaluators should have considered conflict-sensitive measures while planning and implementing the evaluation. In a conflict situation, the methodology should be carefully considered, both by ensuring the anonymity of sources (not requiring their names or documentation and being mindful of where, when, and how the evaluator(s) contact sources) to avoid harm to them, and in using methods and instruments that are sensitive to the sources' experiences, including past trauma. Again, the OECD DAC Network on Conflict, Peace and Development Co-operation provides guidance on conflict sensitivity in evaluation, stating, "As a policy or program should be conflict sensitive, so should the evaluation process itself. Evaluations carried out before, during, or after a violent conflict must be conflict sensitive because they are themselves interventions that may impact on the conflict. In this respect, it is important to understand that questions asked as part of an

[31] OECD DAC, *Evaluating Peacebuilding Activities in Settings of Conflict and Fragility*, pp. 47, 50.

evaluation may shape people's perception of a conflict. . . . The evaluation report must explain what measures were or were not taken to ensure the conflict sensitivity of the evaluation itself and any impact that taking or not taking them may have had on the results of the evaluation."[32] While it is possible that evaluation teams did consider gender-and conflict-sensitive measures when planning and implementing the evaluations, no evaluation report documents those, which, as mentioned above, is important for consumers to be able to better assess the quality of information provided by the evaluation and for potentially improving the quality of future evaluations.

Supporting Evidence:

All of the USAID/CMM's Evaluative Learning Review evaluations also lacked adequate conflict considerations and gender sensitivity. The report states, "Insufficient conflict considerations and gender sensitivity mean that the evaluations risk doing harm and not taking into account all of the relevant variables and, therefore, not fully understanding the interaction of the program and context."

Finally, none of the evaluations mention assessing the conflict analysis or sensitivity of the program. With the limited sample size of this review, it is impossible to tell whether that is a norm for the field. It seems logical that an evaluation of programming occurring in an area of conflict, whether or not the programming is not directed at the conflict, should assess that the program did no harm to participants or the existing conflict situation, particularly in light of the fact that, as the OECD DAC Guidance on evaluating peacebuilding asserts, "all activities in a fragile and conflict-affected setting must be conflict sensitive."[33] At a minimum, the field of evaluation for peacebuilding and conflict resolution programming should consider further discussions on reasonable expectations of assessing programs' conflict sensitivity.

5 Conclusion and Recommendations

This meta-review demonstrates that of the seven evaluations on inter-religious action that were assessed, there are some common strong foundations which include explicitly stating the evaluation purpose, utilizing evaluation questions, and implementing mixed methods.

[32] *Id.*, pp. 35-36.

[33] OECD DAC, *Evaluating Peacebuilding Activities in Settings of Conflict and Fragility*, p. 35. *See also* CDA Collaborative Learning Projects. "Reflecting on Peace Practice Project," 2004, pp. 18-21.

However, there is still room for improvement in providing well-defined evaluation criteria, clearly specifying the evaluation's intended users, and using evaluation approaches. If these three elements are better developed, the assessments could move from being more like research projects to evaluative processes in which program strengths and weaknesses are measured against criteria and findings can be utilized by a specific audience. The evaluations also lacked strong evidence to support their conclusions and generally suffered from a dearth of valid and reliable information. A more robust design which includes baseline and comparison group data (or at least perspectives from non-participants) can increase the ability of evaluators to accurately assess results and enhance the credibility of the final products. Finally, no evaluation referred to conflict- or gender-sensitive processes, which not only limits the depth of analysis but has the potential to cause harm to the people involved in the evaluation.

As it set out to do, this meta-review, with support from other key reports such as the USAID/CMM Evaluative Learning Review, provides a better understanding of what evidence exists and demonstrates the need for more robust independent evaluations and the need for enhanced evaluative thinking to increase the use of evaluations for accountability and learning. The analysis above and the following recommendations can be utilized by evaluation commissioners, evaluands, and evaluators to improve the evaluation of inter-religious action in support of peacebuilding.

- **The evaluand and/or intended users should be specific about the evaluation foundations to ensure that the evaluation produces high quality findings that serve the intended purpose.** While the purpose and evaluation questions in the sample analyzed in this study were relatively strong, evaluations could benefit from identifying more specific intended users and by improving the use of evaluation criteria. Specifying intended users is a good exercise for the evaluand as it will help determine the direction of the evaluation and will assist the evaluators in designing the evaluation so that findings and recommendations are practical and useful for the intended users. Evaluation criteria should be carefully chosen by the evaluand and should be well defined in the evaluation Terms of Reference. Moreover, additional management and oversight of the evaluation process can ensure that the intended criteria and questions actually guide the evaluation and produce specific, trustworthy findings.

- **Evaluators, working with evaluands should increase the robustness of evaluation design for more valid and reliable data and, therefore, more credible evidence**. This can be done by using baseline data, gathering outside perspectives or utilizing comparison groups, and fully integrating mixed methods to increase validity and reliability of evaluation information. Using baseline information and a comparison group will provide the highest likelihood of producing reliable data. Recognizing that full experimental designs are resource-intensive and not always appropriate, feasible or necessary, there are other ways to measure the change a program achieved, such as contribution analysis, theory-based evaluation methods, case studies, Most Significant Change (MSC) studies, outcome mapping, and others. This increased rigor will help improve the field's evidence base and credibility.

In terms of mixed methods, evaluators and evaluation consumers should be willing to accept the use of quantitative methods in evaluating conflict resolution and peacebuilding programs, but should also implement qualitative methods in a way that is rigorous and can stand up to scrutiny. Some methods that may be used to increase the rigor of qualitative methods are purposive sampling to help mitigate selection bias, triangulation, multiple coding (checking of coding strategies and interpretation of data by different people), and respondent validation (checking interim findings with respondents). Moreover, when mixed method approaches are implemented, the results should be used in a way that complements each approach's strengths and utilizes triangulation to improve data quality. However, in order to increase design robustness, evaluation commissioners and/or implementing organizations need to make sure there are sufficient resources to incorporate these elements into methods. Also, the evaluand could enhance the management of the evaluation to ensure that evaluators are aware of the expectations regarding rigor and that these elements are implemented with sufficient quality.

- **Evaluators should implement methodologies that go beyond self-reported data and actually independently measure changes in attitude and behavior.** If the evaluation truly seeks to learn whether a program has had effects on the population outside of the participants, the evaluation should be designed (and also appropriately resourced!) to measure those effects instead of speculating on them. As was mentioned often in this analysis, including non-participant perspectives through interviews or focus group discussions is one important way to get a broader sense of changes that have occurred. Another method which may be used could be media monitoring; for example, watching for increased numbers articles, radio programs, websites that focus on elements of peacebuilding and conflict resolution or, alternatively, decreased numbers of those media inciting violence or intolerance.[34]

[34] Garfinkel, Renee. "What Works? Evaluating Interfaith Dialogue Programs." United States Institute of Peace Special Report, July 2004. Washington, D.C.: USIP, p.10.

Supporting Evidence:

"There are some obvious and simple measures of success, such as the number of participants attending, or the number willing to return or who refer others to the program. There also quantitative measures of attitude change, which rely on self-report to questionnaires. Both of these are important. But what really makes a difference is what people do following the program that they did not do before. Behaviors of various sorts can be observed and quantified once they have been identified as target behaviors."

- **All stakeholders should work to ensure that conflict- and gender-sensitive evaluation designs and processes are implemented and clearly described in the evaluation report.** This should be set as a field-wide expectation and both evaluation commissioners and evaluands should take responsibility for stipulating that specific conflict- and gender-sensitive processes be implemented and documented. These expectations should be written into the evaluation Terms of Reference and should be carefully managed by the evaluand.

- **Commissioners, evaluands, and evaluators can all work to build the evaluation capacity of relevant parties such as evaluators and implementing organizations.** Building the capacity of evaluators perhaps obviously entails many activities that are currently happening such as creating curricula and conducting trainings, connecting professionals in various learning fora, and continually improving the evidence base for inter-religious programming and evaluation.[35] Increasing the evaluation capacity of implementing organizations can also improve the quality of evaluations.

  An Empowerment Evaluation approach is one way to improve the evaluation capacity of implementing organizations. In that approach, program participants and staff "jointly examine issues of concern, while an external evaluator performs the role of a coach or extra facilitator…"[36] This approach is designed to improve programs using a form of self-evaluation and reflection that provides the capacity-building benefits of an internal evaluation with the oversight of an external evaluator. This approach could be coupled with what Church and Rogers call a learning facilitator role, which, when compared to an operative or consultant role, means that the evaluator has a broader mandate than simply

[35] For information on some of these activities, see the Alliance for Peacebuilding website, particularly opportunities and resources under the title "Continuing the Evolution of the Field: The Peacebuilding Evaluation Consortium." Available at: www.allianceforpeacebuilding.org/our-work/about-our-work/peacebuilding-evaluation/.

[36] Church and Rogers, *Designing for Results*, p. 115.

conducting an evaluation and submitting a report; an evaluator who is a learning facilitator conducts the evaluation, contributes to the utilization of the findings, and also facilitates organizational learning. This role may, for example, facilitate workshops with staff to develop an implementation plan based on the evaluation, develop lessons or questions that are applicable beyond the project, establish an ongoing learning system for the project team, and assist with new program development.[37] The intensiveness of this role and the amount of time an evaluator tasked as a learning facilitator ultimately spends with the project team can help build the evaluative thinking of the staff and can help ensure that the results of the evaluation are actually used.

This meta-review was limited to analyzing the strengths and weaknesses of the evaluation methods and products of the seven evaluations of inter-religious peacebuilding programming provided by AfP, and not to synthesize the learnings about inter-religious action that emerge from them. With the currently available information, it would be difficult to conduct an effective meta-analysis; this is due both to the variety of program concepts, strategies and contexts represented in this small sample, and to questions about the trustworthiness of the findings stemming from the overall weaknesses in the validity and reliability of the evidence provided in the majority of the reports. Nonetheless, from the reviewed evaluations and insights from other evaluative pieces of work like the USAID/CMM Evaluative Learning Review, some interesting topics to explore in inter-religious action may be:

- **Type of change:** what are different programs' effects on individual attitude and behavior change versus sociopolitical change? Also, how does the program duration impact the amount of change at different levels (individual, group, sociopolitical)?

- **The focal point of inter-religious programming:** how do attitudes or behaviors toward the 'other' change when opposing sides of a conflict are brought together with the sole purpose of conflict resolution (e.g. through inter-religious dialogue) versus with a specific objective and task at hand (such as documenting vandalism of religious sites)? How do attitudes or behaviors toward the 'other' change if the specific objective of inter-religious action is a development outcome such as reducing malaria or decreasing child marriages?

- **Theory of change:** how do programs with an explicit theory of change differ from similar programs without one? How do evaluations of programs with explicit theories of change differ from those without one?

[37] Church and Rogers, *Designing for Results*, 113.

Annex 1: Overview of Evaluated Programs

Program Concept[38]	Activities	Duration	Program Scope
Religious leaders delivered messages to educate and motivate congregants to eradicate malaria through use of bed nets	Train leaders who, in turn, deliver messages to congregants	Unknown. At least one year	10 leaders directly trained; 20,000 faith leaders involved; 3 states
Inter-religious dialogue and action used to decrease child marriage by addressing many issues of concern to girls, boys, and communities, such as low levels of education, child neglect, sex tourism, and poverty	Training, capacity-building and community outreach including primary school clubs, savings and loans clubs for community members, vocational training and support for rescued girls	2 years	40 savings and loans groups 8 schools with total 800 club members; 1 Vocational Training Institute where the program placed 10 girls
Promotion of religious freedom and prevention of radicalization through youth-centered media and educational activities	Establishing community radio stations and producing programs; establish video competitions and documentary productions	2 years	intended to reach 25,000 students from 10 communities
Interfaith dialogue and teambuilding among youth served as a foundation for social cohesion and help promote a culture of tolerance and activism as well as combat/delegitimize racism	Dialogue groups and a cumulative activism initiative (plus internships for some youth)	Almost 3 years	6 universities
Build the capacity of civil society to prevent and resolve inter-religious conflicts	Leadership trainings, collaborative dialogues, community interventions, radio programming	2 years	300 people trained, 4 one-day dialogues with total of 40 participants, 30 community interventions, 14 roundtables plus 52 radio dramas plus 120 PSAs aired
Decrease the number of attacks on religious sites and improve inter-religious and interethnic relations by involving religious leaders and the media to document attacks and actively speak out against them	Monitoring attacks and other incidents on objects of religious significance and facilitating responses by religious communities, authorities, media	2 years	
By changing the attitudes of key religious leaders through training and joint activities, their broader communities will develop more tolerant, positive attitudes toward each other and conflict and ethnic tensions will significantly decrease.	Conflict transformation skills development and in-depth dialogue for religious leaders, implementation of 53 community projects by leaders in inter-faith teams. development of inter-faith councils, advocacy at national level	Approximately 18 months	Approximately 80 key religious leaders (Hindu, Buddhist, Christian, and Muslim) from 3 regions; additional 80 young religious leaders; 5,000 community members through inter-faith projects.

[38] These summaries of program concepts were inferred by the author from information included in the evaluation reports. They may or may not accurately depict the program designers' or implementers' understanding of the program concept. The summaries are provided here to give the reader context and a frame of reference for the type of programming evaluated.

Annex 2: Evaluation Questions

This table depicts the evaluation purposes and questions from the five evaluations which provided questions. The evaluations are not in any particular order and project-specific information has been omitted for the sake of anonymity.

Stated Purpose: Focus on assessing 1) the appropriateness and relevance of the project, 2) the validity of the three-tiered theory of change, 3) the most significant results, 4) the sustainability of the project impact
Evaluation Objective 1: Evaluate the appropriateness and relevance of the project: - How relevant and appropriate was [the program] to the current context and targeted age group? - What was the experience of gaining participation of women, minorities, and marginalized groups in this program? **Evaluation Objective 2: Test the validity of the project's three-tiered Theory of Change.** - What has changed in the way participants perceive: A. themselves; B. others of their own group; and C. those of the other identity? Why/how? - Which are the most useful skills gained by participants? How have they used them to date? - Have group participants increased their willingness to participate in relationship-building with the 'other'? If so, why and how? - What is the nature of the inter- and intra-group dynamics that were created as a result of the project? (What were they based on? What have they resulted in?) - Were there any turning points in the relations? - What challenges do group participants still face in building relationships/being able to work cooperatively with others (intra- and inter-group)? - To what extent and how, if at all, have the project's joint outreach/advocacy activities engaged local authorities (municipal/university/other)? What has been the result? - How have the [effects of the] project's joint outreach/advocacy activities been perceived by local authorities/community/other students? - To what extent did [the project] influence the wider community? - What reactions do participants get from their family or peers as a result of their participation in the project? How do they respond to these reactions? **Evaluation Objective 3: Learn which have been the most significant results of the project, and understand how they were achieved.** - Which aspects of the project have had the most significant impact—intended or unintended—for participants? Which factors have contributed to this impact? **Evaluation Objective 4: Explore the sustainability of project impacts.** - What effects—positive or negative, intended or unintended—does the project leave behind, among individuals, informal groups, community institutions, and partnering organizations? Which of these, if any, are likely to endure in three to five years? - What are group participants' future plans? Does the [project] experience relate to those plans? If so, how?

Stated Purpose:

…to capture changes in attitudes and behaviors as well as evidence that these changes are related to the project

- Is there evidence that attitudes are changing and in the desired direction? If so, is the change linked to the program? And if so, how/why? Is there evidence that behaviors are changing and are participants doing more with participants in the program from other faiths? Outside the program? If so, is there evidence that the change is linked to the program? And if so, how/why?

- Is there evidence that the broader society is being exposed to more tolerant attitudes? Is there evidence of broader societal attitudinal changes? Is there evidence that youth are becoming involved in reconciliation or conflict prevention activities? Is there evidence that this engagement is changing their attitudes or behaviors?

- Is there evidence that the program is affecting overall levels of conflict/violence in any way?

Stated Purpose:

This evaluation is aimed at assessing the degree to which the objectives and activities were met in accordance with the specific targets developed for each, providing a better understanding of the impact of the interventions.

Effectiveness:

- To what extent were [implementers] able to adapt to changing context and conflict environment?
- What outputs were produced and were they of the appropriate quality?
- What was [the INGO’s] value added in the partnership with local NGOs and institutions?
- What were the key factors that influenced the achievement or non-achievement of the objective and outcomes?
- How has the program contributed to thinking and the dialogue process between community leaders?
- Were the intended outcomes achieved? Specifically, to what extent is the program contributing to a change in attitude, skills, and behavior of the targeted population?

Impact:

- How has the program contributed/not contributed to how citizens envision conflict and peace?
- Have the program activities prevented further escalation of the conflict?
- What changes have taken place on the secondary beneficiaries (motorcycle taxi drivers, youth, and women) as a result of the program?

Stated Purpose:

to assess the effectiveness of this new institution and its centerpiece program

Malaria Program Impact and Activities

- "Are faith leaders having a significant impact on net-utilization and net-hanging rates?"
- "How effective is the training-of-trainers model? How can it be improved?"
- "How can coordination between partners be improved to ensure that interfaith action is maximally effective?"
- How is the program working at the congregational level? Does the training provide religious leaders with a deep enough understanding of key issues/messages? Are they delivering the messages?"

- "Are the training and data-collection tools appropriate and effective?"

Christian-Muslim Relations

- Did encouraging in interfaith training help Muslim and Christin faith leaders establish greater trust and understanding of each other?
- Is there evidence that this contributed to any positive transformation of Muslim-Christian relations at the local level?
- What are the strengths/weaknesses of the interfaith element of the program?
- How can the opportunities for Muslim-Christian collaboration be strengthened throughout the course of the state-level work?

Stated Purpose:

To see how the program affected and made significant changes on [participants] and surrounding communities in reference to their religious perspectives and understandings

Relevance:

1.a. To what extent was the project's approach relevant in promoting religious freedom and countering radicalization through youth-centered media and educational activities? Was the set of activities sufficient? To what extent did the different categories of activities complement each other?

1.b. Did the project target the right group of beneficiaries?

1.c. What positive or negative unexpected or expected results did the project lead to on the youth and on the surrounding host-communities?

Effectiveness:

2.a. Did students' and teachers' knowledge and skills on community radio operation, radio programming and peace and tolerance issues improve, and to what extent did they use the learned skills to promote religious tolerance and resolve conflict in their community?

2. b. To what extent did the project achieve its intended results? What major factors contributed to achieving, or not achieving, its objectives (factors of success and challenges)?

2. c. To what extent did the project empower the students and teachers? To what extent did the skills learned through the training and awareness activities promote religious tolerance and counter radicalization in the surrounding communities?

2.d. Did the project foster cooperation between the kiai, students, teachers, and the surrounding community to work together to promote religious tolerance and counter radicalization?

2.e. To what extent did the project change the attitude of the host-community and increased religious tolerance?

Sustainability:

3.a. Which steps are planned or have been taken to create long-term processes, structures and institutions for the successful promotion of peace and tolerance and prevention of radicalization and extremism in participating areas and their surrounding communities?

3.b. Have the participating communities developed independent initiatives on promoting religious tolerance and countering radicalization?

Annex 3: About the Author

Jennie Vader is a development professional, specializing in the Analysis, Design, Monitoring, Evaluation, and Learning (ADMEL) of development programs. She received her BA in International Political Economy from Colorado College and her MA in NGO Management from the Fletcher School of Law and Diplomacy at Tufts University. Jennie has completed numerous assessments like the one presented here, utilizing desk reviews and other methodologies to analyze evaluation quality and organizations' capacity and efficacy in Monitoring and Evaluation. While her primary focus has been in development, she has an extensive academic background in peacebuilding and conflict resolution, including skills in conflict and gender analysis and designing, monitoring, and evaluating peacebuilding programs.

Annex 4: Meta-Evaluation Terms of Reference

Purpose: This meta-review is part of a larger effort undertaken by the Alliance for Peacebuilding to 1) improve the evaluation of inter-religious action in support of peacebuilding; and 2) understand what evidence exists on what is effective in inter-religious peacebuilding; and 3) build better evidence-based policy and practice. This review aims to assess the "state of play" in evaluation of inter-religious action. This review aims to begin to understand what the current trends are in evaluation of inter-religious action and assess quality of evaluations, with a view to identifying areas of strengths and area for further development and improvement of evaluation of inter-religious action.

Scope of meta-review: The consultant will analyze 10-15 evaluations of inter-faith dialogue programming/processes over the last 5 years, focused on programs funded by the GHR Foundation, and, where those are inadequate, programs implemented by GHR partners and AfP members not funded by GHR. GHR will supply evaluations. Any additional evaluations needed will be provided by CDA Collaborative Learning Projects. The evaluations to be focused on will be identified in consultation with the GHR Foundation and AfP, with consideration to identifying evaluations of programs by organizations with resources/expertise in evaluation and comparable program types. The consultant will participate in initial conversations with AfP and GHR to determine the scope and focus of the meta-review, as well as the standard of evidence, and quality criteria to be used.

Meta-review elements:

The meta-review will analyze patterns and trends in terms of evaluation approaches and methodologies, and to identify strengths and weaknesses of the evaluation methodologies and products. The meta-review will be conducted based on a review of the documentation alone, with consultation, when appropriate with the GHR Foundation, but will not require interviews with evaluators or program staff.

Questions include:

Comparative information about evaluations:

- Who were the users and who was/were the evaluator(s)?
- What was the evaluation purpose (e.g., learning, accountability, etc.)?
- What were the evaluation questions?
- What evaluation criteria (if any) were used (e.g., OECD DAC criteria like relevance, effectiveness, impact, sustainability, etc.)?
- What were evaluation approaches and methodologies?
- What were data collection methods?

If these elements are not addressed in the evaluation, the report will note their absence, or identify any implied information, and discuss the implications.

Comparative analysis of strengths/weaknesses/quality of evaluations:

- Was the information gathered valid and reliable? Are reliability problems and limitations reported, analyzed and acknowledged?
- Are sources of information properly identified and listed? Is there a variety of sources? Are potential biases identified, acknowledged or analyzed?
- What are the limits and shortcomings of the evaluation approach and methodology? Is the evaluation design appropriate for the questions? Are they identified within the evaluation? What are other limits that an informed observer can identify)?
- To what extent are conclusions supported by data and evidence (according to standards to be decided in consultation with AfP and GHR)? Are the findings specific and supported with strong quantitative and/or qualitative evidence? And have plausible alternative explanations of the evidence been explored and explained?
- Was the evaluation process gender- and conflict-sensitive?
- Was conflict analysis appropriately incorporated into the evaluation?

General recommendations

- Analysis of any patterns and conclusions regarding evaluation quality, challenges, gaps and questions for further investigation or development in further meta-evaluation and inquiry
- Possibilities for meta-analysis: to what degree can any learnings and observations of findings/conclusions be seen as trustworthy?

Annex 5: References

Evaluations Reviewed

Countering & Preventing Radicalization in Indonesian Pesantren: Final Evaluation Report. Search for Common Ground. Indonesia, July 2014. Conducted by Lanny Octavia and Esti Wahuyuni.

Dialogue and Action Project: Final Assessment Report. Catholic Relief Services, Coast Interfaith Council of Clerks, the Catholic Diocese of Malindi and the GHR Foundation. Kenya, June 2013. Conducted by International Center for Research on Women.

Gemini Final Evaluation Report. Catholic Relief Services and Sadaka Reut. Israel, July 2014. Conducted by Nell Bolton and Carol Daniel Kasbari.

Inter-Religious Cooperation for Community Development and Social Empowerment in Trincomalee and Batticaloa Districts and Padaviya Division: Final Evaluation Report. Karuna Center for Peacebuilding. Sri Lanka, May 2013.

Monitoring and Responses to Attacks on Religious Buildings and Other Holy Sites in BiH. Nansen Dialogue Center. Sarajevo, March 2013.

Preventing Inter-Religious Violence in Plateau State Nigeria: Final Evaluation. Search for Common Ground. Nigeria, January 2013.

The Faith Effect: Nigeria. Center for Interfaith Action on Global Poverty and Nigerian Inter Faith Action Association. Nigeria, 2011. Conducted by Wise Solutions, LLC.

Other References and Documents Consulted

Allen, S. et al. 2014. *Evaluative Learning Synthesis: USAID/CMM's People-to-People Reconciliation Fund, Annual Program Statement (APS).* Washington, D.C.: Social Impact and USAID.

Bamberger, M. 2012. "Introduction to Mixed Methods in Impact Evaluation." Impact Evaluation Note No. 3. *Impact Evaluation Guidance Note and Webinar Series*. Washington, D.C.: Interaction. http://www.interaction.org/impact-evaluation-notes (accessed March 16, 2015).

Barbour RS. 2001. "Checklists for improving rigor in qualitative research: a case of the tail wagging the dog?" *BMJ: British Medical Journal* 322 (7924): 1115-1117.

Blum, A. and Kawano-Chiu, M. 2012. "'Proof of Concept' – Learning from Nine Examples of Peacebuilding Evaluation. A Report on the 2011 Peacebuilding Evaluation Evidence Summit." Washington, D.C.: United States Institute of Peace and Alliance for Peacebuilding.

CDA Collaborative Learning Projects. 2004. "Reflecting on Peace Practice Project." http://www.conflictsensitivity.org/sites/default/files/Reflecting_on_Peace_Practice.pdf

Centre of Excellence for Evaluation. 2012. "Theory-Based Approaches to Evaluation: Concepts and Practices." Ottawa: Treasury Board of Canada Secretariat.

Church, C. and Rogers, M. 2006. *Designing for Results: Integrating Monitoring and Evaluation in Conflict Transformation Programs*. Washington, D.C.: Search for Common Ground.

Garfinkel, Renee. 2004. "What Works? Evaluating Interfaith Dialogue Programs." *United States Institute of Peace Special Report.* Washington, D.C.: USAIP.

OECD DAC. 2012. *Evaluating Peacebuilding Activities in Settings of Conflict and Fragility Improving Learning for Results*. *Organisation for Economic Co-operation and Development.* Available at http://www.oecd-ilibrary.org/development/evaluating-donor-engagement-in-situations-of-conflict-and-fragility_9789264106802-en (accessed March 16, 2015).

— 2010. "Quality Standards for Development Evaluation." DAC Guidelines and Reference Series. Paris: Organisation for Economic Co-operation and Development.

— 2010. "Glossary of Key Terms in Evaluation and Results Based Management." Paris: Organisation for Economic Co-operation and Development, DAC Working Party on Aid Evaluation.

— No date. Criteria for Evaluating Development Assistance, www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm (accessed March 15, 2015).

— 1991. "Principles for Evaluation of Development Assistance." Paris: Organisation for Economic Co-operation and Development.

Rogers, M., Bamat, T, Ideh, J. 2008. *Pursuing Just Peace: An Overview and Case Studies for Faith-Based Peacebuilders*. Baltimore: Catholic Relief Services.

Ryan, G. 2005. "What Are Standards of Rigor for Qualitative Research?" Paper submitted at the National Science Foundation's Workshop on Interdisciplinary Standards for Qualitative Research, May 19-20, 2005.Available at www.wjh.harvard.edu/nsfqual/Ryan%20Paper.pdf (accessed March 16, 2015).

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. & Befani, B. 2012. *Broadening the Range of Designs and Methods for Impact Evaluations: Report of a study commissioned by the Department for International Development*. DFID Working Paper 38. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/67427/design-method-impact-eval.pdf.

USAID Office of Conflict Management and Mitigation. 2009. "Religion, Conflict, and Peacebuilding: An Introductory Programming Guide." Washington, D.C.: USAID.

van Brabant, K. 2010. "Peacebuilding How? Criteria to Assess and Evaluate Peacebuilding." Geneva: Interpeace.

White, H. and Phillips, D. 2012. *Addressing attribution of cause and effect in small n impact evaluations: toward an integrated framework.* 3ie Working Paper 15. New Delhi: International Initiative for Impact Evaluation. http://www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf.