INTERNATIONAL AID EVALUATION: AN ANALYSIS AND POLICY PROPOSALS

Thomaz Kauark Chianca, Ph.D.

Western Michigan University, 2008

Evaluation has been intertwined with international aid work since its inception in the late 40's-early 50's, but it is still an area with considerable room for improvement. If, as is often alleged, evaluations of international development efforts are methodologically weak they are misleading international agencies about the real impact of the sizable amount of resources being spent. A recent study by Chianca, described in this thesis, with a sample of 50 US-based international non-profit organizations (INGOs) illustrates the serious situation of the structure and practice of evaluation in those agencies.

A number of efforts to improve this situation have been put in place. Some of them have greater focus on methodological solutions and push for the development of more rigorous impact evaluations using experimental or quasi-experimental designs. Other efforts, while maintaining perspective on the importance of adopting more rigorous evaluation methods, have instead prioritized the establishment of principles and standards to guide and improve evaluation practice. Studies involving thorough analysis of the main efforts to improve international aid evaluation and of the most prominent evaluation standards proposed to the development field are scarce.

This dissertation is a contribution to the field in several ways: (i) it provides a general synthesis of the current movements to improve aid evaluation; (ii) it describes and assesses some of the most prominent standards for aid evaluation; (iii) in particular, it presents a thorough assessment of the most widely adopted set of

evaluation criteria worldwide, the five OECD/DAC evaluation criteria, with specific suggestions for improving them; (iv) it discusses results of a survey of INGOs on their evaluation principles and practice, and their feedback on the evaluation standards recently proposed by InterAction (the largest coalition of US-based INGOs); and (v) in the light of the preceding, it provides InterAction and other aid agencies with concrete suggestions to improve future revisions of their evaluation standards and guidelines.

INTERNATIONAL AID EVALUATION: AN ANALYSIS AND POLICY PROPOSALS

by

Thomaz Kauark Chianca

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation

Western Michigan University
Kalamazoo, Michigan
April 2008

ACKNOWLEDGMENTS

Finishing this dissertation represents the completion of a long and exciting journey I have made with my family since 2003. I left Brazil and a carrier as an evaluation consultant because I wanted to learn more and become a top-notch evaluator with international experience. If I am able to say after almost 4 years and one-half that I have successfully accomplished my objective, it is only because I have received the support from many people. Without them I doubt I would ever get this far.

My wife, Claudia P. Ceccon, and my three daughters, Carolina, Gabriela and Mariana Ceccon Chianca are the first I would like to thank. They all had to give up so many things, especially family and friends in Brazil, to follow me in this adventure. I want to thank them for their patience, especially at this final stage, and for their unconditional love regardless of my many imperfections. I would never be able to do this without their support.

My parents, Rosemary (in-memoriam) and Eclécio Chianca, are the ones to be primarily blamed for my inquietude to pursue new learning adventures. Their life example of leaving their hometowns and families to pursue further education and overcome poverty has certainly left a mark on me. My in-laws, Jo and Claudius Ceccon, for all their unconditional support during all these years, especially for taking in Carolina after she decided to move back and start Med school in Rio. Maria Belaniza and Gustavo Botelho, my sister and brother in-law, also deserve special thanks to provide a shelter and lots of love for Gabriela who also moved back to Brazil before I was able to finish my studies.

I was only able to come to WMU because of some interesting conjunction of factors. Thanks to Dr. Jim Sanders, I was first introduced to the evaluation world in 1995. He was the one who invited me to apply to the program. Dr. Jane Davidson was who accepted my application, facilitating the arrangements for me to come and serving as my first advisor. Dr. Teri Behrens was the person offering me a job at the Evaluation Unit at the W.K. Kellogg Foundation that made it possible for me to have a decent life with my family for the initial two-years. My special thanks to Mrs. Mary Ramlow who was instrumental in helping me stay on track in my program and navigate through all the university bureaucracy during all the years I stayed at WMU.

Given the tailor-made characteristic of the program, putting together and following a PhD study plan is not an easy task. For helping me go through this process and for serving on my written and oral comprehensive exams committee, I would like to thank Drs. Jane Davidson, Michael Scriven (committee chair), Daniel Stufflebeam, David Hartman, and Paul Clements.

I was lucky enough to put together a great dissertation committee. Dr. Michael Scriven (chair), Dr. Paul Clements, and Mr. Jim Rugh comprise the best group of scholars and practitioners I could ever dream to have on my side and help me explore my rather complex dissertation topic. Their guidance, flexibility and patience were fundamental!

I am a strong believer that one will best learn by putting into practice what she learns. For my great happiness, I was able to work in different projects with some of the best evaluators in the world while I was pursuing my studies. Of course those “extra activities” slowed a bit my degree completion schedule, but undoubtedly added

quality to my training. Those were unforgettable and exciting experiences that took me to diverse places in the world; from rural Michigan to the highlands of Tibet, the dry forest in Peru, the Northern mountains in Thailand, and several other wonderful places. For making this possible I want to thank Dr. Michael Scriven for his tireless and ingenious work leading the Heifer International impact evaluations 1, 2 & 3 (and possibly 4 and 5!); Dr. Dan Stufflebeam for supervising my work in 2003/2004 at the W.K. Kellogg Foundation (WKKF); Dr. Paul Clements for working with me in "Vida no Vale Project" in Brazil and in writing the Uttar Pradesh Sodic Lands paper; Dr. Robert Brinkerhoff for orienting the Success Case Method evaluation of the WKKF Learning Partners initiative; and Dr. Teri Behrens for supervising my working at the WKKF Evaluation Unit and for co-authoring a few still unpublished papers.

The many friends we made in Michigan were also essential in making our experience much more enjoyable and easier—many thanks to Allison Downey & John Austin (for all the great moments and especially for helping me explore my musical vein), Jim & Sue Sanders (for everything, especially the lovely Thanksgiving parties), Gary & Anna Miron, Chris & Daniela Coryn-Schroeter, Arlen, Janet, Thelma, & Amy Gullickson, Caryn & Joe King, Karen & Craig Russon, Nadini Persaud, Karin Ladley, Lili Rodrigues & Rigo Rincones, Tony & Hyan Berkley, Dawn Winstone, Teri Behrens, Frank Taylor, John & Julie Risley, Anne & Gustavo, Luz Delgado, and so many others that I have very close to my heart.

I also could not have made it through without the support from my peers at the Interdisciplinary Ph.D. in Evaluation. Special thanks to Amy Gullickson, Chris Coryn, Todd Harcek, Nadini Persaud, Daniela Schroeter, Krystin Martens, Mina Zadeh, Paul

Lamphear, John Risley, Ron Visscher, Kristin Richardson, Michelle Woodhouse, Anne Cullen, Brandon Youker, Cristian Gugiu, Lori Wingate, Ryoh Sasaki, Wes Martz, Wiilis Thomas, and Tererai Trent. They have provided ideas and encouragement during all these years. Many thanks from the bottom of my heart to Amy Gullickson and Todd Harcek for kindly accepting the hard task of correcting my broken English in this dissertation; I am sure they had much more work than they ever anticipated.

Heartfelt thanks to all staff from the Evaluation Center that have provided great support for all the projects I have managed during the past few years: Mary Ramlow, Christine Hummel, Joe Fee, Sally Veeder, Patti Negreviski, Gary Miron, Lori Wingate, Anne Cullen, Liesel Ritchie, and Arlen Gullickson.

I am indebted to some members of the InterAction's Evaluation Interest Group, especially, Jim Rugh, Carlisle Levine, Menno Wiebe, Megan Steinke, and Heather Dolphin. They have provided very helpful feedback and great encouragement during all the hard process of designing, implementing, and analyzing results of the survey with 50 representatives from international NGOs which was a key component of my dissertation.

I also would like to thank my colleagues in Brazil for helping me in different ways during my temporary stay in the U.S.: Ivete Pomarico de Souza, João Carlos Monteiro, Thereza Penna-Firme, Anna Thereza Leão, Eduardo Marino, Daniel Brandão, and Rogério Silva.

My exciting experience in the U.S. had also its many roadblocks and distresses, but I have no regrets—would certainly do it all over again. Now let's see what new

adventures life will bring my way.

Thomaz Kauark Chianca

TABLE OF CONTENTS

Table of Contents—continued

CHAPTER

CHAPTER

Table of Contents—continued

CHAPTER

Table of Contents—continued

Table of Contents—continued

APPENDICES

LIST OF TABLES

CHAPTER I

INTRODUCTION

In 2005, the equivalent of 106 billion U.S. dollars from affluent countries was officially devoted to aid to developing countries (United Nations 2006). Each year, approximately 165 U.S.-based International Non-Governmental Organizations (INGOs)[1], members of the American Council for Voluntary Action (InterAction), mobilize more than $4 billion, just from private donors, in additional aid contributions (InterAction 2007). These funds are used to support and/or implement development, relief, or advocacy initiatives in every developing country in the world. Donors pose hard questions about how their substantial investments are used. They want to know whether their contributions are meeting the needs of the people in the recipient countries. They want to be certain appropriate measures are being taken to ensure those resources are been used with probity and with the most possible efficient means. Solid evaluation policies and practice are, undoubtedly, a main strategy to providing acceptable and consistent answers to these important questions.

Even though evaluation has been intertwined with international aid work since its inception in the late 40's-early 50's, it is an area that has room for improvement and, by its very nature, demands it. However, the quality of evaluations in development aid has been considered by scholars and practitioners quite disappointing overall. Some have argued that evaluations of international development efforts are methodologically weak and, therefore, are not providing reliable information that can help improve the work done by donor agencies and determine the impact of the resources being spent (Clements

[1] A list of acronyms used throughout this dissertation can be found in Appendix A.

2005a; Leading Edge Group 2007; Savedoff et al. 2006). The Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP) has conducted four annual independent meta-evaluations (2001-04) regarding the quality of samples of evaluations of humanitarian responses from its members. ALNAP has found that, even though improvements have gradually occurred overtime and that evaluation has become more deeply integrated in the sector, "the quality of the evaluations themselves still leaves much to be desired" (ALNAP 2006, p. 3).

The literature contains studies showing mixed results in terms of the quality and usefulness of evaluations of INGO interventions. Three[2] publicly available studies commissioned by CARE International of samples of evaluation reports of projects supported by that agency throughout the world (the CARE MEGA[3] evaluations) are good examples. The independent evaluators responsible for the studies indicated that, overall, a great proportion of the evaluations reviewed lacked rigorous designs and focused primarily in measuring projects' outputs rather than impacts or outcomes (Goldenberg 2001, p. 1; Goldenberg 2003, p. 8; Russon 2005, p. 1-3). They also recognized that there was evidence of increasing improvements in the quality of the assessed evaluations, especially between the first (1994 to 1999) and second (2000 to 2002) studies,

Their perceptions corroborate findings from Kruse et al. (1997) from their study involving the review of 60 reports of 240 projects conducted in 26 developing countries:

> … in spite of growing interest in evaluation and growing numbers of evaluation studies, there is still a lack of firm and reliable evidence on the impact of NGO development projects and programmes. Most impact assessments rely on qualitative data and judgements and most are undertaken very rapidly. The majority have [sic] been content to report on and record outputs achieved and not outcomes achieved, or broader impact (p. 7).

[2] A forth MEGA evaluation was finalized in July 2007 by Jim Rugh, but was not summarized in this dissertation due to time constraints.
[3] Meta-Evaluation of Goal Achievement of CARE Projects

A study by Chianca with a sample of 50 U.S.-based INGOs, conducted as part of this dissertation (see detailed results on Appendix B), helped provide additional information about the current situation of evaluation principles and practice in the sector. The study reveled that (i) less than one half of the agencies (44 percent) reported having any system to collect evaluation reports of programs, projects or other efforts they sponsor or implement; (ii) about one-fourth (28 percent) of the agencies indicated that they periodically synthesize and share findings from the evaluations they sponsor or conduct; (iii) only 8 percent indicated having conducted any formal meta-evaluation of their evaluations; (iv) more than one half of the agencies (54 percent) reported having less than one-third of their programs evaluated by external professionals with evaluation expertise; (v) only 16 percent of respondents indicated that more than two-thirds of their efforts are evaluated by external evaluators; (vi) 52 percent of the agencies claimed to have developed their own monitoring and evaluation (M&E) policies, guidelines or standards; and (vi) 38 percent indicated that their agencies have adopted, to some extent, M&E policies, guidelines or standards developed by other organizations.

A number of efforts to improve the situation of the high proportion of low-quality evaluations of international aid interventions have been put in place by different agencies or consortium of agencies. The underlying assumption is that by improving evaluations, aid agencies will be able to become more effective in helping to meet the needs of the people they serve. Even though sharing similar motivations and objectives, these efforts have different ways to approach the problem. Some of them have greater focus on methodological solutions and push for the development of more rigorous impact evaluations using experimental or quasi-experimental designs (Savedoff et al. 2006; J-PAL 2007; SEGA 2006; MDRC 2007; World Bank 2007a; World Bank 2007b). Other agencies, while maintaining perspective on the importance adopting more rigorous evaluation methods, have instead prioritized the establishment of principles and standards to guide and improve evaluation practice (OECD 1991; InterAction 2005).

A minority of organizations within the ones advocating primarily for "rigorous impact evaluation" such as the Abdul Latif Jameel Poverty Action Lab and the Scientific Evaluation for Global Action, support the exclusive use of randomized control trials (RCTs) as the only acceptable method to assess impact. Their position has generated lively debates in the development field. The majority opposing this idea contends that evaluation questions should be the determining factor when choosing the appropriate method for impact evaluations (NONIE 2007; 3IE 2007).

Efforts in the direction of creating standards for the evaluation of aid interventions have been initiated by many of the most prominent donor agencies including the World Bank and the Development Assistance Committee of the Organization for Economic Co-operation and Development (OECD/DAC)—the organization representing most of the existing bilateral[4] donors. Until now, the OECD/DAC evaluation framework has been the most widely adopted in the field. The importance of standards to improving professional practice has been well described by Picciotto (2006):

> Such rules [standards] underlie the social contract that allows professionals (and the organizations that employ them) to enjoy public trust, practice their craft without undue interference and charge for services rendered. On the supply side, standards enhance the professional stature of those who operate in conformity with them and promote good practices. On the demand side, they facilitate comparisons among providers of services, thus helping customers secure value for money (p. 33).

In the INGO arena, specific evaluation standards and principles are less frequently found. In the United States, InterAction is the organization making important efforts to lead INGOs to develop and adopt evaluation standards. The Evaluation and

[4] Agencies representing a donor country and responsible for establishing individual cooperation efforts with low- or middle-income countries; for example, the U.S. Agency for International Development (USAID), the Swedish International Development Cooperation Agency (SIDA), and the U.K. Department for International Development (DFID).

Program Effectiveness Working Group (EPEWG) has recently proposed a comprehensive revision of InterAction's evaluation standards and guidelines. This revision, if approved by InterAction's Board of Directors, has a real possibility of influencing a relevant number (165 plus) of major INGOs based in the U.S.

The purposes of this dissertation

Studies involving thorough analysis of the main efforts to improve international aid evaluation and of the application of most prominent evaluation standards to the development field are scarce. This dissertation aims at making a contribution to the field in several ways: (i) to provide a general synthesis of the current movements to improve aid evaluation; (ii) to describe and assess the most prominent standards for aid evaluation; (iii) to present a thorough assessment of the prevailing and most adopted set of evaluation criteria worldwide (the five OECD/DAC evaluation criteria) with specific suggestions for improving them; (iv) to discuss the results of a survey with 50 INGOs on their evaluation principles and practice, and their feedback on the recently proposed InterAction M&E standards and guidelines; and (v) in the light of the previous, provide InterAction with concrete suggestions to be considered for future revisions of their M&E standards and guidelines.

This dissertation is divided into five main chapters. First, we will discuss the main efforts in place to improve development aid evaluation and analyze their main limitations and potentials to accomplish their aims. The second chapter is dedicated to study the evaluation standards proposed by donor agencies, the United Nations (UN) system, and evaluation and research networks. Embedded in Chapter III is a thorough analysis of the OECD/DAC evaluation criteria and suggestions for improving them. Chapter IV is dedicated to assess the evaluation standards proposed by INGOs, based on the survey with 50 INGOs members of InterAction, and the most recent set of standards and guidelines proposed by InterAction. Chapter V will present the results of the previously

mentioned survey specifically on the perceptions of INGOs about the new M&E standards and guidelines proposed by InterAction. Finally, Chapter VI will bring together concepts and conclusions from the previous chapters as the basis to propose a set of evaluation standards that should help InterAction and other aid agencies take their M&E standards to the next level. Concluding remarks about possible implications and level of adoption of the suggested new set of standards will be presented, along with additional suggestions and ideas for future investigations in this area.

CHAPTER II

EFFORTS TO IMPROVE THE QUALITY OF INTERNATIONAL DEVELOPMENT EVALUATIONS

There are many efforts currently in place trying to contribute to improve the quality of evaluation in the development world. The following is an analysis of the most prominent and documented efforts in place at the moment. Even though the identification of those efforts was based on an extensive search of the current literature on development aid and on consultation with experts in the field, including his dissertation committee, there might be some unintentional omissions.

The efforts have been classified in five different groups taking into account the organizations leading the efforts: (i) consortia of organizations, (ii) multilateral[5] and bilateral agencies, (iii) INGOs, (iv) professional organizations and networks, and (v) research groups.

Consortia of organizations

Three initiatives have been classified to this group. All of them have been founded and lead by representatives from diverse organizations including multilateral and bilateral donor agencies, UN agencies, INGOs, national government agencies, and research institutes.

[5] International agencies supported by several nations and responsible for coordinating cooperation among more than two states (e.g., the World Bank, the United Nations Development Programme, the African Development Bank)

The International Initiative for Impact Evaluation (3IE)

The 3IE evolved from an initiative developed by the Center for Global Development (CGD) and funded by the Bill and Melinda Gates Foundation and the William and Flora Hewlett Foundation. 3IE was officially created in March 2007 (Leading Edge Group 2007) with ambitious objectives:

> - *identify enduring questions* about how to improve social and economic development programs through structured consultation with Member Institutions and others in order to catalyze comparable studies on selected issues and ensure that studies promoted by 3IE are needed, relevant and strategic;
>
> - *identify programs that represent opportunities for learning* so as to encourage impact evaluations in those instances where studies are feasible, findings can affect policy, and results, when combined with other sources of information, will advance practical knowledge;
>
> - *adopt quality standards* to guide its reviews of impact evaluations through periodic technical consultations;
>
> - *finance the design and implementation of impact evaluations* that address questions of enduring importance to policymaking in low- and middle-income countries;
>
> - *Prepare or commission syntheses* of impact evaluations to link the findings from individual studies with broader policy questions;
>
> - *advocate* for the generation and use of impact evaluations;
>
> - *share and disseminate information* about opportunities for learning, planned studies, designs, methods, and findings; and
>
> - *promote the mutual development of capacity* to conduct rigorous impact evaluations and to use evidence in policymaking in low- and middle-income countries (p. 5).

Members of 3IE include organizations either implementing or funding social and

economic development programs in developing or transitional countries. The list of current agencies interested in participating in the institute include: Mexican Ministry of Health, Ugandan Ministry of Finance, UK Department for International Development, Netherlands Ministry of Foreign Affairs, Canadian International Development Agency, African Development Bank, Bill & Melinda Gates Foundation and William and Flora Hewlett Foundation (Leading Edge Group 2007). Rugh indicated that CARE International has recently become a member of 3IE and that other INGOs are also considering joining this new organization (J. Rugh, personal communication, November 13, 2007 2:45 pm).

The initiative brought together a group of experts to study the reasons for good impact evaluations of development initiatives being so rare and to find possibilities to solving the problem. The expert group generated a report "When Will We Ever Learn? Closing the Evaluation Gap" (Savedoff et al. 2006) which generated some debate in the field, possibly for two main reasons. First they made critiques to current evaluation practice in the sector which spills to all organizations working with development efforts, but especially for the bilateral and multilateral donor agencies who fund most of the aid programs. Second, when defending more rigorous designs to evaluation, they favored random allocation as the primary method of choice for evaluations. More recently, after some harsh critique from the community and probably from further discussions with the different agencies interested in joining the initiative, including bilateral donors (e.g., DFID), they have given up being so explicit about this position. They are being more inclusive in the final version of their founding document stating that the evaluation design should be the most feasibly rigorous one to answer the evaluation questions posed. As a brand new organization and counting on the support of powerful agencies, it will be important to follow whether 3IE will live up to its ambitious goals.

Network of Networks on Impact Evaluation Initiative (NONIE)

As the push for more rigorous methods for assessing impact of development aid was gaining increasing contorts of privileging RCTs and major private donors, such as the Bill and Melinda Gates Foundation, started to support such initiatives, many international development agencies started to voice their discontent in relation to that position. Those dissident voices were publicly heard in major conferences, especially at the 2007African Evaluation Association. Also a movement among donor agencies contrary to the "RCT dictatorship" started to take shape and became formally structured in May 2007, when the Network of Networks on Impact Evaluation Initiative (NONIE) was created (NONIE 2007). NONIE's main objective is "to foster a program of impact evaluation activities based on a common understanding of the meaning of impact evaluation and approaches to conducting impact evaluation" (p. 1).

The primary members of NONIE include the Evaluation Network of the Development Assistance Committee of the Organization for Economic Co-operation and Development (OECD/DAC)[6], the United Nations Evaluation Group (UNEG)[7] and the Evaluation Cooperation Group (ECG)[8]. Representatives from developing country governments (that have partnerships with bilateral, multilateral and UN system agencies) and from existing national or regional evaluation networks can become members of the organization only by invitation from any of the three founding organizations.

In order to fulfill its primary mission of preparing guidance and providing useful resources for impact evaluations, NONIE has established a task team charged with: (i) preparation of impact evaluation guidelines; (ii) establishing collaborative arrangements

[6] OECD/DAC Evaluation Network brings together representatives from evaluation units of 18 bilateral development agencies (e.g., USAID, DFID, SIDA, CIDA, etc.)

[7] UNEG is a network of UN 43 units responsible for evaluation including the specialized agencies, funds, programs and affiliated organizations.

[8] ECG was created by the heads of the evaluation units from the seven existing multilateral banks: African Development Bank, Asian Development Bank, European Bank for Reconstruction and Development, European Investment Bank, Inter-American Development Bank, International Monetary Fund, World Bank Group.

for undertaking impact evaluation, leading to initiation of the program; and (iii) developing a platform of resources to support impact evaluation by member organizations. The task team has already put some of its work on their website including a database with summaries of impact evaluations implemented by one of the network's members, and more resources are expected to be available in the near future.

With many shared objectives with the 3IE group a movement to approximate both organizations has started (Clarke & Sachs 2007). Two statements in the 3IE founding document have clearly contributed to create a positive attitude on NONIE's part towards pursuing collaborative efforts with that group. First, different from the initial general perception of the field, 3IE acknowledged that different methods can be used to conduct rigorous impact evaluations, besides RCTs. The second statement indicated 3IE's interest to find common ground to collaborate with NONIE, as it evolves, especially in terms of:

- defining enduring questions related to the design and conduct of impact evaluations that should be collectively tackled,
- coordinating impact evaluations being conducted in the same countries by level of inquiry or type of program being evaluated,
- sharing databases of ongoing and completed impact evaluations,
- sharing methodological papers and guides, and
- sharing materials and methods for building capacity of partners in designing and conducting impact evaluations. (Rockefeller Foundation 2007, p. 4)

3IE and NONIE have also recognized that there are serious threats to the success of both organizations if they do not pursue a collaborative agenda. Those threats include (i) waste of scarce resources to accomplish same objectives (e.g., development of guidelines and quality standards for impact evaluation, building up databases, etc); (ii) increase in transactional cost for partner countries by asking them to join separate networks and creating confusion by promoting different approaches to impact evaluations

to the same partners; and (iii) reduction in the likelihood of commitment and provision of resources by donor agencies due to lack of coherence between these two organizations.

A pertinent question is whether those organizations should remain as separate entities or whether they should join forces to form a stronger single organization. According to the joint statement produced by Jeremy Clarke, from DFID and representing NONIE, and Blair Sachs, from the Bill and Melinda Gates Foundation representing 3IE (Clarke & Sachs 2007), the two organizations should maintain their own identities and seek funds from different sources. They should, however, establish a clear agenda for collaboration (p. 3). The authors indicated three aspects that are common to both organizations and 13 others that are unique to one or the other organization. Table 1 presents the commonalities and differences presented in the joint statement.

Table 1. Unique and common functions of NONIE and 3IE[9]

FUNCTION	NONIE	3IE
General		
Advocacy and promotion of impact evaluation	No	Yes
Identifying enduring questions and priorities for more impact evaluation work	Yes	Yes
Setting Standards for impact evaluation	No	Yes
Methods		
Alternative approaches to impact evaluation, e.g., on policy influence and macroeconomics, institutional development	Yes	No
Applications of impact evaluation to new Aid Instruments and programs	Yes	No
Guidance on methods and approaches	Yes	Yes
impact evaluation Program Delivery		
Technical support and consultancy to design specific impact evaluations	No	Yes
Mobilizing and providing additional resources for impact evaluation	No	Yes
Financing pool for new impact evaluation proposals from developing countries	No	Yes
Implementing a program of impact evaluations		
Of donor support	Yes	No
Of developing country policy and programs*	No	Yes
Capacity Building in developing countries	No	Yes
Community of Practice and Support		
Network of donors	Yes	No
Network including non state actors and think tanks in developing countries	No	Yes
Resource Platform : Database and website resources	Yes	Yes
Quality Assurance of Impact Evaluations	No	Yes

* 3IE could examine donor support as it contributes to wider programs and are open to direct proposals from donor members.

[9] Adapted from Clarke & Sachs 2007, p. 6.

The analysis in the joint statement presents a few surprises, especially in regards to NONIE's scope of work. If both agencies are committed to increase the number and improve the quality of impact evaluations, it is hard to understand why NONIE does not have as part of its functions the promotion of impact evaluations, development of standards for impact evaluations, and investment in building the capacity of evaluators from developing countries. Since the organizations are still on their infancy, their foci may get clearer as they move along, and some of these apparent inconsistencies might fade away. Nonetheless, both organizations have clear potential to make important contributions to improving the quality of the evaluation of aid interventions. Keeping a continuous flow of communication between the organizations will be essential to increase their impact and, especially, to avoid unnecessary duplication of efforts, imposition of overload and confusion of agencies in developing countries.

Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP)

ALNAP was created in 1997 as one of the efforts to improve performance and accountability of humanitarian interventions which derived from the Joint Evaluation of Emergency Assistance to Rwanda. The 60 full-members of ALNAP meet twice a year and comprise representatives from UN agencies, international NGOs, donor governments, the Red Cross Movement, academics and independent consultants. There are also 600 observing members that are included on a mailing list and kept informed about the main work by ALNAP. Eight full members are elected for ALNAP's Steering Committee and a Secretariat is hosted by the Overseas Development Institute (ODI) in London (ALNAP 2007).

The main activities of ALNAP include its biannual meetings, a yearly publication (the Review of Humanitarian Action), and a wealth of evaluation-related information available in their website. ALNAP has developed specific materials for training

evaluators to work on evaluation of humanitarian action (EHA) that includes (i) a course manual (with background reference documents, definitions of terms, checklists and tables, individual and group exercises, course evaluation, etc), (ii) session summaries (with objectives, suggested timings, answers to exercises, etc), and (iii) PowerPoint slides covering the relevant topics for each session.

ALNAP has also made publicly available a database of evaluation reports of humanitarian action interventions. As of 10/21/07 the database had links to 675 completed reports of evaluations supported by the full-members and other agencies. A small number of those reports are only accessible to professionals belonging to ALNAP's full-member agencies, in accordance with the wishes of the organizations commissioning those evaluations.

Clearly another very important contribution by ALNAP to the field was the development, since 2001, of annual evaluations of a sample of reports from/for? EHA. To orient those meta-evaluations a system was created, the Quality Proforma, with a list of key criteria related to the main aspects to be considered. These aspects include: (i) the terms of reference for the evaluation, (ii) methods, practice and constraints of the evaluation, (iii) analysis made by the evaluators of the context to which the intervention is responding, (iv) evaluation of the intervention, and (v) assessment of the evaluation report. The system also proposes a rating scale ranging from A (good) to D (poor)—no rubrics were provided to anchor the scale (ALNAP 2005). The meta-evaluations are conducted by two independent consultants using the Quality Proforma framework. Meta-evaluations for 2001 through 2004 are posted in their website.

ALNAP has also provided support to efforts for designing and conducting joint evaluations of large humanitarian responses. The most prominent project currently supported by ALNAP is the Tsunami Evaluation Coalition (TEC) which was created in 2005 as the primary driver to evaluate the response of the main relief agencies to the 2004 Tsunami in Asia. TEC brings together 46 different agencies and has released

five thematic reports[10] and one synthesis report examining the how well the response occurred between the first eight and 11 months after the Tsunami. This synthesis report not only draws on the five thematic reports but also in the findings from more than 140 additional reports developed by the agencies involved in the effort (TEC 2007).

Multilateral and bilateral organizations

Five efforts to improve evaluation aid were identified as being lead by donor multilateral and bilateral agencies and by the UN system of agencies.

The World Bank's impact evaluation initiatives

The World Bank (WB) has led, individually, several initiatives to improve the number and quality of development evaluation. PovertyNet is probably the most prominent example of such efforts by the WB. It is a website providing a wealth of resources and analyses for researchers and practitioners on key issues related to poverty, including monitoring and evaluation of poverty reduction strategies (World Bank 2007a). In terms of evaluation, the website offers free access to: (i) guidelines for conducting impact evaluation in particular sectors (e.g., water and sanitation) or under specific constraints (e.g., low budget), (ii) examples of impact evaluations conducted for the World Bank, and (iii) a series of methodological papers dealing with issues relevant to impact evaluations.

Embedded in PovertyNet is the Development Impact Evaluation (DIME) initiative. DIME brings together diverse areas within the World Bank (e.g., thematic networks, regional units and research groups) to coordinate clusters of impact evaluations of strategic interventions across countries in different regions of the world. These

[10] TEC thematic evaluations: (i) Coordination of international humanitarian assistance in tsunami-affected countries; (ii) The role of needs assessment in the tsunami response; (iii) Impact of the tsunami response on local and national capacities; (iv) Links between relief, rehabilitation and development in the tsunami response; and (v) Funding the tsunami response.

evaluations are oriented towards increasing the number of WB impact evaluations in strategic areas, helping develop impact evaluation capacity not only among WB staff but also from government agencies involved in such initiatives, and building a process of systematic learning on effective aid interventions.

Regionally, the WB has a special effort to mainstream rigorous impact evaluation within its supported initiatives in education, malaria, health, infrastructure, and community driven development. The initiative is known as the Africa Impact Evaluation Initiative. It is aimed at building the capacity of national governments of over 20 countries in Africa on conducting rigorous impact evaluations (World Bank 2007b).

The Evaluation Cooperation Group (ECG)

In 1996, the seven existing multilateral development banks[11] created a forum at which their head of evaluation units can meet on a frequent basis to harmonize their work on evaluation issues. Representatives from the United Nations Development Programme (UNDP), the Evaluation Group and the Evaluation Network of the Development Assistance Committee of the Organization for Economic Co-operation and Development (OECD/DAC) are observer members. The main objectives listed by the ECG include:

1. strengthen the use of evaluation for greater effectiveness and accountability,
2. share lessons from evaluations and contribute to their dissemination,
3. harmonize performance indicators and evaluation methodologies and approaches,
4. enhance evaluation professionalism within the multilateral development banks and to collaborate with the heads of evaluation

[11] African Development Bank, Asian Development Bank, European Bank for Reconstruction and Development, European Investment Bank, Inter-American Development Bank, International Monetary Fund, World Bank Group

units of bilateral and multilateral development organizations, and

5. facilitate the involvement of borrowing member countries in evaluation and build their evaluation capacity (ECG 2007).

The ECG website is targeted primarily to the member agencies (most information seems to be on a password protected area) and not much to the external public—even though there are a number of publications on monitoring and evaluation by the member agencies that are made freely available.

The United Nations Evaluation Group (UNEG)

The UN system has also developed its own effort to improve the quality of evaluations and mainstream evaluation functions within their member agencies. The United Nations Evaluation Group (UNEG) was formed as a network of professionals responsible for monitoring and evaluation in 43 units within the UN system including specialized agencies, funds, programs and affiliated organizations. The UNDP has the responsibility to chair UNEG and facilitate opportunities for members to "share experiences and information, discuss the latest evaluation issues and promote simplification and harmonisation of reporting practices" (UNEG 2007).

UNEG is playing an important role in the ongoing UN organizational reform by providing guidance on how to structure a UN-wide evaluation system that will help make the evaluation work within the agency more coherent and with higher quality. Some of the most relevant contributions from UNEG to the establishment of a more coherent evaluation system within the UN agencies were the creation of a set of evaluation norms and one of evaluation standards. Those documents set basic rules to be followed by all UN agencies and that should facilitate collaboration among them on designing and conducting evaluations (UNEG 2005a and UNEG 2005b). Those evaluation norms and standards will be discussed in more details later on in this dissertation.

OECD/DAC Network on Development Evaluation

Probably the oldest effort to bring donor agencies together around evaluation issues was the on by the Development Assistance Committee of the Organization for Economic Cooperation and Development[12] (OECD/DAC). In the late 60's and throughout the 70's, it was one of the first development agencies to officially address some key issues about evaluation methodology, and to organize a series of seminars bringing together evaluators from different parts of Europe. In 1980, a sub-group was officially formed to address the issue of aid effectiveness and, within a context of the world petroleum crises, was faced with the challenge of determining the effectiveness of the international aid provided by the OECD member countries. The Group was unable to provide a reasonable answer to the query since the findings from the evaluations commissioned by the different OECD bilateral aid agencies targeted lessons learned. Thus, those evaluations did not provide trustworthy assessments of impacts that would make it possible to draw overall conclusions about the value of aid supported by OECD members. Regardless of this not so successful start, the Group, instead of being terminated, was promoted to the status of a Working Group on Aid Effectiveness with broader aims including strengthening collaboration among evaluation units of bilateral and multilateral agencies, providing guidance on aid effectiveness to DAC based on lessons learned, and building evaluation capacity in developing countries (Cracknell 2000).

A milestone for the OECD/DAC Working Group on Aid Effectiveness's work was the development of the "Principles for Evaluation of Development Assistance"

[12] The Organization for Economic Cooperation and Development (OECD) is an economic counterpart to the North Atlantic Treaty Organization (NATO) and was created in 1947, then called "Organization for European Economic Co-operation" (OEEC), to co-ordinate the Marshall Plan for the reconstruction of Europe after World War II. Currently with 30 country members (with the strongest economies in the world), it is dedicated to help its members "achieve sustainable economic growth and employment and to raise the standard of living in member countries while maintaining financial stability – all this in order to contribute to the development of the world economy." (OECD 2007)

(OECD 1992). Those principles have had great influence in the way evaluation functions have been structured in aid agencies. They have also served as the basis for the establishment of the five evaluation criteria to assess aid interventions which have been widely adopted by OECD/DAC members and, therefore, have significantly fashioned the design and implementation of aid evaluations. There will be a special section in this dissertation that will provide a comprehensive analysis of the five DAC evaluation criteria.

Among the many relevant works the OECD/DAC Network on Development Evaluation is currently doing, it is worth mentioning (i) the DAC Evaluation Resource Centre (DEReC) a free and comprehensive "online resource centre containing development evaluation reports and guidelines published by the Network and its 30 bilateral and multilateral members"; (ii) several publications including the DAC evaluation quality standards, guide to manage joint evaluations, and evaluating conflict prevention and peacebuilding activities; (iii) a follow up study on the extent to which decisions of the Paris Declaration on Aid Effectiveness[13] are being adopted by the different aid agencies; and (iv) leadership on the establishment of NONIE.

International Non-Governmental Organizations (INGOs)

INGOs have participated in the creation of the currently most prominent joint efforts for improving international development aid evaluation including 3IE, NONIE, and ALNAP[14]. The latter seems to be the one where INGOs have a most active participation. The work done by InterAction seems to be the only major movement lead

[13] A high-level meeting in Paris on March 2005 involving Ministers of developed and developing countries responsible for promoting development and Heads of multilateral and bilateral development institutions to define "far-reaching and monitorable actions to reform the ways [they] deliver and manage aid as [they] look ahead to the UN five-year review of the Millennium Declaration and the Millennium Development Goals (MDGs) later [in 2005]." (Paris Declaration 2005) The main actions defined include issues related to ownership, harmonisation, alignment, results and mutual accountability in development aid.

[14] ALNAP has also developed a set of minimum standards for good practice in disaster response (the Sphere project); the Humanitarian Accountability Partnership (HAP) is a membership organization, similar to InterAction, created to enforce the adoption of such standards by agencies working in the field.

exclusively by INGOs in the direction of fostering increased quality of evaluation of aid interventions.

American Council for International Voluntary Action (InterAction)

On the realm of International Non Government Organizations (INGOs), the American Council for Voluntary International Action (InterAction) is playing a major role in improving evaluation principles and practice among U.S.-based nonprofit agencies working internationally on development, relief, advocacy and technical assistance. InterAction congregates more than 165 of such agencies, mobilizing more than 13 billion U.S. dollars every year from private and public donors to support projects in all developing and transitional countries[15]. InterAction has created important opportunities for INGOs to conduct serious discussions about monitoring and evaluation (M&E) issues relevant to their work and has made many important efforts to help their member agencies improve their M&E functions.

The Evaluation Interest Group (EIG) is one example of such efforts. For 14 years, EIG has brought together M&E staff and consultants from INGOs several times during the year for meetings[16] on relevant themes such as implications of Theories of Change to evaluation and effects of U.S. Government new foreign policy to USAID's M&E requirements from INGOs. Once a year, an intensive two and one-half day meeting, called the "Evaluation Roundtable" is held in the same city where the annual conference of the American Evaluation Association takes place, usually a few days prior to the beginning of the conference. The Evaluation Roundtables have been an important venue for the exchange of experiences, collective evaluation capacity building, and generation

[15] This estimation was done by Chianca based on the ,most recent and publicly available information about the InterAction members' annual expenses reports. Sources of information included the agencies' annual reports, the Charity Navigator website, and InterAction's Member Profile (2004-05).

[16] Those are usually half-day, bi-monthly meetings hosted at InterAction's headquarters in Washington DC; possibilities to call in are made available for EIG members unable to participate in person.

of new ideas to advance evaluation policies and practice within INGOs. EIG has also an electronic discussion listserv (IAEVAL) with more than 300 members.

All InterAction members are required to follow financial, operational, programmatic, and ethical standards developed by InterAction (the PVO[17] Standards) in order to maintain their membership status. The enforcement of the standards is done through bi-annual self-certification processes that require agencies to provide documented evidence that they are in fact complying with the different standards so they can renew their membership or, if new members, join InterAction. The specific standards dedicated to M&E in the current version of the InterAction Standards are quite limited, not enough to provide members with the necessary guidance to establish and maintain good M&E systems.

A committee—the Evaluation and Program Effectiveness Working Group (EPEWG)—was created in 2004 to provide InterAction with ideas to help member agencies establish strategies to demonstrate the effectiveness of their work to themselves, their primary stakeholders and the general public. The Working Group produced a position statement, approved by InterAction's Board of Directors on September 20, 2005, which laid out five key actions all members should commit to follow in order to demonstrate agencies' effectiveness:

1. Articulate its own criteria for success in bringing about meaningful changes in people's lives, in terms of its mission and major program goals.

2. Regularly evaluate its progress towards such success.

3. Mainstream relevant monitoring and evaluation in agency policy, systems and culture.

4. Allocate adequate financial and human resources for its strategic evaluation needs.

[17] Private Voluntary Organizations, which is a less used name to call INGOs.

5. Collaborate with partners and stakeholders in developing mutually satisfying goals, methods, and indicators for project and program activities (EPEWG 2005, p. 6).

The Position Statement called for a revision of the InterAction standards based on the five proposed actions that have direct implications to the monitoring and evaluation functions of member agencies. The EPEWG took responsibility to develop a new set of standards related to M&E which they completed in 2006. Since then their ideas have been submitted for review by members through an ample process that included several EIG meetings and a consultative survey answered by representatives from 50 member agencies. EPEWG has just sent (October 4, 2007) the final version of the new M&E standards to be reviewed by the InterAction's proper decision-making channels (PVO Standards Committee and Board of Directors) for possible inclusion as part of their PVO Standards and Self-Certification process. The proposed new InterAction M&E standards will be discussed and assessed in detail later on in this dissertation since they are an essential element for this study.

The EPEWG has serious plans for strengthening InterAction's role as a leading force to contribute for the advance of evaluation in INGOs. The ideas being discussed among the EPEWG members are ambitious but quite promising. They include (i) providing support to InterAction members in strengthening their M&E policies, principles, standards, strategies, systems and staff capacities; (ii) developing strategies to tackle the issue of impact evaluations as a multi-agency effort; and (iii) development of a peer accountability process (J. Rugh, personal communication, July 16, 2007 5:11 pm).

Professional Associations and Networks

Chianca identified four professional organizations that are making specific contributions to advance evaluation aid. Two of them were associations (one formed by individuals and one formed by national and regional evaluation organizations) while the

other two were open networks—one has worldwide influence while the other concentrates its work in Latin American and the Caribbean countries.

International Development Evaluation Association (IDEAS)

Created in 2002, the International Development Evaluation Association (IDEAS) is a membership-based organization congregating evaluators from different countries with the main objective of contributing to improve the quality and expand the practice of evaluation of development aid, especially in developing and transitional countries. In June 2007, IDEAS reported to have 441 members—more than one-half (236) from countries located in Africa, Latin America and Asia. They have organized their first biannual conference in New Delhi, India, on April 2005; their second biennial conference, initially scheduled to be a joint meeting with the Latin American and Caribbean Evaluation Network (RELAC), on May 2007, had to be postponed to 2008 due to difficulties in obtaining needed financial support. IDEAS has led or co-hosted other relevant events such as the symposium on “Rethinking Development Evaluation” (Gland, Switzerland, July 2004), the symposium on “Parliamentary Oversight for Poverty Reduction” involving parliament leaders in Southeast Asia and Africa (Cambodia, October 2005), and two workshops on “Country-Led Evaluations in the Czech Republic in June 2006 and in Niger as part of the Fourth Conference of the African Evaluation Association, January 2007 (IDEAS 2005; IDEAS 2007). IDEAS has an electronic discussion list open only to members and a website with up-to-date information about main events, publications and other resources relevant to international development evaluation. IDEAS has ambitious plans to expand its membership; they aim at having 1,000 individual members by the end of 2008.

International Organization for Cooperation in Evaluation (IOCE)

The International Organisation for Cooperation in Evaluation (IOCE) is an umbrella organization that brings together the national and regional associations, societies and networks of professional evaluators from around the world. IOCE was created in 1999 with a grant from the W.K. Kellogg Foundation that supported the development of the two initial meetings of leaders from the existing evaluation professional associations. IOCE works to increase communication and collaboration among member agencies aiming at strengthening evaluation theory and practice worldwide through "cross-fertilization of ideas, high professional standards, and an open and global perspective among evaluators" (IOCE 2007). In November 2006, IOCE had 12 official members including all five existing regional organizations (Africa, Australasia, Europe, Latin America and the Caribbean, and Russia and the Newly Independent States)[18], and seven national organizations (United States, Canada, Italy, Belgium, Malaysia, Pakistan, and Sri Lanka)[19]. IOCE has still great potential for growth since there were 62 evaluation professional organizations listed in their website in November 2006 (see Appendix C for complete list).

The main priorities for IOCE are (i) support for emerging evaluation associations, societies and networks through provision of resources to guide their organization, consolidation and growth, and participation in regional and/or national evaluation events, and (ii) promotion of international debates about evaluation in "different cultural contexts – nationally and internationally –including issues of social justice and human rights" (IOCE 2007). Most of IOCE activities are conducted using web-based resources to

[18] African Evaluation Association; Australasian Evaluation Society; European Evaluation Society; International Program Evaluation Network (Russia & Newly Independent States); Red de Evaluacion de America Latina y el Caribe (ReLAC).

[19] American Evaluation Association; Canadian Evaluation Society; Italian Evaluation Society; Malaysian Evaluation Society; Pakistan Evaluation Network (PEN); Sri Lanka Evaluation Association (SLEvA); Wallonian Society for Evaluation (Belgium).

maintain costs as low as possible. Even though not officially dedicated to the field of development aid evaluation, given its international nature and diversity of membership, IOCE has engaged in activities tackling issues relevant to development evaluation and has clearly the potential to contribute to improve practice in the field by supporting and strengthening evaluation organizations throughout the world.

MandE News

Another important player among the existing relevant efforts to improve evaluation in development aid is the Monitoring and Evaluation News (MandE News). Created in 1997 by Rick Davies, an independent consultant with vast experience in international development aid, as one of the first websites dedicated to monitoring and evaluation issues in development aid. The website development and maintenance was supported for 8 years (until 2005) by several UK-based INGOs including Oxfam UK, Save the Children UK, and ActonAid UK, among other 7 agencies. It provides a wealth of information for professionals working in international aid monitoring and evaluation including summaries of relevant documents and book, plus indication of important events and training opportunities.

Perhaps the most successful project supported by MandE News is its main electronic discussion list with more than 1,100 members worldwide—Davies claims that the listserv has the majority of its subscribers from countries in Africa and Asia. It is clearly one of the largest listserv dedicated to the field currently active[20]. MandE News also manages other two electronic discussion lists, one on network analysis and evaluation (with 110 members), and one on a new monitoring technique created by Davies (Davies and Dart 2005) that does not use indicators—the 'Most Significant Changes' (with 630 members). Other important features of the website include

[20] The only other similar listserv we are aware of that is larger than MandE News is PREVAL with more than 1.400 members (see description below).

information on (i) special topics (e.g., working with the Logical Framework, the 'Basic Necessities' survey, and transparency: enabling public M&E), (ii) M&E training providers, (iii) specialist M&E websites (e.g., evaluation capacity building, micro-credit systems, peacebuilding), (iv) evaluation societies and networks, (v) M&E units within aid agencies, (vi) evaluation centers, and (vii) M&E glossary.

PREVAL

Probably one of the most prominent regional efforts to advance international development evaluation is PREVAL—Spanish acronym for "Program for Strengthening the Regional Capacity for Monitoring and Evaluation of IFAD's Rural Poverty Alleviation Projects in Latin America and the Caribbean". Even though being supported by the UN's International Fund for Agriculture Development (IFAD) and originally focused on staff and consultants working on their projects in the region, PREVAL has gone way beyond its original intent by becoming an open network involving M&E professionals working in the region. PREVAL's website has one of the most comprehensive collections of information on development evaluation available in Spanish, both original production from professionals from the region and translations from English. Their quarterly newsletter is a useful resource with important information about the evaluation scene in Latin America and the Caribbean (LAC) including trainings opportunities, key papers, new books, news on professional evaluation organizations, and highlights of the work by IFAD in M&E in the region. PREVAL also provides M&E capacity building seminars throughout the region, offers a searchable database of individual consultants and firms working on evaluation in LAC and has an electronic listserv with more than 1,400 subscribers. Another important facet of PREVAL's efforts to improve/strengthen development evaluation in the region has been its support for the creation of national M&E organizations in different countries, and also of the regional organization: RELAC, the Latin American and Caribbean Evaluation Network.

Research Groups

There are at least four research groups that can be considered as making important contributions in the area of international development evaluation. They are the ones that go beyond selling their specialized evaluation services to other organizations in the sector by dedicating part of their time to train development evaluators, advocate for higher-quality work on development evaluation, make available resources and key information to support other groups, and serve as a reference to other professionals and agencies in the field. While three of the identified agencies are directly connected with well known universities, one of them (MDRC) is an independent nonprofit organization.

The Abdul Latif Jameel Poverty Action Lab (J-PAL)

The Abdul Latif Jameel Poverty Action Lab (J-PAL), created in 2003 as part of the Massachusetts Institute of Technology (MIT), is dedicated to research on development and poverty using randomized controlled trials. It is comprised of more than 30 researchers (directors, members and staff), most of them PhD graduates from Harvard University and MIT. J-PAL seems to be expanding quite intensively in the last few years. Signs of its growth can be perceived in the two recently opened regional offices, one in France, to cover Europe, and another one in India, to cover Southeast Asia. Also, since its inception, J-PAL has completed 27 projects and there are, at the moment, 54 ongoing projects in several different countries involving a diverse cadre of content areas including education, health, employment, microcredit, local governance, etc. After reviewing the brief descriptions in their website of a random sample of 10 of their current studies, it is clear that all of them are focused to answer a few very specific impact questions.

J-PAL's influence in the field is also marked by the well-established training courses they offer on a yearly basis on the use of randomized trials in evaluation. They report that evaluators from 30 different countries have attended their 5-day training

sessions offered during the summer 2007 in Nigeria, USA, and India (J-PAL 2007). With no doubt J-PAL has found an important niche to work and has been quite successful in not only attracting new contracts for designing and implementing randomized studies, but also influencing a great number of evaluators and agencies working in the international development field.

The Scientific Evaluation for Global Action (SEGA)

The Scientific Evaluation for Global Action (SEGA), hosted at the Center for International and Development Economics Research (CIDER) at the University of California—Berkley, is another clear example of U.S.-based agencies dedicated to promote the use of randomized control trials to evaluate international development projects. SEGA brings together more than 25 economists and public health researchers from the Departments of Economics, Agricultural and Resource Economics, Political Science, School of Public Health, and the Haas School of Business at UC Berkeley and international health and development centers at UCSF and UCSD.

Apparently SEGA and J-PAL have some significant ties. At least 10 completed or ongoing projects listed in their websites were/are joint efforts among members from both organizations. The evaluation of components of the Mexican conditional cash transfer project to stimulate, among other positive behaviors, school attendance and retention (Progressa) and the evaluation of the primary school deworming project in Kenya are good examples of such close collaboration (SEGA 2006).

Manpower Demonstration Research Corporation (MDRC)

The other one is MDRC—a 34 year-old research organization with offices in New York and Oakland, CA. MDRC congregates 32 senior experts in areas of K-12 education, higher education, families and children, workers and communities, and welfare and

barriers to employment. They claimed to have helped pioneered the use of RCTs in the evaluation of public policies and programs targeted to low-income people. Even though the bulk of their work is within the U.S. borders, MDRC has also been involved in international projects and has been a reference for international development agencies in the used of randomized designs to assess social and development policies or programs. Their website indicates that MDRC has almost 60 ongoing or recently completed projects; they also make freely available a large number of resources to evaluators including 22 working papers on research methodology, 22 “how-to” guides, 8 video achieves, 13 policy briefs, among others (MDRC 2007).

Centre for the Evaluation of Development Policies (EDEPO)

Another organization with a high profile in the field of international development evaluation is the Centre for the Evaluation of Development Policies (EDEPO). The Centre is based at the Institute for Fiscal Studies (IFS), a leading independent research institute on economic analysis in the UK, and at the Department of Economics at University College London. They have a cadre of 42 completed or ongoing research projects since the inception of the center in 2004. Most of the projects listed in their website are research studies targeted to answer specific impact and explanation questions about a given program. They are not explicitly vocal about the use of RCT as “the” method of choice for impact evaluations and seem to have been quite eclectic in the research designs they use. A good example of such diversity of methods can be noticed in the following description of one of their ongoing studies:

> Much of the literature focuses upon documenting the ex-post impact of an income shock and efforts to use historical risk are made difficult by the need to identify valid instrumental variables to account for endogeneity. This project uses a more "direct" approach by asking household heads to assign probabilities to different incomes. Whilst these types of questions can be difficult to implement amongst a population with low levels of

> literacy and numeracy, careful design and explanations can enable this. This project analyses the plausibility of estimates of expected income and income risk obtained from this method using questions contained in the first and second follow up surveys of the Familias en Accion survey. It will also look at ways of improving the method for future surveys of a similar nature. This project will also look at the impact of perceived income risk upon other outcomes of interest, notably investments in human capital (EDEPO 2007).

EDEPO has 14 members (10 staff, three research fellows, and one research associate) and has made available 21 research papers in their website. They do not seem to offer much training opportunities (there are just a few presentations posted in their website), however their impact in the development evaluation field can probably be better inferred by the half a dozen very influential international organizations they work with, including the World Bank, the UK Department for International Development (DFID), and the Inter American Development Bank.

Summary and reflections about the efforts to improve development evaluation

In this section, we have discussed 16 efforts in the direction of improving international development evaluation that have been considered the most prominent at the moment. Three of them are joint efforts involving a number of different types of agencies (e.g., donors, INGOs, UN agencies, research groups); four are led by multilateral and bilateral organizations; one has INGOs as the leading agencies; four were created by professional associations or networks; and four comprise international development research groups and think tanks.

Table 2. Summary of current efforts to improve international aid evaluation

Type	Name	Members
Consortia of organizations	• International Institute for Impact Evaluations (3IE)	Mexican Ministry of Health, Ugandan Ministry of Finance, DFID, CIDA, Netherlands Ministry of Foreign Affairs, African Development Bank, CGD, Gates Foundation, Hewlett Foundation
	• Network of Networks for Impact Evaluation (NONIE)	OECD/DAC Development Evaluation Network, UN Evaluation Group, Evaluation Cooperation Group (multilateral development banks)
	• Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP)	60 full-members including UN agencies, INGOs, donor governments, Red Cross Movement, academics and independent consultants.
Multilateral and bilateral agencies	• PovertyNet, Development Impact Evaluation (DIME), and African Impact Evaluation Initiative	Diverse areas within the World Bank Group
	• The Evaluation Cooperation Group	Heads of evaluation units from the multilateral development banks (AfDB, AsDB, EBRD, EIB, IADB, IMF and WB)
	• United Nations Evaluation Group	43 units within the UN system including specialized agencies, funds, programs and affiliated organizations
	• OECD/DAC Development Evaluation Network	30 heads of evaluation units of bilateral and multilateral development agencies
INGOs	• Evaluation and Program Effectiveness Working Group (EPEWG)	M&E staff and consultants from INGOs members of InterAction
Professional associations and networks	• International Development Evaluation Association (IDEAS)	400+ evaluators working or interested on international development issues
	• International Organization for Cooperation in Evaluation (IOCE)	Five regional evaluation organizations (Africa, Australasia, Europe, LAC, and Russia & NIS), and seven national organizations (U.S., Canada, Pakistan, Italy, Belgium, Malaysia, and Sri Lanka)
	• Monitoring and Evaluation News (MandE News)	International development evaluators; initial institutional support from several INGOs
	• Program for Strengthening the Regional Capacity for Monitoring and Evaluation of IFAD's Rural Poverty Alleviation Projects in Latin America and the Caribbean (PREVAL)	IFAD staff and consultants and hundreds of evaluators working with poverty reduction initiatives in LAC
Research groups	• Abdul Latif Jameel Poverty Action Lab (J-PAL)	30 researchers, most PhD graduates from Harvard University & MIT
	• Scientific Evaluation for Global Action (SEGA)	25 economists and public health researchers from UC Berkeley, UCSF and UCSD

Table 2 – Continued

Type	Name	Members
Research groups (cont.)	• Manpower Demonstration Research Corporation (MDRC)	32 senior experts in areas of K-12 education, higher education, families and children, workers and communities, and welfare and barriers to employment
	• Centre for the Evaluation of Development Policies (EDEPO)	14 researchers and faculty from the Department of Economics at University College London

The OECD/DAC development evaluation network is probably the most influential effort in place given the many substantial contributions they have made to the field (e.g., the OECD/DAC five evaluation criteria), its longstanding work since the 1970's and the broad composition of its membership—all bilateral agencies are active members and the largest multilateral agencies are observers. Those factors make its work likely to reach most agencies conducting development work in the world.

In the INGO realm, at least in the U.S., InterAction seems to be the most active movement to improve the quality of international development evaluation. Given its size and level of representativeness[21], it has the potential to influence a large number of INGOs and make an important contribution to the international aid evaluation field.

While most of the reviewed efforts have more holistic approaches in their strategies to help the field move forward, at least six of them are solely focused on improving the quality of impact evaluations. Those efforts include 3IE, NONIE, WB's impact evaluation initiatives, J-PAL, SEGA, and MDRC. The broad move towards results-based management (RBM) among public sector institutions in the mid-1990s[22] is considered one of the main drivers for the larger efforts—3IE, NONIE and WB (Ofir 2007). The overall disappointment in the field with the lack of rigor in many evaluations

[21] As mentioned in the description of InterAction (page 19), its member agencies expend more than 13 billion U.S. dollars per year in international aid work.

[22] The RBM trend was lead by governments from developed countries such as the U.S., U.K. and Australia that started to refocus the way their agencies operate, with "improving performance (achieving better results) as the central orientation" (Binnendijk 2001, p. 6). It did not take long for the OECD member governments to require their international development agencies (bilateral agencies) to adopt this framework, as did most of the UN and other multilateral agencies such as the World Bank (UNFPA 2007).

of development aid and the still quite profuse focus of such evaluations on measuring aid interventions' outputs instead of outcomes/impacts can also be considered as important factors influencing the creation of such efforts.

It is hard to think anyone would argue against the importance of conducting assessments of expected outcomes of an aid intervention using robust designs as supported by the agencies promoting those efforts. However, an unbalanced focus on outcome measurement, especially the ones that only try to measure a few variables overtime, carries the risk of reducing the evaluation function to a single criterion exercise—i.e., finding out whether the expected or planned outcomes were actually achieved. To determine the quality, value and importance of an aid intervention, however, a thorough evaluator needs to rely on several criteria that go much beyond measuring outcomes. Ethicality, side-effects (negative and positive), sustainability, exportability, environmental responsibility, and cost of the intervention are some of the key elements that need to be considered in any professional evaluation (Scriven 2007).

In a presentation at the 2006 InterAction Forum, Chianca (2006) presented an illustration of what can be missed if the focus on measuring results is the sole criterion in an evaluation of an aid intervention:

> Let's suppose a given INGO has as its mission reduce poverty in developing countries by supporting small farmers through ecologically sustainable practices and new technology. Indeed a series of impact evaluations of a significant cross-section of their programs shows that the program beneficiaries are significantly increasing their income—let's assume here, for the matter of this exercise, that strong evidence has been found linking the program activities to the observed outcomes. This should certainly be a major factor demonstrating the organization's effectiveness, right?
>
> Now, let's suppose that we have an independent evaluator assessing some of the programs supported by this organization and we found out that in many instances:

(i) the parents are taking their children out of school, because they need their help with the larger crop they have as a result of the training, technical support, and input they received from the programs—a clear pernicious side-effect;

(ii) many beneficiaries are selected based on level of friendship with community leaders or their specific ethnic group—a clear ethical issue;

(iii) most programs are using an outdated technology, wasting resources that could have been used to benefit more people in the communities—a clear process issue;

(iv) the programs are significantly more expensive than comparable alternatives—a clear cost issue;

(v) the programs are helping participants increase their income by producing larger crops of specific products that even though in the short term will assure revenues to beneficiaries, given clear signs from the market (overseen by the planners at program inception), are not likely to last for very long—a clear flaw in the needs assessment;

(vi) most main effects of the programs are not likely to last for long after the support from the international NGO ends—a clear sustainability problem; and (to close with a positive perspective)

(vii) beneficiaries are being able to employ other community members that otherwise would not be employed, helping almost double the impact of the programs in reducing poverty—a clear positive unpredicted (and unreported) impact.

Well, after taking into consideration those different factors, maybe our perception of how effective this INGO really is might change considerably… [the] main message … is that by focusing primarily on measuring the attainment of goals, evaluations will miss essential aspects that have a lot to say about the quality, value and importance of a program or an organization. If they are not adequately taken into account, conclusions about effectiveness of programs can become very sloppy, not to mention the planning of follow up actions based on these findings.

Within the group of agencies dedicating their efforts to improve impact evaluations, there are some who have been advocating very strongly for the use of RCTs as the "golden standard" for aid evaluation. The agencies openly pushing this agenda include most of the research centers described earlier—J-PAL, SEGA and MDRC. The 3IE has moderated its initial, more radical (pro-RCT), position after receiving heavy criticisms from the aid evaluation community, including during the most recent conference of the African Evaluation Association (J. Rugh, personal communication, electronic message, February 7, 2007, 8:14 am).

There is little dispute of the qualities of RCTs as a powerful method for assessing expected outcomes (causal effects) of a program and that identifying such outcomes is an important part of current program evaluation practice (Donaldson & Christie 2005). However, there are serious problems with the idea that RCTs should become the hegemonic method for determining impact (and causal relationships) of programs, including aid interventions.

The American Evaluation Association in its response to the U.S. Department of Education's notice of proposed priority to the use of RCTs to evaluate their programs, titled "Scientifically Based Evaluation Methods: RIN 1890-ZA00"[23], made clear some of those problems. The main arguments used include: (i) RCTs are not the only method capable and scientifically rigorous enough to determine causal linkages between observed outcomes and an intervention (e.g., epidemiological studies linking lung cancer to tobacco and rats infestation to bubonic plague); (ii) RCTs can only deal with a limited number of isolated factors that are less likely to capture the multitude of complex factors influencing outcomes, being therefore less effective than other methods that are sensitive

[23] Notice of proposed priority by the U.S. Department of Education, released in Dec 4, 2003, establishing the focus of Federal funding on "expanding the number of programs and projects Department wide that are evaluated under rigorous scientifically based research methods [aka, RCTs]…" (USDE 2003). In practice, this notice meant that virtually all funding for evaluation in the Department of Education would go to experiments using random allocation.

to contextual factors (culture, local conditions, etc.) and open to capture unpredicted causal factors; (iii) there are situations when RCT designs need to be ruled out for ethical reasons (e.g., denying benefits to participants); and (iv) there are many cases when there is not enough data to fulfill the minimum requirements of sample size to develop a RCTs (AEA 2003).

Davidson (2006) lists a number of important evaluands that would not get evaluated in case a radical option, such as the one defended by the Department of Education that generated the AEA response, that only evaluations using RCTs designs would be funded. Her list includes: (i) nationwide programs implemented at the same time (lack of control groups); (ii) programs that are complex, always changing, and differently implemented in different places (instability of measures); (iii) programs targeting small groups/minorities (sample too small); and (iv) truly innovative policies/programs (unpredicted outcomes). She also indicates that formative evaluations focusing on assessing the quality of processes and early results would not lend themselves to RCTs.

In the international development field, RCTs have been used in quite limited situations when interventions are discrete and, apparently, homogeneous. Examples of such use include public health (school deworming), educational technology (use of flipcharts) and conditional cash transfer[24] initiatives (Kremer n.d.). In reality, however, most aid interventions involve several complex components and are marked by (i) heterogeneity in delivery of services/benefits, (ii) possibility of being influenced by several different actors at non-predictable times (e.g., new government policies or programs), and (iii) need for constant adaptation to the changing environment. One could argue that some specific and smaller aspects or parts of those interventions may lend themselves to RCT studies, but not the interventions in their entirety.

[24] Programs that provide small financial support to poor families in exchange for the adoption of some specific measures such as keeping children at school, and taking infants or pregnant women to regular medical visits.

Instead of focusing on improvements of impact assessments, this dissertation aims at contributing to improve international aid evaluation, especially within INGOs, by building on the knowledge and work of more holistic approaches, and proposing improved sets of evaluation standards to guide evaluation practice. A thorough assessment of the most prominent sets of evaluation standards for international aid organizations are discussed in the next chapter.

CHAPTER III

EVALUATION STANDARDS BY DONORS, THE UN SYSTEM AND EVALUATION NETWORKS

Efforts in the direction of establishing guidelines, standards and/or criteria targeting the improvement of evaluation practice within the development sector are longstanding trends. The World Bank's Operations Evaluation Department (OED) was certainly one of the pioneers in this area. In 1976, the OED issued "Standards and Procedures for Operations Evaluation" which provided specific guidance for the evaluation processes conducted at the end of a project (Project Completion Reports) and after a few years of project completion (Project Performance Audit Reports) (Willoughby 2003, p. 11). After this first set of standards several others have been developed by different agencies within the development community.

This chapter will analyze the most prominent evaluation standards[25] proposed by multilateral and bilateral agencies. We begin with the OEDC/DAC criteria, then continue with the evaluation standards of USAID and five multilateral agencies: (i) the UN Evaluation Group (UNEG), (ii) the European Commission Agency for External Cooperation (EuropeAid), (iii) the World Bank's Independent Evaluation Group (IEG), (iv) the Multilateral Development Banks (MDB), and (v) the Global Environment Facility (GEF). We conclude with a summary of the standards proposed by these organizations, classifying them under a specific analytical framework developed specifically for this study.

[25] We have adopted a broad definition for evaluation standards which includes any written orientation to ensure good practice in evaluation. Standards, guidelines and principles are used interchangeably throughout this dissertation.

The OECD/DAC evaluation principles, criteria, and standards

Our analyses will begin with an assessment of the evaluation criteria for development interventions proposed by the OECD/DAC (1991) including specific suggestions of ways for strengthening them. This special analysis is justifiable since those criteria have been the most influential and commonly adopted framework for evaluating development aid projects among both bilateral and multilateral agencies for more than 15 years. Brief analyses of adaptations of the OECD/DAC criteria for two specific areas (humanitarian assistance and peacebuilding) and of the OECD/DAC evaluation standards, still under test phase, that provide guidance on the conduct of evaluations and for reports will follow.

This section will encompass (i) a brief historic background and description of the OECD/DAC evaluation principles and the five evaluation criteria, (ii) a thorough assessment of the five evaluation criteria, (iii) analysis of the adaptations of the OECD/DAC criteria to humanitarian and peacebuilding sectors, and (iv) analysis of the new OECD/DAC evaluation standards.

Historical context and description of the OECD/DAC evaluation principles and criteria

Efforts in the direction of establishing guidelines, standards and/or criteria for improving evaluation practice within the development sector are longstanding trends. The World Bank's Operations Evaluation Department was certainly one of the pioneers in this area. Specifically, in 1976, this department issued "Standards and Procedures for Operations Evaluation" which provided specific guidance for the evaluation processes conducted at the end of a project, the Project Completion Reports, and after a few years of project completion, the Project Performance Audit Reports (Willoughby 2003, p. 11). However, to date, the evaluation standards for development aid, established in 1991, by

the Development Assistance Committee (DAC) from the Organization for Economic Cooperation and Development (OECD) have been by far the most influential work in the field of development evaluation.

DAC was established by the OECD to improve cooperation between the governments of its 30 members (the most affluent nations in the world such as the US, Japan and Germany) and governments of developing or transitional countries. In late 1992, the OECD/DAC released a document (OEDC 1992) devising key principles for aid management. Monitoring and evaluation functions formed a substantial part of those principles.

Since their inception, the OECD/DAC evaluation guidelines have shaped the way most donor agencies and their clients/grantees commission or design and conduct program evaluations. These guidelines are based in six general principles:

1. All aid agencies should have an evaluation policy.
2. Evaluations should be impartial and independent.
3. Evaluation results should be widely disseminated.
4. Evaluation should be used—feedback to decision-makers is essential.
5. Donor and recipient agencies should be partners/cooperate with the evaluation—strengthen recipient agencies and reduce administrative burden.
6. Evaluation should be part of the aid planning from the start—clear objectives are essential for an objective evaluation (p. 132)

The five criteria to evaluate development interventions (relevance, effectiveness, efficiency, impact, and sustainability) are undoubtedly the most known and adopted features that emerged from the OEDC/DAC evaluation guidelines.

The great acceptance and influence of the DAC criteria can be partially explained by the powerful and influential composition of its Committee. More than 30 heads of

evaluation units from virtually all bilateral[26] and multilateral[27] agencies have a seat in the Committee. The agencies represented by these professionals have adopted the five criteria. Even though some of those agencies have introduced small adaptations, interpretations, or expansions, the underlying core ideas of the criteria have been maintained.

INGOs have also been affected by the DAC criteria partially because several of them operate grants from bilateral and multilateral donors and these funders request the integration of the five criteria into the INGO evaluations. There are signs, however, that some INGOs have also integrated the ideas of the DAC criteria independently from official requirements from donors. INGOs that traditionally do not operate with large direct support from donor agencies, such as Heifer Project International, have also adopted the five criteria as part of some of their requests for proposals (RFPs) for evaluations.

The establishment of the DAC criteria can be considered, at the time of its inception, a great step forward in the direction of improving the quality of development evaluations. These criteria shifted the focus of development evaluations away from solely assessing program outputs or use of funds according to what was proposed, or from the adoption of the economic rate of return (ERR)[28] estimation as the single criterion to assess an aid intervention. Instead, these criteria proposed considering a broader set of key elements.

The five DAC evaluation criteria are based on the conception that evaluation is an

---

[26] Agencies representing a donor country and responsible for establishing individual cooperation efforts with low- or middle-income countries (e.g., US Agency for International Development—USAID, Swedish International Development Cooperation Agency—SIDA, UK Department for International Development—DFID)

[27] International agencies supported by several nations and responsible for coordinating cooperation among more than two states (e.g., the World Bank, the United Nations Development Program—UNDP, the African Development Bank)

[28] Interest rate at which the cost and benefits of a project, discounted over its life, are equal. (Business dictionary 2007) Generally speaking, the higher a project's internal rate of return, the more desirable it is to undertake the project. (Investopedia 2007)

assessment “to determine the relevance and fulfillment of objectives, developmental efficiency, effectiveness, impact and sustainability” of efforts supported by aid agencies (OECD 1992, p. 132). The OECD/DAC members view these criteria as essential in guiding development aid evaluation. The following are the current definitions of the criteria provided at the OECD/DAC (2006) Website:

> Relevance: The extent to which the aid activity is suited to the priorities and policies of the target group, recipient and donor. In evaluating the relevance of a programme or a project, it is useful to consider the following questions: To what extent are the objectives of the programme still valid? Are the activities and outputs of the programme consistent with the overall goal and the attainment of its objectives? Are the activities and outputs of the programme consistent with the intended impacts and effects?
>
> Effectiveness: A measure of the extent to which an aid activity attains its objectives. In evaluating the effectiveness of a programme or a project, it is useful to consider the following questions: To what extent were the objectives achieved / are likely to be achieved? What were the major factors influencing the achievement or non-achievement of the objectives?
>
> Efficiency: Efficiency measures the outputs -- qualitative and quantitative -- in relation to the inputs. It is an economic term which signifies that the aid uses the least costly resources possible in order to achieve the desired results. This generally requires comparing alternative approaches to achieving the same outputs, to see whether the most efficient process has been adopted. When evaluating the efficiency of a programme or a project, it is useful to consider the following questions: Were activities cost-efficient? Were objectives achieved on time? Was the programme or project implemented in the most efficient way compared to alternatives?
>
> Impact: The positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended. This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators. The

> examination should be concerned with both intended and unintended results and must also include the positive and negative impact of external factors, such as changes in terms of trade and financial conditions. When evaluating the impact of a programme or a project, it is useful to consider the following questions: What has happened as a result of the programme or project? What real difference has the activity made to the beneficiaries? How many people have been affected?
>
> Sustainability: Sustainability is concerned with measuring whether the benefits of an activity are likely to continue after donor funding has been withdrawn. Projects need to be environmentally as well as financially sustainable. When evaluating the sustainability of a programme or a project, it is useful to consider the following questions: To what extent did the benefits of a programme or project continue after donor funding ceased? What were the major factors which influenced the achievement or non-achievement of sustainability of the programme or project? (p. 1-2)

The five criteria tackle very important aspects of an evaluation. They have the relevant feature of being applicable to the ample range of aid interventions from single projects or groups of projects (programs), to large scale sector interventions (e.g., investment in a country/state health system) or the whole portfolio of interventions supported by a donor agency in a country or state. Also, these criteria are clearly more comprehensive than the set that was commonly used (and still is quite preponderant) to assess the work of international development agencies which comprise measuring outputs, monitoring resources' application, and, where more sophisticated, estimating a project's economic rate of return.[29]

Since its implementation, the DAC criteria have remained relatively unchanged. In 1998, a report was released by the OECD (1998) that included the results of a comprehensive study commissioned by the DAC Working Party on Aid Evaluation focusing on members' experiences with the application of the 1991 "Principles for

[29] ERR estimations are especially common in evaluations of interventions supported by the World Bank.

Evaluation of Development Assistance".[30] The report concluded that the principles were still valid and sound. However, because of changes in the general aid context in many donor countries, the report suggested the need to rethink some of the interpretations and applications of the principles (p. 7).

An assessment of the OECD/DAC evaluation criteria

Given the importance and level of influence of the DAC criteria in the development world, it is appropriate to submit them to independent scrutiny. Three sensible questions to orient a reflection on the five criteria include: (i) Are they sufficient to provide a sound assessment of the quality, value and significance of an aid intervention? (ii) Are they necessary? and (iii) Are they equally important?

To address the first question is to consider whether key elements related to determining merit, worth or significance of an aid intervention were left out of the criteria definitions. To do so, the first step included a careful comparison between the DAC criteria and one of the most comprehensive and current set of program evaluation criteria proposed by Scriven (2007)—the Key Evaluation Checklist (KEC). The results from this initial exercise were critically reviewed and expanded by a group of 10 professional evaluators with broad experience in international development programs and diverse background (public health, community socio-economic development, management, engineering, public administration, political sciences, and education). These 10 professionals, currently pursuing doctoral degrees in evaluation at the Interdisciplinary Ph.D. in Evaluation at Western Michigan University, created a taskforce on international development evaluation and conducted eight meetings over a 4-month period to specifically discuss improvements to the OECD/DAC evaluation criteria.

The overall conclusions were that:

[30] Those are the six overall evaluation principles mentioned earlier in this paper (p. 2) under which the five OECD/DAC criteria were developed.

- The definition of ‘relevance’ currently focuses primarily on the ‘goals and priorities’ of donors or country/local governments, instead of focusing on ‘meeting the needs’ of the targeted population. This criterion should be refocused to address the needs of the intervention’s impactees.
- Similarly to ‘relevance’, the definition of ‘effectiveness’ focuses on determining the extent to which the intervention met its ‘goals’, and not the ‘needs’ of aid recipients. This criterion should be refocused or possibly subsumed under the ‘impact’ criterion, since goals cover only the expected positive results from an intervention.
- The current definition of ‘sustainability’ is limited to prospective (likelihood of) sustainability and do not make any reference to retrospective sustainability (how sustainable it has been). Furthermore, it only mentions the need to consider environmental and financial aspects of sustainability, leaving out other essential elements to the sustainability of interventions such as political support, cultural appropriateness, adequacy of technology, and institutional capacity.
- ‘Efficiency’ even though tackling some of the right issues, falls short on the coverage of ‘costs’ (e.g., non-monetary costs) and ‘comparisons’ (e.g., creative alternatives). Furthermore, the term ‘efficiency’ often gets defined as ‘least costly approach’ but it is a limited definition given the way evaluations are structured. Cost-effectiveness seems a better term to define this criterion.
- Two key criteria are missing: ‘quality of process’ (e.g., ethicality, environmental responsibility) and ‘exportability’ of whole or part of the aid intervention, meaning the extent to which it could produce important contributions to other aid interventions (e.g., via use of its innovative design, approach, or product, and cost savings).

We will now address these points in some detail.

Relevance, Effectiveness and Impact

The main issues emerging from the analyses related to ‘relevance’ and ‘effectiveness’ have the same conceptual root. The DAC criteria seem to assume that the evaluation should be conducted to determine whether the program met the aid intervention goals in order to determine its success. As discussed in the literature (e.g., Davidson 2005, Scriven 1991), using goals as the primary guide to evaluations can be quite misleading because measuring program goals may not necessarily determine the value of the program to the recipients.

With no doubt, program goals are important for planning and monitoring functions. They provide the necessary orientation to managers regarding how the intervention should be implemented and the specific indicators that should be tracked over time in order to measure important aspects of the project outcomes, and to determine how well the intervention is evolving. However, measuring the level of goal achievement can not be considered a sound basis for an evaluation of an intervention because goals, if not grounded in a sound needs-assessment, reflect only the expectations of program designers, managers, and other stakeholders. As such, goals are not necessarily connected to the real needs of the targeted populations. Of course, there are cases where goals are defined based on well-designed needs-assessment, thus making them soundly aligned with the main existing needs. Nevertheless, what is at stake in an evaluation, and should make up the primary aim for an evaluator, is the search for what is really happening as a result of the aid intervention, regardless of what was initially intended by the program managers or other stakeholders. Also, often times, depending on the context, goals can be set too low or too high, and thus not provide a good parameter for evaluating an intervention.

In the definition provided by OECD/DAC for assessing ‘relevance’ of an aid intervention, the evaluator is challenged to consider whether the program design,

activities, and outputs are aligned with the policies and priorities of a target population, fund recipients, and donor agencies. In practice, this discussion usually explains how an aid intervention relates to the donors' and governments' strategies. It certainly helps to establish the context and significance of the intervention for the donors and governments, but it is not necessarily evaluative. While the call for considering priorities of the target group may lead evaluators to take into account people's needs, the other components of the definition are directly connected with the established goals either by the recipient countries or by the donor agencies. This approach can blur the perspective of the evaluators and divert their attention from the core function of the criterion—which should be to determine whether the intervention's design, activities, and initial results are adequate to respond to existing needs. It seems reasonable to make adjustments in the definition of this criterion by focusing the definition on program recipients' needs.

A similar argument applies to 'effectiveness'. In this case, the OECD/DAC definition indicates that the level of goal achievement (or the likelihood of their achievement) should be used as one of the main criteria to determine the merit of an aid intervention. As explained above, program goals can be misleading and a focus on them can sidetrack evaluators from what is really essential, i.e., determining if an evaluand[31] is producing meaningful outcomes that are addressing existing needs instead of fulfilling pre-established goals. Again, if the goals are perfectly aligned with people's needs, then measuring the achievement of the goals will certainly point evaluators to the right direction. However, a good evaluator should never take for granted that the program goals adequately reflect the needs of the target population. Revising the definition of 'effectiveness' to encompass this perspective is another option for improving the DAC criteria.

A more radical possibility could involve the dissolution of this criterion, assuming that it could be subsumed under 'impact'. The logic for the latter option is that 'impact'

[31] Whatever is being evaluated (e.g., programs, projects, policies, etc)

requires a careful and comprehensive assessment of the results produced by an intervention including expected and unexpected, positive and negative impacts. One could argue that the search for the positive and expected impacts would correspond to the revised version of the definition of 'effectiveness' and, consequently, eliminating the necessity for a stand-alone criterion.

It is also relevant to recognize that the concept of 'need' overlaps substantially with 'impact'. It does not seem possible for a project to have highly cost-effective impacts and not address a real need of a population or group. Furthermore, a project should not be negatively assessed for not addressing all the needs of the beneficiary population/group. Implementing agencies are not necessarily competent to address needs outside their area of expertise. Only in some cases, e.g., emergencies, can a project be properly criticized for not addressing the population's most pressing needs (however these may be identified).

Sustainability

The definition offered by the OECD/DAC for 'sustainability' has missed important elements. First, it seems to ignore evaluative studies conducted several years after the original funding has been withdrawn—retrospective studies. Second, while it clearly addresses economic and environmental aspects of sustainability, it falls short in discussing several other essential elements of sustainability such as political support, socio-cultural adequacy, technological appropriateness, and institutional capacity. For instance, if an intervention does not take into consideration the specific culture of a given region or community, even if initial results are positive, the likelihood of maintaining a program intervention will sharply decrease when the initial funding is withdrawn. This is especially relevant to programs that require direct participation of program recipients to achieve success—e.g., in a water and sanitation intervention, community groups are responsible to organize and pay for maintenance of water pumps and pipes. This aspect is

also relevant to the possibly new ‘quality of process’ criterion, since delivering culturally inappropriate activities or services can considerably decrease an evaluator’s assessment of the quality of an aid intervention.

Making those dimensions explicit in the definition of sustainability will certainly strengthen it. It is interesting to note that one of the OECD/DAC members, the Danish International Development Agency (DANIDA), has already included aspects beyond financial and ecologic issues. They identified seven determinant factors for sustainability of aid interventions including: policy support measures, choice of technology, environmental matters, socio-cultural aspects, institutional aspects, economic and financial aspects, and external factors (DANIDA 2006, p. 57).

Efficiency

‘Efficiency’ has been defined by OECD/DAC as the determination of whether aid interventions use “the least costly resources possible in order to achieve the desired results” (OECD 1992; p. 1). The definition clearly states that in order to arrive at good conclusions about efficiency, it is necessary to conduct a cost analysis and compare the intervention with possible alternatives.

There are many important components in a cost-analysis besides direct money cost that are, unfortunately, quite often overlooked in development evaluations. It seems appropriate to urge evaluators to take into consideration non-monetary costs (e.g., participants’ time or stress), as well as other important types of cost including indirect, start-up, close-down, maintenance, and opportunity costs (Scriven 2007).

In terms of assessing alternatives to an aid intervention, it could also be valuable to call the evaluators’ attention to think broadly, and not restrict themselves to the most obvious comparisons. Evaluators should be challenged to consider possibilities that are both less expensive than the current intervention, and more expensive, as long as these alternatives produce reasonably similar results. Thinking about existing alternatives,

including options that could be logically predicted for the future, would also expand the evaluator's perspective in determining the value of the intervention under consideration.

Complementing the current version of the criterion with some specific guidance on what to look for on 'cost' and 'comparisons' could make the criterion even stronger. Furthermore, the term efficiency has been associated more with 'least costly approach' which is a limited definition given the broaden meaning of the criterion. 'Cost-effectiveness' is a more comprehensive term and seems to better define the many concepts embedded under this criterion.

Missing criteria

Finally, 'quality of process' and 'exportability' are key criteria that are missing in the OECD/DAC list. It can be argued that some components of the aid intervention's process are already contemplated under 'efficiency' (e.g., how the intervention is performing in terms of using resources to produce results)[32], and, to some extent, under 'relevance' (e.g., how important the activities and outputs are in terms of addressing people's needs). However, there are a number of very important process elements left out from the five criteria that can be determinant in assessing the quality of an intervention. Those aspects include (i) ethicality (e.g., are any ethical norms not observed in the delivery of services to recipients or in treating staff?), (ii) environmental responsibility (e.g., are the activities completed by the intervention producing current or future damage to the environment?), (iii) scientific soundness (e.g., does the program follow sound scientific knowledge or accepted 'best practice' guidance of the relevant sector, based on research and evaluations of similar interventions?), (iv) adoption of alleged specifications (e.g., is the intervention delivering what was promised?), (v) coverage (e.g., are the

[32] Indeed, some could make the argument that 'efficiency' should be subsumed under the new 'quality of process' criterion. However, this would make the new criterion overweighed, with too many and too important aspects embedded in it. Keeping them separate might be a better solution to avoid the risk of overshadowing some important aspects.

targeted people being covered? do men and women, boys and girls have equal access to benefits? is the intervention covering an appropriate number of recipients?), (vi) responsiveness (e.g., is the intervention adequately responding to the changing environment?), and (vii) stakeholder participation (e.g., do men and women, and/or boys and girls or relevant sub-groups in the society have equal opportunities to participate in program decisions and activities?), and (viii) cultural appropriateness (e.g., are the services and activities being delivered in accordance to local cultural norms?). Failing to provide credible answers to these (and other similar) questions will certainly affect the quality of the evaluation of any aid intervention.

'Exportability' is the other important aspect missing from the five criteria. It determines the extent to which an aid intervention as a whole or some of its elements (e.g., innovative design, approach, or product) is transferable (e.g., could be potentially worth or produce a key contribution) to another setting (Davidson 2005, p. 6). A positive response to the previous question will clearly affect the way an evaluator will determine the importance or significance of an aid intervention, and also the way s/he will assess the intervention's sustainability[33]. It is important to note, however, that the meaningful application of this criterion will require from evaluators broaden knowledge outside the intervention being evaluated, e.g., other similar (or not) aid interventions, and a certain doses of creativity for considering possible applications of successful ideas to other settings. Another caution is the need to avoid confounding 'exportability' with 'replicability'—a criterion loosely and, sometimes, harmfully used in development aid. More often than desired, managers push for the full transferability of a successful aid intervention to other settings, without careful consideration of the specific socio, economic, and cultural specificities with disastrous results.

The addition of 'quality of process' and 'exportability' to the existing list of DAC

[33] Considering sustainability in a broader perspective than only the continuation of program activities beyond donor initial funding.

criteria will make them much stronger.

The relative importance of the OECD/DAC criteria

The current definition of the five criteria implies that they all have the same level of importance. A reasonable question to ask is whether the criteria should have different weights in determining the overall assessment of an intervention. For instance, should the impact produced by a project receive higher weight in comparison to the other criteria in the overall summative assessment about that project?

Even though the establishment of weights for the criteria seems to present some relevant benefits, the accomplishment of such a task is not easy, if at all feasible. It is possible to defend that producing substantial positive 'impact', in many situations, is a more crucial criterion to determine merit and worth of a project than other criteria. For instance, if a project eradicated hunger in a poor region, even if it did not present very good performance in terms of efficiency, sustainability, effectiveness, or relevance, it might still be considered a good project. However, this statement can only be taken seriously if the project's performance in the other criteria was not at a level considered unacceptable. With this in mind, the answer to the appropriateness of weighing the criteria will have first to address the issue of 'bars'.

A 'bar', according to Scriven (1991), is the minimum acceptable level of performance on a criterion below which an intervention will be considered fully unacceptable regardless of its performance on other evaluation criteria. In considering the five DAC criteria, 'impact', 'efficiency' and 'sustainability' criteria should have minimum acceptable levels of performance ('bars') associated with them. If 'quality of process' is included in the DAC criteria, it should also be considered a good candidate for setting 'bars'.

As for the 'impact' criterion, a 'bar' should be established at the dimension 'negative side-effects'—i.e., if an aid intervention is affecting the people or the

environment in any serious detrimental way, then the aid intervention should be considered unacceptable regardless of how well it performs in other criteria (e.g., being efficient, having high quality of implementation, producing positive impacts). 'Bars' should be established for 'efficiency' at the level of waste of scarce resources or high costs (monetary and non-monetary costs). For instance, if an aid intervention is producing good results in meeting people's needs, but, in order to do so, it is requiring much greater resources than what would be acceptable, or, to access benefits, participants need to spend too much time or encounter serious distress (all at unacceptable levels) then the intervention can not be deemed acceptable.

'Sustainability' is also an important dimension that can require 'bars'. An aid intervention will likely be seen as an unwise investment of scarce resources if the positive outcomes produced by the intervention disappear (or are likely to disappear) right after the original funding is withdrawn and the situation of project participants returns to its original, or even less desirable condition. Of course one may argue that the benefits produced during the intervention's lifetime were so significant that they might have overshadowed the lack of sustainability in the future (e.g., several lives were saved). Also, 'sustainability' will only be essential to the extent to which meaningful outcomes are produced by the project for a reasonable cost with no or a minimum/acceptable waste of resources without incurring any ethical negative impact. There is benefit in placing a 'bar' on sustainability, but only after the evaluand clears the 'bar' in the other four criteria.

At least two components of the 'quality of process' criterion—ethicality and environmental responsibility—constitute particularly important features of any evaluand and should have 'bars' associated with them. Discrimination of participants or staff based on gender, religion, ethnicity, sexual orientation, etc, is a serious ethical issue and could justify failing a given intervention even if it performs well in other criteria. Similarly, if an aid intervention is producing important immediate benefits to participants (e.g.,

increase in people's income) but placing environmental conditions into jeopardy, its acceptability becomes questionable. If there are unavoidable damages to the environment due to extreme reasons (e.g., survival), then the program must consider a strong plan for implementing effective measures that will progressively recover the damages.

'Quality of process' also has other components that even though not as crucial as ethics or environmental responsibility, can certainly influence the performance assessment of any evaluand. The main examples include provision of alleged services (if these services address a local need), following acceptable standards of practice in the field, and adoption of most current scientific knowledge.

It is possible to identify a 'bar' for 'relevance', but only in pre-formative or formative evaluation processes. Those are evaluations conducted during the design and implementation phases of an intervention; they provide an opportunity for organizations to use their findings to introduce changes (improvements) to the aid intervention early in the design phase or as its ongoing during the implementation phase. If the evaluand is found not to address existing needs then it is reasonable to conclude that it is not performing at a minimum acceptable level and, therefore, should be immediately revised.

It is hard to defend that 'effectiveness' should lend itself to the establishment of bars. Even if the project's goals and objectives are connected to the needs of the participants, not achieving some of the goals (in part of in full) might not provide grounds to determine that the intervention was unacceptable. This is the case because the intervention might still have provided some important (unexpected) benefits to the participants which were not thought out as objectives/goals of the intervention.

Returning to the issue of weighing, relative to the other criteria, whether the aid intervention is producing meaningful changes in people's lives certainly carries much weight and places the 'impact' criterion on a possible superior position in terms of importance. If an intervention is producing significant impact, even if it is not very

efficient or the original objectives are not being achieved as planned, as long as it clears the ‘bars’ for the other criteria, it will probably be considered a good intervention; while the reverse will not be true—if an intervention is very efficient, but is not really producing relevant impact then it will probably not be considered as good. However, providing a correct numeric weight to ‘impact’ is tricky, since there are no clear grounds to establish that value—should it be weighted 50%, 100% or another percentage more that the other criteria? One way to display a higher level of importance for ‘impact’ in relation to the others would be to set a higher ‘bar’ that criterion.

The five OECD/DAC evaluation criteria have been an important step forward to make the evaluation of aid interventions more comprehensive. However, there are some key issues related to focus (need to refocus ‘relevance’ and ‘effectiveness’ on needs of potential beneficiaries and not on funders’ and/or governments’ priorities), omissions (need to include ‘quality of process’ and ‘exportability’ as part of the criteria) and importance determination (need to establish ‘bars’ for some key criteria) that should be addressed so the DAC criteria can, once again, lead the international aid evaluation field to a more advanced position.

One of the organizations that have re-focused the DAC criteria is ALNAP. We turn next to analyzing their efforts.

ALNAP’s reinterpretation of the OECD/DAC criteria for evaluation of humanitarian action

The great contribution of ALNAP to the field of standards for international aid evaluation was the clarification and expansion of the evaluation criteria specific to humanitarian assistance interventions first proposed by the OECD/DAC (1999). In their 2006 publication (ALNAP 2006), they provide clear guidance on the interpretation of the criteria and include specific real-life case studies of the application of the criteria in evaluations of humanitarian action (EHA). There are some significant differences

between the criteria to evaluate humanitarian action and the five OECD/DAC criteria for evaluating development assistance. Three criteria are basically maintained (effectiveness, efficiency, and impact) with two conceptual refinements: (i) the idea of timeliness (whether the goals were achieved in the expected timeframe) was included within the concept of effectiveness, and (ii) the idea of breadth of impact—macro (sector) and micro (household).

The concept of 'Relevance' is significantly revised by embedding in it the perspective of 'Appropriateness':

> Relevance/Appropriateness--Relevance is concerned with assessing whether the project is in line with local needs and priorities (as well as donor policy). Appropriateness is the tailoring of humanitarian activities to local needs, increasing ownership, accountability and cost-effectiveness accordingly (p. 20).

The criterion of 'Sustainability' is dropped since there is no consensus whether humanitarian action should support longer-term needs. However, there is consensus that connections between emergency aid, recovery and development should be established by humanitarian interventions, and they should be assessed on the extent to which they are able to promote such connections. With that in mind a new criterion was created and named 'Connectedness':

> Connectedness--Connectedness refers to the need to ensure that activities of a short-term emergency nature are carried out in a context that takes longer-term and interconnected problems into account (p. 21).

Two completely new criteria were proposed to deal with important issues not addressed by the original OECD/DAC five criteria: 'Coverage' and 'Coherence':

> Coverage--The need to reach major population groups facing life-threatening suffering wherever they are (p. 21).

> Coherence--The need to assess security, developmental, trade and military policies as well as humanitarian policies, to ensure that there is

> consistency and, in particular, that all policies take into account humanitarian and human-rights considerations (p. 33).

Another important aspect debated by the developers of the EHA criteria is ‘Coordination’. It refers to the practical measures taken by the different agencies involved in a given humanitarian action to align their work. The alignment could include the creation of groups to coordinate their activities such as discussions about geographical targeting and sharing of key information. The difference between ‘Coordination’ and ‘Coherence’ lies solely on the practicality aspect of the coordination. Coherence is only concerned about assessing whether there is alignment between the policies of the different actors involved in the emergency intervention (p. 54). In the most recent reinterpretation effort, ‘Coordination’ was neither made a stand alone criterion nor subsumed under ‘Coherence’. Rather, it was included as one of the sub-criterion under ‘Effectiveness’, since, apparently, it seems to be implied that coordination among donors, NGOs, and government agencies will be one of the goals of a humanitarian action.

ALNAP also established eight aspects that should be considered as cross-cutting themes within all seven criteria: (i) attention to local context (as a determining factor in the results); (ii) utilization of human resources (expertise of field staff, staff turnover, communications, training, etc.); (iii) protection of recipients?(states and individuals protecting people in war); (iv) participation of primary stakeholders in all stages of interventions; (v) amplification of coping strategies and resilience with regard to emergencies; (vi) observation of gender equality; (vii) HIV/AIDS (e.g., interactions between food insecurity and HIV/AIDS in countries with HIV/AIDS prevalence); and (viii) consideration of environmental responsibility.

OECD/DAC evaluation criteria for peacebuilding

Probably the most recent documented effort to develop evaluation standards/criteria for aid interventions has been a collaboration between two networks

supported by OECD/DAC—the DAC Network on Conflict, Peace and Development Co-operation (CPDC) & the DAC Network on Development Evaluation. The main objective of the effort was to develop a set of evaluation criteria specific to evaluations of conflict prevention and peacebuilding interventions (CPPB). To do that, they draw heavily on the existing OECD/DAC five evaluation criteria and on the OECD/DAC seven evaluation criteria for humanitarian action adapted by ALNAP. The Collaborative for Development Action Inc. (CDA), was asked to prepare the foundation paper with input from representatives from both OECD/DAC networks—evaluation and peacebuilding (OECD 2007).

CDA is a firm based in Cambridge, Massachusetts with 20 plus years of experience in peacebuilding projects. They have proposed eight[34] evaluation criteria for the CPPB interventions. Six of them were directly adapted from the OECD/DAC original and EHA criteria (relevance/appropriateness, effectiveness, efficiency, impact, sustainability, and coverage), and two were new criteria (linkages and consistency with CPPB values). Instead of offering specific descriptive definitions, they have indicated a set of key evaluation questions for each one of the criteria.

The relevance/appropriateness criterion explores the capacity of the CPPB intervention to adapt to the rapid changing environment in order to remain relevant/appropriate to the current context. Some specific evaluation questions proposed to guide this part of the process include: Does the CPPB intervention address key driving factors or constituencies of the conflict? Has the effort responded flexibly to changing circumstances over time?

Effectiveness, efficiency and impact were maintained mostly within the same general line of inquiry as in the original criteria. However, sustainability and coverage gained specific questions directly related to the realm of CPPB interventions—e.g., (i)

[34] Coherence has been pointed out as a possible criterion to be included, but is not yet a consensus among members of the group and therefore has not been included in this analysis.

sustainability: “Will the parties to a negotiated agreement honor and implement it? Are effective mechanisms in place to facilitate implementation? Have those who benefit from ongoing conflict or would resist movement towards peace (“spoilers”) been addressed adequately?”; (ii) Coverage: “Are there “hidden conflicts” that receive little or no international attention? Is sufficient attention being paid to emerging violence and conflict prevention in all potentially violent regions?” (p. 18).

The two new criteria can be considered related to the quality of the intervention implementation process or outputs. ‘Linkages’ refers to the level of success of the intervention in establishing needed connections between the key actors in the efforts for preventing conflicts and building peace. The key orienting questions for assessing this criterion include:

> 1) Are individual and grassroots projects or programmes linked to higher levels (national, regional) and to parallel efforts in other domains (micro-macro, across sectors)?
>
> 2) Are country-level initiatives addressing regional/international dimensions of the conflict or linking to efforts that are?
>
> 3) Are interventions focused on key decision makers or power brokers linked with efforts to engage larger populations and constituencies—and *vice versa*?
>
> 4) Are efforts aimed at promoting individual changes in behavior, skills and attitudes linked with change efforts at the socio-political level?
>
> 5) Are different efforts contradictory or undermining each other? (p. 18)

‘Consistency with CPPB values’ is a criterion that measures the extent to which the agencies responsible for managing and implementing CPPB efforts are following basic ethical principles. It also relates to whether those agencies are serving as a role model to other organizations and individuals in the region as far as basic CPPB principles. Examples of such principles include being respectful to divergent opinions

and making sure decisions are made in participatory ways involving all groups (different ethnicities, religions, etc).

OECD/DAC quality evaluation standards

The OECD/DAC Development Evaluation Network has recently developed a set of standards related to the quality of evaluations. As described above, the five OECD/DAC evaluation criteria, their humanitarian action and peacebuilding adaptations were designed to evaluate the quality of aid provided. In contrast, the DAC Quality Evaluation Standards (QES) were created to ensure that evaluations will be conducted in a quality way and will produce quality products (OECD 2006c). QES developers expect the adoption of the standards by the member agencies will facilitate collaboration on joint evaluations and also comparisons and better use of the evaluation findings across agencies.

The QES consist of 32 standards under ten general headings. For the purposes of this dissertation, we will classify the standards under the framework we have been using to analyze all standards (see detail on p. 86). Also, some standards will be collapsed for simplification purposes and since they have some overlap. However, their original number will be included at the end (between parenthesis) to facilitate identification. Twenty standards were considered as refereeing mostly to the quality of the evaluation process and its main product, the evaluation report. They were classified under 11 headings developed by Chianca:

(i) consistent conclusions—should answer evaluation questions and be consistent with findings and clearly distinct from recommendations and lessons learned (9.1. , 10.1. & 10.3.);

(ii) actionable recommendations and meaningful lessons learned—clearly distinct from each other and from the conclusions (9.3. & 10.3.);

(iii) systematic and clear data presentation, analysis, and interpretation (10.2.);

(iv) focused executive summary—succinct and covering main conclusions, recommendations and lessons learned (10.5.)

(v) description and assessment of the intervention logic (2.2.)

(vi) discussion of context—social, economic, political (including countries' and agencies' policies, stakeholders' involvement, local arrangements for project to function, etc.) (3.1., 3.2., 3.3. & 3.4.)

(vii) discussion of methodology—reliability and validity of data and findings, issues of attribution and contributing/confounding factors, strategies for stakeholder consultation, and sampling (4.1., 4.2., 4.3. & 4.4.);

(viii) transparency, reliability and accuracy of information sources (5.1. & 5.2.);

(ix) incorporation of stakeholders' comments regarding the evaluation process and findings (8.1.);

(x) quality control—internal and/or external formative metaevaluation[35] (8.2.); and

(xi) evaluation conducted on a timely fashion and within budget (9.2.).

Four standards were identified as being relevant to the behavior and quality of the evaluation team:

(i) competent and diverse evaluation team—complementary technical skills and content knowledge, gender and geographic origin balance (4.5.);

(ii) independence—from management, implementers and beneficiaries (6.1.);

(iii) ethical—respect rights and welfare of all stakeholders, ensure confidentiality (7.1.); and

(iv) disclosure of disagreements among team members (7.2.).

[35] The evaluation of evaluations. Evaluations should be evaluated on five core dimensions of merit: validity, utility, conduct, credibility, and costs. In other words, evaluations should produce valid justifiable conclusions, be useful to the client and other relevant audiences; be conducted in an ethical, legal, professional, and otherwise appropriate manner; be credible to relevant audiences; and be as economical, quick, and unobtrusive as possible (Davidson 2005, pp. 242-43). The OECD/DAC Quality Evaluation Standards described above offer another set of criteria for metaevaluations.

Finally, three standards were classified as related to actions of evaluation commissioners:

(i) provision of clear direction—on evaluation's rational, purpose, objectives, scope and major questions (1.1., 1.2., 1.3., 2.1. & 2.4.);

(ii) ensuring free and open process—establish necessary measures so evaluators are able to work freely and without interference, having access to all relevant data (6.2.); and

(iii) ensure use—program managers will provide easy access to evaluation reports to intended users of the evaluation and will have a follow up plan on the implementation of the recommendations (9.4.)

The QES are currently in the application test phase. The testing phase will last for three years (2006-2009). The OECD/DAC evaluation network expects to learn from the experience gained from the member agencies and other interested organizations in applying the QES so they can arrive to a more definitive version in the near future.

The USAID evaluation standards

The main focus of this dissertation is on evaluation principles and practice in U.S.-based INGOs. Many such agencies have projects funded by USAID. In fact, almost one-fourth (22 percent) of the 50 INGOs responding to a survey by Chianca (see survey details on Chapter V, p. 113) indicated that their agencies have adopted required M&E guidelines from USAID. Given the large influence of USAID evaluation policies to the INGO community, we will present an in-depth analysis of them. Curiously, our careful review of several USAID documents and the website revealed that the agency does not have a general set of standards, principles or guidance to orient evaluations of the aid interventions they fund.

In the following sections, we will describe our review hat included several different sources within USAID that provide any guidance to the conduct and assessment

of evaluations of interventions funded by them. We will first analyze the only document with a general evaluation policy applicable to the whole agency, the Automated Directives System number 203 (ADS 203). Second, we will assess the contents of the USAID's website dedicated to evaluation, the EvalWeb. Third, we will discuss some of the USAID evaluation guidelines for specific areas such as disaster assistance, food emergencies, and child survival and health. Fourth, we will present the perspectives of some InterAction members about USAID's evaluation requirements. Finally, we summarize the evaluation standards emerging from the different sources reviewed and draw general conclusions about the current situation of evaluation standards at USAID.

The Automated Directives System number 203 (ADS 203)

USAID has a system with all policy directives and required procedures to be followed by all agency's employees, the Automated Directives System (ADS). ADS is divided into six series organized by function: agency organization & legal affairs (series 100), programming policy (series 200), acquisition & assistance (series 300), personnel (series 400), management services (series 500), and budget & finance (series 600). Series 200 contains USAID's policy and guidelines defining how the agency designs programs to achieve development results, implements those programs, and assesses them. It includes policies related to operations and development. Operations policies encompass procedures and methods to plan, achieve, assess, and learn from programs, and are covered in ADS 200 to 203. Development policies define the content of the agency's programs and are covered in ADS 200, and 204 to 209 (USAID 2004a, p.2).

ADS 203 (USAID 2004b) seems to be the most up to date document describing the agency's monitoring and evaluation policies. The document (ADS 203) contains the agency's policy directives and required procedures to its operating units (not for grantees or contractors). The policy provides guidance on how these units should assess the extent to which the activities developed are achieving the intended results. It also indicates how

the units should systematize and share learning from these assessments with other operating units within the agency and with other development agencies (p. 3).

Among the many topics covered in ADS 203 such as development of performance management plans for strategic objectives, selection of performance indicators, and standards for data quality, there is one section dedicated to evaluation. This evaluation section provides a definition for evaluation[36], indicates when an evaluation should be conducted by an operating unit, presents tips on evaluation methodologies[37], and discusses ideas on how to document and share evaluations. Besides the need to measure goal achievement, no information on what other essential elements should be considered in the evaluation of a program, or specific standards to ensure the quality of evaluation processes or of the work by evaluators. This document, clearly, is not directed to orient INGOs in their efforts to assess the programs funded by USAID.

USAID's EvalWeb

In an effort to find out what USAID requires from INGOs in terms of evaluation, a comprehensive search through its website was conducted. It is important to recognize how rich and inclusive the USAID website is. The agency has a system that collects all major documents, including all evaluation reports from both operating units within the agency and funded projects, and makes them publicly accessible. The search of their website led to some key documents that helped shed light into the question of what is required from INGOs in terms of evaluation by USAID.

---

[36] "An evaluation is a relatively structured, analytical effort undertaken to answer specific program management questions. An evaluation can provide a systematic way to gain insights and reach judgments about the effectiveness of specific activities, the validity of a development hypothesis, the utility of performance monitoring efforts, or the impact of other changes in the development setting on achievement of results." (USAID 2004b; p. 23-24)

[37] They clearly favor "rapid appraisal techniques" and strongly encouraging customers' and partners' participation in designing and conducting evaluations.

The agency has a specific section of its website dedicated to monitoring and evaluation issues, called EvalWeb. The first page contains an account of a commissioned study indicating that there has been a significant decrease in the number of evaluation reports submitted to the USAID clearinghouse. It also indicates the agency's senior leadership is sponsoring an ongoing organizational strategy to reverse this situation (USAID 2007).

A list of very suggestive questions, in the section "evaluation tools and resources" seemed to be the one with the most promising information for our research. First we followed what appeared to be the most obvious link to find the information we were looking for: "What are the requirements for USAID evaluations?"; this link led us back to ADS 203, the document directed to operating units already described above.

Our second best option was "How do I conduct a USAID evaluations?" (Sic); the result was quite misleading, since it led to a table titled "USAID Mechanism for Conducting Evaluations" (USAID 2005). The table presented a list of names and contact information for external agencies working in different areas (e.g., democracy and governance, basic education, general business, and macroeconomic) that have been contracted by USAID to perform evaluations and other consulting services. No guidance on how to conduct evaluations of USAID programs was provided there. One potential intention of the webpage is to provide professionals interested in conducting evaluations of USAID funded programs with an opportunity to find examples of ongoing evaluations and, possibly, seek guidance from the contractors.

The other links in the "evaluation tools and resources" section did not take us anywhere with relevant information for our study, except for the one titled "Where can I find a glossary of key terms in evaluation?" The link took us to 3 definitions for evaluation, one from the OECD/DAC, another one from Danida, and one from USAID. This is how evaluation was defined by USAID in that webpage:

> An evaluation provides a systematic way to assess program performance and impact. Program impact is really looking at the question, to what extent have the intermediate results lead to achievement of the strategic objective? Program performance includes:
>
> effectiveness—to what extent were the intermediate results achieved (such as increased agricultural yields by poor farmers);
>
> efficiency—are there ways to achieve the results for less cost or in less time;
>
> sustainability—have the institutions, finances, and personnel requisite for the continued success of the activities been established; and
>
> replicability—to what extent are the circumstances surrounding the intermediate results typical (USAID 2007, p. 1).

These evaluation criteria mostly resemble the ones proposed by the OECD/DAC (1991). However, they omit 'Relevance' and present different interpretations for two of the criteria: (i) 'impact' is limited to goal achievement—they do not consider unexpected or negative impacts, and (ii) 'effectiveness' is restricted to achieving intermediary goals. They also include a new criterion: 'replicability'. The concept of 'replicability' was defined similarly to "generalizability" as proposed by Scriven (2007, p. 14), but with a more limited perspective. From the few words defining it, replicability seems to deal with the issue of how typical the program's context is, so inferences can be made about how its results might be replicated. Generalizability, one the other hand, subsumes the concepts of longevity/durability, sustainability, and exportability[38]. There are no clear signs, however, that the generalizability criteria are required by USAID to be adopted in their evaluation of initiatives supported by them.

[38] Not only in the sense of replication of the assessed program in similar contexts, but especially in the possibility of exporting the program as a whole or some of its components to quite varied contexts and counting on different resources

USAID evaluation guidelines for specific areas

An expanded search at the EvalWeb website, using keywords such as guidelines for program design, monitoring and evaluation, took us to a few interesting documents. The first one was the "Guidelines for Unsolicited Proposals and Reporting", issued in December 15, 2006, by the Office of U.S. Foreign Disaster Assistance—OFDA (USAID 2006a). The Guide is intended to assist organizations in the preparation of proposals for new grants and award modifications and their submission to OFDA. The one page in the document dedicated to monitoring and evaluation provides a reasonably detailed account of what should be monitored in the programs being proposed:

> The monitoring plan should directly relate to the overall and sector-specific information provided in the Program Description section, including the objectives, activities, and planned indicators. … USAID/OFDA will use this information to evaluate whether its resources are being used effectively. The monitoring plan should specify the following:
>
> • The source, method, and time-frame for data collection;
>
> • The office, team, or individual identified to undertake monitoring-related tasks;
>
> • The quality assessment procedures that will be used to verify and validate the measured values of actual performance;
>
> • The known monitoring limitations, the impact the limitations may have on program implementation, and the plans for addressing these limitations as applicable;
>
> • The plans for data analysis, reporting, review, and use;
>
> • The funds spent per sector against the total amount, in order to assess the rate of spending in relation to program implementation (p. 30).

In contrast with this quite detailed description of monitoring requirements, there was only a short paragraph dedicated to evaluation. The paragraph did not provide any specific guidelines on what key elements of a program should be assessed, except to say that an evaluation should identify program outcomes and impact, lessons learned, and best practices. It was indicated that they encourage independent evaluations. The decision on what evaluation standards to follow is left up to the discretion of applicants (or the evaluators).

The document also described what is expected from successful applicants after their grant is approved in terms of reporting to USAID. The first requirement is for performance baseline data which describe "the prevailing conditions of a beneficiary population and/or the situation at the onset of the disaster or program, the magnitude of the problem, and/or the needs that the Recipient's program will address" (p. 41). The other requirements are: (i) quarterly performance reporting (cumulative achievements and a comparison of actual accomplishments against the objectives and indicators), (ii) informal reporting and quantitative data collection (periodic updates on program activities), (iii) financial reports, and (iv) annual/final report (account of program impact based on the pre-defined indicators measured at baseline).

A quite useful appendix in this document presented specific indicators that should be monitored by grantees for the different areas covered by USAID initiatives. The areas covered include: Agriculture and Food Security, Economy and Market Systems, Health, Humanitarian Coordination and Information Management, Logistics/Relief Commodities, Nutrition, Protection (e.g., gender related violence and child-friendly spaces), Risk Reduction (Natural and Technological Disasters), Shelter and Settlements, Water, Sanitation, and Hygiene, Cross-Cutting Themes (e.g., cash distribution, capacity building, HIV/AIDS, Internally Displaced Populations).

The other document retrieved, "PVO[39] Guidelines for Title II Emergency Food Proposals and Reporting" (USAID 2003), did have any specific guidance for INGOs on M&E. The document, still in draft form, was written by the Office of Food For Peace (FFP) which is responsible for managing all Title II food assistance involving other countries. The UN World Food Program (WFP), INGOs, and, sometimes, local government agencies are the main recipients of Title II grants. However, even when WFP remains the principal recipient of U.S. Title II commodities in an emergency program, INGOs very often serve as implementing partners. All INGOs interested in implementing Title II programs need to formally agree to comply with FFP norms including the guidelines being discussed here. The document had a section on monitoring and evaluation. Monitoring was described as tracking measures of the delivery of commodities and other performance indicators, and evaluation as the function of measuring the achievement of agreed upon objectives. Here is the specific description of how the program should be evaluated according to them:

> A. Programs will be evaluated on the basis of stated objectives agreed to with FFP, as part of the yearly program performance review and USAID R4 process. This review will encompass the following:
>
> (1) factors affecting program performance and summary of data on progress toward achieving the FFP/Emergency Relief Strategic Objective (SO), including data on Intermediate Results (IRs);
>
> (2) expected results for the next year (in the event that the program is extended beyond one year);
>
> (3) special concerns/issues; and
>
> (4) proposed change or refinements to the program objectives, if necessary (p. 20-21).

[39] PVO is an acronym for Private Voluntary Organizations, term still used by a few agencies, including USAID and InterAction, to refer to INGOs.

The last document reviewed was a set of evaluation policies for the Child Survival and Health Grants Projects (CSHGP). CSHGP is a USAID program that work with U.S.-based INGOs and their overseas NGO partners to promote sustained improvements in child survival and health outcomes (USAID 2006b). The document first laid out the three main enabling factors for a successful evaluation: (i) participation of all program partners and stakeholders; (ii) having a good program design; and (iii) project staff committed to take action based on the evaluation data. While (i) and (iii) seem reasonable aspects to ensure greater possibilities for use of the evaluation findings, (ii) is quite questionable as a standard for good evaluation. Even when a program is badly designed, it doesn't mean a high quality evaluation of the program can't be conducted.

After providing guidance on the evaluation process (when to evaluate, how to staff an evaluation, etc.), it presented the specific contents of a final evaluation report. This section shed light on the key issues expected to be covered by grantees on their final evaluation report:

(i) Achievement of project objectives (present summary chart comparing baseline and final data for project indicators including health outcomes, capacity development of local partners, health services improvements, policy changes achieved through the project)

(ii) Unexpected successes and constraints

(iii) Potential for sustainability

(iv) Potential for scalability or expansion of impact

(v) Lessons learned

(vi) Quality of project management including planning, finances, information management, personnel, training, logistics, and technical support

(vii) Results highlights: documenting (i) innovative ideas (creative solutions to common problems that seem effective but still need to be better studied) (ii) promising practices (solutions to problems that work well for one

> organization and have potential to work for others), and (iii) best practices (solutions to problems with evidence of both effectiveness and replicability, and are often supported by peer reviewed literature and international standards) (p. 15).

General perceptions from some InterAction members about USAID evaluation requirements

The last piece of information to complete the quilt about M&E standards within USAID was drawn from informal discussions during the 2006 meeting of members of InterAction's Evaluation Interest Group (EIG) in Portland, OR. EIG is comprised of M&E specialists working in or for INGOs. A number of them are responsible for evaluations of USAID grants.

The main concern some of these professionals expressed was the contractual requirement for them to monitor a large set of output indicators. Such a requirement was reported to consume a considerable part of the program's evaluation resources. EIG members indicated that in some cases more comprehensive external evaluations of USAID funded projects were possible. However, the emphasis on tracking output indicators, besides overwhelming program managers and internal M&E experts, usually would not generate useful data for impact or outcomes evaluations.

A summary of USAID evaluation standards

Based on the above analysis, it seems fair to conclude that USAID does not have a general set of evaluation standards/guidelines/policies to orient the evaluation work at the aid intervention level. Nonetheless, if we combine all the aspects mentioned in the different sources examined, it is possible to develop a reasonable short summary of evaluation standards for USAID:

A. Standards for evaluating aid interventions:

a. Effectiveness: determining the extent to which objectives and intermediate results were achieved.
b. Impact: extent to which intermediate results lead to achievement of the strategic objective; includes also unexpected successes and constraints (note: if we can interpret “constraints” as side effects, then this seems like a fair definition of impact).
c. Efficiency: are there ways to achieve the results for less cost or in less time.
d. Sustainability: assessing whether the institutions, finances, and personnel requisite for the continued success of the activities have been established.
e. Replicability: the extent to which the circumstances surrounding the intermediate results are typical.
f. Potential for scalability or expansion of impact.
g. Quality of project management.
h. Innovation: Innovative, promising, and best practices

B. Standards for ensuring quality of evaluation processes:
   a. Participation: involving all program partners and stakeholders.
   b. Sound program design (note: this is clearly not a necessary condition or an attribute of a good evaluation—one can certainly produce a good evaluation of a badly planned program)

C. Standard for ensuring intended evaluation users commitment:
   a. Utility: project staff committed to take action based on the evaluation data.

General conclusions about USAID evaluation standards

Even after this compilation effort, it is clear that there are still many important standards that have not been considered even in our summary of the USAID criteria. Those standards include (i) relevance to participants’ needs, cost, ethicality, and

comparisons (related to standards of evaluation of programs), (ii) metaevaluation, feasibility, validity, and cost-effectiveness (for standards for assessing evaluations), and (iii) all criteria for ensuring quality of evaluators (credibility, independence, systematic inquiry, etc). Our analysis, however, brought to light one interesting aspect that does not seem to be explicitly included in any other set of standards we have examined so far for this study. “Innovative, promising and best practices” seem to be relevant attributes to determine the significance of a program—if a program has developed creative and effective solutions to common problems than the overall merit of the program should be increase. One could argue that this aspect could be subsumed under “generalizability” or “exportability” since they are also directly connected to determining the importance of an evaluand, however, they do not speak exactly to the same issue.

It seems also fair to say that USAID’s orientation to evaluation is more focused on accountability than learning; on compliance (with grant agreement) than relevance of intervention. They put greater priority to the measurement of goal achievement through the monitoring of specific predetermined indicators. This is, undoubtedly, a quite limited perception of evaluation functions. An interesting observation by Rugh is that USAID usually prefers not to include their performance or the performance of the U.S. government as part of the evaluations they commission (J. Rugh, personal communication, November 13, 2007 2:45 pm).

Another interesting aspect to note is that USAID does not openly require or even suggest the adoption of OECD/DAC criteria and standards to evaluate their programs. That is quite different from many of the other 17 bilateral international development agencies who are also members of the OECD/DAC and that have participated in the development of their evaluation standards. The latter have embedded those standards in their individual evaluation policies. Finally, it is curious to note the several different ways definitions and guidelines for evaluation are proposed by different areas within

USAID, which, once more, reinforces the idea that these issues seem far from being well resolved internally.

Evaluation standards in the UN System

The UNEG developed in 2005 two foundational documents that establish the norms and standards to govern the evaluation functions in all UN agencies (UNEG 2005a; UNEG 2005b). Their main objective was to make a significant contribution to harmonize and simplify the way evaluation is structured and implemented by the different UN agencies. To develop the standards, they claim to have drawn on several sources including existing guidelines proposed by UN agencies (e.g., UNICEF, UNDP, ILO, etc.), the OECD/DAC criteria, bilateral and multilateral agencies (e.g., DANIDA, World Bank, EC), professional associations (e.g., AEA, AfrEA), and ALNAP. The UNEG standards lay out the basic principles for the creation of an institutional framework for managing, conducting and using evaluations in each agency. They also provide specific guidance on the competencies and work ethics any evaluator involved in evaluations of aid interventions supported by the UN agencies should have.

A considerable part of the UNEG standards are specific to the functioning or creation of an evaluation unit within an UN agency—e.g., “The Head of evaluation must have the independence to supervise and report on evaluations as well as to track follow-up of management’s response resulting from evaluation” (UNEG 2005a, p. 8). Since this dissertation focuses on standards that can be applicable to aid evaluation in general, we will not include in the following analysis the standards that are too specific to the functioning or creation of evaluation units. On the other hand, there are also some standards that are way too general and were also left out of the analysis—e.g., “Each evaluation should employ design, planning and implementation processes that are inherently quality oriented, covering appropriate methodologies for data collection, analysis and interpretation” (p. 9).

Also, UNEG has differentiated norms from standards, a distinction we decided not to observe in the analysis since they both relate to aspects that should be followed or assessed to ensure high-quality evaluations. Their standards and norms are presented in a specific order in the two documents analyzed; here, however, they are presented in the way that better fits the analytical framework we have been using to discuss all sets of standards. In terms of specific criteria to be taken into consideration when evaluating programs, UNEG supports the use of the five OECD/DAC criteria, plus two new ones: "value-for-money" and "client satisfaction". Even though there are no further descriptions for these two new criteria, the former appears to be quite similar to 'Efficiency' according to the OECD/DAC, (i.e., the best use of available resources). The latter could be argued to be one indicator of 'Relevance' which currently deals with the extent to which the priorities of the recipients of the aid intervention are met, and might not deserve the status as a stand alone evaluation criterion. Also, they propose the assessment of the extent to which gender equity and human rights as cross-cutting themes were observed by the evaluand. Aspects to be assessed would include the evaluand's efforts to promote gender equality and gender-sensitivity, and its attempts to reach marginalized, vulnerable and hard-to-reach groups (UNEG 2005b, p. 19-20).

UNEG indicates that for humanitarian response interventions, besides the five OECD/DAC criteria, coverage, coordination, coherence, connectedness and protection should also be considered. Three of them, coherence, connectedness, and coverage are among the special criteria suggested by the OECD/DAC (1999) to evaluate complex emergencies[40]. 'Coordination' and 'Protection' are also part of the criteria proposed by OECD/DAC but not as stand alone criteria; both are considered as essential sub-criteria under 'Effectiveness'. In the recent interpretation by ALNAP (2006), 'Protection' is considered a cross-cutting theme, relevant to all criteria.

[40] The OECD/DAC criteria for evaluation of humanitarian assistance are discussed in detail in the analysis of the interpretation and further development ALNAP made of these criteria.

UNEG has identified several standards related to the responsibilities of intended evaluation users, especially the commissioners of evaluations—in the UN case, the governing boards of the different agencies. The standards include[41]:

(i) Evaluation policy: all agencies should have an evaluation policy reflecting the overall norms and standards defined by UNEG;

(ii) Adequacy of resources: enough resources should be allocated so evaluation functions can operate effectively and with independence;

(iii) Ensuring impartiality and independence: evaluators should be protected from pressures that might compromise their independence by locating evaluation functions away from management, and protecting evaluators from possible negative repercussions for career development;

(iv) Hiring competent evaluators;

(v) Commitment to use evaluation findings: create mechanisms (e.g., action plans) to follow-up on implementation of evaluation recommendations;

(vi) Ensuring learning: make evaluations available and create mechanisms to systematize and disseminate lessons to improve practice internally and externally (UNEG 2005a, p. 6-7)

(vii) Issuing good Terms of Reference for evaluations: clearly providing the purpose and context as well as establishing realistic scope and describing the process and the product of the evaluation (2005b, p. 10-11).

In terms of standards directly relevant to the quality of evaluations, UNEG indicates at least five[42]:

(i) Impartiality: "…absence of bias in due process, methodological rigor, consideration and presentation of achievements and challenges. It also implies that the views of all stakeholders are taken into account." (UNEG

[41] The headings for the standards were created by Chianca, since they were not defined by UNEG.
[42] The last two standards' headings (stakeholder consultation and quality evaluation report) were created by Chianca since they were not separately defined by UNEG.

2005a, p. 8)

(ii) Transparency: “Evaluation Terms of Reference and reports should be available to major stakeholders and be public documents” (p. 10).

(iii) Contribution to Knowledge Building: “Evaluations should be conducted and evaluation findings and recommendations presented in a manner that is easily understood by target audiences” (p. 11).

(iv) Stakeholder consultation: “Stakeholders should be consulted in the planning, design, conduct and follow-up of evaluations” (UNEG 2005b, p. 11).

(v) Quality evaluation report: “The final evaluation report should be logically structured, containing evidence-based findings, conclusions, lessons and recommendations, and should be free of information that is not relevant to the overall analysis. The report should be presented in a way that makes the information accessible and comprehensible” (p. 16).

In terms of standards directly relevant to the evaluators’ capacities and behaviors, the documents indicate at least seven:

(i) Independence: “…members of an evaluation team must not have been directly responsible for the policy-setting, design, or overall management of the subject of evaluation, nor expect to be in the near future” (UNEG 2005a, p. 8)

(ii) Technical competency: “Evaluators must have the basic skill set for conducting evaluation studies...” (p. 9)

(iii) Evaluation Ethics: “Evaluators must respect the right of institutions and individuals to provide information in confidence and ensure that sensitive data cannot be traced to its source. Evaluators must be sensitive to beliefs, manners and customs of the social and cultural environments in which they work. … evaluators must be sensitive to and address issues of

discrimination and gender inequality. … wrongdoing … cases must be reported discreetly to the appropriate investigative body" (p. 10).

(iv) Evaluability assessment: evaluators should "identify if there is clarity in the intent of the subject to be evaluated, sufficient measurable indicators, assessable reliable information sources and no major factor hindering an impartial evaluation process" (p. 9)

(v) Clear evaluation design[43]: Evaluators should provide an evaluation design that clearly indicates (a) the evaluation criteria to assess the evaluand (cost-effectiveness should be assessed to the extent possible), and (b) a sound methodology that will allow a complete, fair and unbiased assessment of the evaluand (UNEG 2005b, p. 11-13).

(vi) Team diversity: "The composition of evaluation teams should be gender balanced, geographically diverse and include professionals from the countries or regions concerned" (p. 15).

EuropeAid evaluation criteria

EuropeAid is the agency that coordinates all external aid provided by the European Commission (EC). They claim to be world's largest aid donor agency. The Joint Evaluation Unit of EuropeAid's Cooperation Office has established in 2006 specific guidelines for external aid projects and programs supported by the EC, and have established that evaluation "consists in judging the results of public actions in order to check their conformity with set objectives" (EuropeAid 2007).

EuropeAid has adopted the five OECD-DAC evaluation criteria with slight reinterpretation for sustainability and impact, and have established two additional criteria derived from requirements by the EC. In defining sustainability, they complement the

[43] This standard is a compound of UNEG's standards 3.6, 3.7, and 3.8.

general definition by OECD/DAC (continuation of benefits of aid intervention after completion of support from the aid agency and likelihood of long-term benefits) with the concept of “resilience to risk of the net benefits flows over time” which is close to Scriven’s (2007) definition of ‘durability’ as a component of the generalizability evaluation criterion on his Key Evaluation Checklist (p. 14). In defining impact, they have made it more comprehensive by indicating as impact “positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended” (EuropeAid 2005, p. 2). The two new criteria are:

> Coherence/complementarity: This criterion may have several dimensions: 1) Coherence within the Commission's development programme; 2) Coherence/complementarity with the partner country's policies and with other donors' interventions; 3) Coherence/complementarity with the other Community policies
>
> Community value added: The extent to which the project/programme adds benefits to what would have resulted from Member States' interventions in the same context (p. 2).

The new ‘Coherence/complementarity’ criterion proposed by EuropeAid does not seem too different from the concept of ‘Relevance’ by the OECD/DAC evaluation criterion—such as that one, it has important weaknesses. It is true that even if an intervention is considered relevant to the donor’s, country government’s and intended beneficiary community’s priorities it might be considered not exactly coherent or complementary to other policies or interventions in place. However, if an aid intervention is clearly addressing important needs of the target population one must wonder whether it is important to determine if the intervention is coherent or complementary to exiting policies or interventions. A reasonable question to ponder is whether it would be justifiable to render a bad evaluation for such an intervention, just because it does not seem to be coherent or complimentary to EuropeAid work in a given country. Similarly

to the discussion about the ‘Relevance’ criterion on the in-depth analysis of the OECD/DAC evaluation criteria presented in a previous section of this dissertation (p. 38), the focus of an evaluation should be on the needs of the target population and not only on the goals of the aid or governmental agencies. Based on the short definition presented in the document reviewed, the coherence/complementarity criterion seems to be closely aligned with goals, and therefore does not appeal as being a good addition to the set of evaluation criteria of aid interventions.

The ‘community value added’ criterion has not been clearly defined. Based on the few questions included in the reviewed document, the criterion seems to be a measure of whether the EC support adds more benefits or value to the intended beneficiaries than if the intervention had been implemented by one of the EC member states. If this interpretation is correct, then the criterion seems aligned with one of the main components (‘comparisons’) of the already exiting ‘efficiency’ criterion under the OECD/DAC. When considering ‘comparisons’ an evaluator should be exploring alternative ways of conducting the intervention that could have lead to similar or better results, using less, more or equal resources. In this case, exploring the possibility of having the intervention implemented by one of the EC bilateral development agencies or the host country itself, instead of the EuropeAid, could certainly encompass one of the possible alternative models. If this rationale is correct then this new criterion seems redundant and, probably, unnecessary.

World Bank evaluation standards

The Independent Evaluation Group (IEG) is an independent entity within the World Bank that reports directly to the Bank’s Board of Executive Directors and is responsible for the evaluation of all major endeavors supported by the World Bank throughout the developing and transitional world. IEG has adopted an “objectives-based approach” to evaluate development interventions which translates basically into

comparisons between what was achieved by the intervention with what it had planned to achieve. In conducting such evaluations, IEG adopts a framework that includes the assessment of some key aspects of an intervention: outcomes, sustainability, and institutional development impact of WB's operations (World Bank 2006).

Those aspects resemble closely the five OECD/DAC evaluation criteria but with different organization and interpretation for some of the criteria. The first main difference is that 'relevance', 'effectiveness' (or 'efficacy', as they call it) and 'efficiency', even though defined similarly to the definition by the OECD/DAC, are presented as sub-criteria under a new criterion called 'outcomes'. This new arrangement appears to present a conceptual problem. If we take the definition of 'outcomes' by the OECD/DAC (2002) and endorsed by the WB, as the "likely or achieved short-term and medium-term effects of an intervention's outputs" (p. 28) it is not clear why 'relevance' has been put under outcomes. Relevance, as defined by the IEG, refers to the alignment of the intervention's stated objectives[44] with the "country needs and institutional priorities" (World Bank 2006). A clearer connection between 'outcomes' and 'relevance' could be established if relevance was referred to as 'achieved outcomes' and not just to 'stated objectives'.

The definition of 'sustainability' appears to be the same as the one defined by the OECD/DAC, but with a small twist. Instead of just presenting the more general definition for sustainability (likelihood of positive results being sustained after the completion of the project funding period), the IEG indicates what is important to be measured by an evaluation is to determine an intervention's sustainability, which is 'resilience to risk'. Here are the questions they indicate are important to be answered by the evaluators:

> At the time of evaluation, what is the resilience to risks of future net benefits flows? How sensitive is the project to changes in the operating environment? Will the project continue to produce net benefits, as long as intended, or even longer? How well will the project weather shocks and

[44] Instead of 'objectives', the OECD/DAC refers to the intervention's 'design' or 'implementation strategy' (OECD 2005)

changing circumstances? (World Bank 2006).

The 'impact' criterion proposed by the OECD/DAC was focused by the IEG to measure impact in terms of institutional development, or more precisely, "the extent to which a project improves the ability of a country or region to make more efficient, equitable and sustainable use of its human, financial, and natural resources" (World Bank 2006). It is interesting to note that the IEG's focus on measuring objectives has excluded (intentionally or unintentionally) any mention to efforts by evaluators to search for unexpected positive impacts, or bad side-effects that might jeopardize any positive intended impact produced by an aid intervention. That omission is a clear difference from the OECD/DAC criteria.

IEG also has added one new criterion to their list: 'Bank and borrower performance'. This criterion complements the OECD/DAC criteria by bringing to the discussion some important aspects related to the quality of the process of the aid intervention that are ignored by the OECD/DAC criteria. The following is how IEG defines this criterion:

> IEG's assessments of Bank and Borrower Performance focus on how good a job each partner has done during the different stages of the project cycle, i.e, project identification, preparation, appraisal and implementation. Bank performance is judged based on the extent to which services provided by the Bank ensured quality at entry and supported implementation through appropriate supervision (including ensuring adequate transition arrangements for regular operation of the project). Borrower performance evaluates the extent to which the borrower assumed ownership and responsibility to ensure quality of preparation and implementation, and complied with covenants and agreements, towards the achievement of development objectives and sustainability (World Bank 2006).

In terms of general standards related to the quality of evaluations, IEG has proposed four: (i) usefulness (evaluations should produce findings that are timely and address current concerns; it should involve managers, borrowers, co-financiers, and

beneficiaries); (ii) credibility (evaluators should have adequate professional capacity and use rigorous methods); (iii) transparency (evaluation reports are available to all major stakeholders and the evaluations are subject to independent annual reviews); and (iv) independence (evaluators should be independent from line management at all stages).

IEG and the OECD/DAC Network on Development Evaluation have recently released a new publication on evaluation of large-scale, multiple partners, and across-countries initiatives, the GRPP—Global and Regional Partnership Programs (World Bank 2007c). They have developed a set of evaluation criteria that builds on the original OECD/DAC five criteria. The original criteria were slightly adapted to best fit the scope of such complex interventions but their definitions were basically maintained.

They have also included two new criteria that directly related to key components of the GRPP which are 'governance and management', and 'resource mobilization and financial management'. The former assesses the structures and processes created to govern these complex partnerships and their effectiveness in performing key activities such as preparing strategies, allocating financial resources, and reviewing proposals for inclusion in the portfolio. To do that they propose the use of seven principles of good governance: legitimacy, accountability, responsibility, fairness, transparency, efficiency, and probity. The assessment of resources mobilization and financial management includes determining how well the processes of soliciting funds by the program, provision of funds by donors, and the allocation and control of funds work.

Multilateral Development Banks' evaluation criteria for public sector operations

The Evaluation Cooperation Group (ECG)[45] was established initially as a working group charged by the Multilateral Development Banks (MDB) to develop methodology,

[45] The Evaluation Cooperation Group was formed by the heads of the evaluation units of the African Development Bank, Asian Development Bank, European Bank for Reconstruction and Development Bank, European Investment Bank, Inter-American Development Bank and the World Bank.

criteria and ratings for evaluating the banks' public sector operations (MDB n.d.). ECG basically adopted the OECD/DAC criteria with a few reinterpretations and additions.

'Relevance', 'effectiveness' (called 'efficacy' by ECG), 'efficiency' and 'sustainability' are considered the core criteria. They propose the creation of a fifth core criteria, the 'aggregate project performance indicator, which would represent the synthesis of the project's overall performance and would be determined by aggregating the ratings provided to the other four core criteria.

ECG proposes four complimentary evaluation criteria. The first two are related to impact: (i) institutional development impact that would measure the extent to which "a project improves or weakens the ability of a country or region to make more efficient, equitable, and sustainable use of its human, financial and natural resources" (p. 8) and (ii) other impacts that would include, among others, impact on poverty reduction and in the environment. The other two criteria relate to the performance of the main agencies involved in the program: the banks and the governments. The banks' performance would be measured by assessing the quality of services provided to ensure good implementation and future operation of the project. The borrowers' performance would be assessed basically by determining the level of success of the agencies in creating the necessary conditions for project's sustainability through fostering participation by the project's stakeholders in addition to its own support.

Another interesting feature of this original work was the effort ECG made to establish grounds for rating each criterion on a scale. The following is their account:

> Rating: For each rated criterion, MDBs use an even number (mostly four, exceptionally six for greater differentiation) of rating scale points. For the sake of validity, credibility, transparency and comparability, they apply a clearly defined rating for each scale point that reflects a pre-defined set of ranked value terms. Scales are symmetrical in nature (with due regard to the need for eliminating non-evaluable and not applicable cases). Evaluators provide a justification for each rating, where necessary or

useful (p. 10).

They recognize that the ratings should be based on the best judgment an evaluator can make based on evidence from both qualitative and quantitative data, and should be well justified in their report. There is also an alert about the risk of limiting organizations' learning if they only focus on the aggregated project performance criterion to assess the projects, given their clear appeal to use in corporate reporting and in comparing projects, regions, sectors and MDBs.

Global Environmental Facility (GEF)

GEF is an organization that congregates country governments from around the world and is dedicated to help developing and transitional countries fund interventions to protect the global environment. Among the 177 member countries, there are 32 who have donated about 6.2 billion U.S. dollars in grants and mobilized other 20 billion U.S. dollars in co-finance from other sources. These monies fund projects on biological diversity, climate change, international waters, land degradation (primarily desertification and deforestation), ozone layer depletion, and persistent organic pollutants in 140 countries since 1991.

GEF has developed an evaluation policy laying out the basic criteria and principles to guide evaluations of all funded initiatives (GEF 2006). In terms of evaluation criteria for assessing aid interventions, GEF decided to adopt the five from OECD/DAC without any adaptations or different interpretations. They have also established other specific standards to ensure and assess the quality of not only evaluation processes and products, but also the evaluators responsible for those evaluations.

Regarding the quality of evaluations, they have defined six criteria: (i) 'Impartiality"—comprehensive and balanced representation of strengths and weaknesses of the evaluand including the views of all stakeholders; (ii) 'Transparency'—about the

evaluation's purpose, criteria and intended use of findings to all stakeholders; evaluation reports easily available and reader-friendly; (iii) 'Disclosure'—lessons from evaluations should be fed back to policymakers, managers, beneficiaries and the general public; managers should disclose all project related information to evaluators; (iv) 'Partnership'—whenever possible, evaluations should be joint efforts with partner agencies working in the funded countries; project managers and local implementing NGOs in participate in GEF evaluation activities; (v) 'Credibility'—evaluation reports should reflect consistency and dependability in data, findings, judgments and lessons learned; and (iv) 'Utility'—evaluations should be well-informed, relevant, timely, and clearly and concisely presented; conclusions and recommendations should be results-and action-oriented (p. 16-18).

They have also set forward three specific standards related to the quality of the internal and external evaluators involved in the evaluations of GEF interventions: (i) 'Independence'—members of evaluation teams should not have been involved in any activity related to the evaluand; for evaluations under the responsibility of project managers, review mechanisms to verify impartiality and rigor should be considered); (ii) 'Ethical'—evaluators must respect confidentiality of individuals and institutions; if wrongdoing is uncovered evaluators should be discreet in providing this information to the appropriate instances; managers should remain open to the findings, and do not allow vested interests to interfere with the evaluation; and (iii) 'Competencies and Capacities'—evaluators should have the necessary range of expertise (technical, environmental, social science and evaluation) to the job; whenever possible GEF evaluations should utilize local expertise, and also support local evaluation capacity building (p.16-18).

Synthesis and discussion

The OECD/DAC five evaluation criteria have been a strong foundation for international development evaluation[46] since 1991. They have been the most prominent and widely adopted criteria used for aid evaluation by most bilateral and multilateral donor agencies, as well as international non-governmental organizations (INGOs). However, critiques of the quality of development aid evaluation are still quite abundant and best practices have matured since the last update of the criteria in 1998. Thus, it is reasonable to question how those criteria can be improved. In this chapter we provided a critical look at the OECD/DAC criteria and proposed recommendations for changes, including: (i) revisions of definitions (refocus 'relevance' and 'effectiveness' to address people's needs and not agencies' goals; possibly subsume 'effectiveness' under 'impact'; improve coverage of 'cost' and 'comparisons' under 'efficiency'; include 'cultural appropriateness' under 'sustainability'); (ii) addition of key missing criteria ('quality of process' and 'exportability'); and discussions about the level of importance of the criteria ('bars' and 'weights' to the criteria).

OECD/DAC has also proposed adaptations of their five criteria to two specific areas: humanitarian action (revised also by ALNAP) and conflict prevention and peacebuilding. In the ALNAP reinterpretation of the five criteria to the humanitarian sector, they (i) maintained three of them (effectiveness, efficiency, and impact), (ii) revised extensively 'relevance' to include the concept of 'appropriateness' (greater focus on local needs and ownership), (iii) substituted 'sustainability' with 'connectedness' (between short-term emergency activities to long-term development), (iv) proposed two new criteria ('coverage' of all affected people and 'coherence' of policies to take into account humanitarian and human-rights), and (v) established eight cross-cutting themes

[46] For the purpose of this paper, the term "international development evaluation" and "aid evaluation" will be used interchangeably.

that should be taken into consideration in the assessment of the criteria (e.g., local context, stakeholder participation, gender equity, environmental responsibility). In the conflict prevention and peacebuilding area, the OECD/DAC criteria were expanded to eight criteria. They maintained four of the original criteria (effectiveness, efficiency, impact, and sustainability), borrowed the 'relevance/appropriateness' and 'coverage' criteria as defined by ALNAP, and added two new criteria: 'linkages' (between key actors to peacebuilding) and 'consistency with CPPB values' (e.g., ethicality, respect for divergent opinions).

USAID is the agency with the greatest influence over U.S.-based INGOs in terms of evaluation. This fact is justified by the substantial number of INGOs that are supported by USAID grants and, therefore, have to follow guidelines for program monitoring and evaluation required by that agency. Our analysis from several sources concluded that USAID does not have a common set of evaluation standards or criteria for grantees to follow. There are, however, a number of evaluation guidelines proposed by different areas within USAID. A compilation by Chianca of such guidelines provided a list of eight possible evaluation criteria to assess USAID funded interventions. Four are consistent with the OECD/DAC criteria: effectiveness, efficiency, impact, and sustainability. (However, it is important to note that we could not find any requirement or suggestion by USAID for evaluators to follow the OECD/DAC criteria.) The other four criteria are new: replicability, scalability or expansion of impact, quality of project management, and innovation. Two other criteria were also identified and refer to the quality of the evaluation process (participation of stakeholders) and obligations of primary intended users of evaluations (commitment of project staff to use findings).

There have been many initiatives among multilateral agencies to establish standards and/or criteria to orient evaluation of the efforts they support. All of those efforts have adopted some version of the OECD/DAC evaluation criteria, sometimes providing specific interpretations, reorganization, and/or expansions of the criteria. The

UN Evaluation Group (UNEG), besides adopting the OECD/DAC criteria, has established a comprehensive set of standards related to the quality of evaluations (e.g. transparency, knowledge building, quality of evaluation report, etc) and to the behavior of the evaluators (e.g., independence, ethics, technical competency, etc). The group has also included a new set of standards, not yet made explicit by other organizations, relevant to the behavior of commissioners or primary intended users of evaluations (evaluation clients and other stakeholders). The standards include: (i) defining an evaluation policy, (ii) providing adequate resources, (iii) ensuring evaluation impartiality and independence, (iv) hiring competent evaluators, (v) committing to use the evaluation findings, (vi) ensuring learning, and (vii) providing clear guidance and realistic scope for evaluations.

EuropeAid, responsible for all aid provided by the European Commission, adopts the OECD/DAC criteria, with some changes. They made minor reinterpretations for sustainability and impact and included two new criteria (coherence/complementarity and community value added). Both new criteria, however, seem to be already included under components of the OECD/DAC ‘relevance’ and ‘efficiency’ criteria.

The World Bank’s Independent Evaluation Group (IEG) basically adopts the OECD/DAC’s five criteria, including some rearrangement and reinterpretation, and proposes two new criteria. Relevance, effectiveness, and efficiency are included as sub-criteria under a new criterion named outcomes. Since relevance relates to stated objectives and not achieved objectives (outcomes), its classification under outcomes seems unclear. The concept of sustainability was expanded with the introduction of the idea of resilience to risk. Impact was refocused to measure changes produced by the intervention in terms of institutional development of the local, national, or regional agencies involved in the process. They have also proposed two new criteria related to the quality of the intervention’s process which are bank performance and borrower performance. For the large-scale, multiple partners and cross-countries initiatives

supported by the WB, IEG has proposed two additional criteria related to process: ‘government and management’ and ‘resources mobilization and financial management’. Finally, in terms of evaluation criteria related to the quality of evaluations (and not of evaluands), IEG established four: usefulness, credibility, transparency, and independency.

The Evaluation Cooperation Group (ECG), formed by the heads of evaluation units of seven Multilateral Development Banks, has proposed similar evaluation criteria for evaluands as the ones proposed by the IEG with minor adaptations. The main contribution they made was an effort to provide a framework for rating the criteria, using a four-point scale.

The Global Environmental Facility (GEF) claims to have fully adopted the five criteria of the OECD/DAC to evaluate their projects. GEF has also adopted the four criteria proposed by the IEG for assessing the quality of evaluations (usefulness, credibility, transparency, and independency) and added two others: disclosure of evaluation information and partnership with managers and local agencies for conducting the evaluations. GEF has also established three criteria related to the evaluators responsible for GEF evaluations: independence, ethicality, and competence/capacities.

Table 3 summarizes the standards proposed by the different organizations, classifying them under a specific framework emerged from Chianca’s review of almost 40 evaluation standards by aid agencies. During this review, it became clear that even though included under one general set the standards actually address different dimensions—four, to be precise:

(i) Standards related to the evaluands, i.e., inherent to the quality of the aid interventions (e.g., program efficiency, impact, ethicality, cost, etc.).

(ii) Standards referring to evaluation processes and products (e.g., timeliness, validity of conclusions, proper descriptions of context and methodology, etc.).

(iii) Standards dealing with the capacity and behavior of the evaluators (e.g., evaluators' ethics, technical capacity, respect to stakeholders, etc.).

(iv) Standards relative to the evaluation commissioners and other primary stakeholders (e.g., commitment to use evaluation findings, provision of adequate resources, etc).

Table 3. Summary of evaluation standards from bilateral and multilateral agencies

Group of standards	Standards	Organizations								
		DAC	ALNAP	CPPB	USAID	UNEG	EUAID	I EG	MDB	GEF
for Evaluands	Relevance	X	.	.	.	X	X	X	X	X
	Relevance/ Appropriateness	.	X	X	.	.	.	.	.	.
	Client satisfaction	.	.	.	.	X	.	.	.	.
	Community value added	.	.	.	.	.	X	.	.	.
	Efficiency	X	X	X	X	X	X	X	X	X
	Value for money	.	.	.	.	X	.	.	.	.
	Effectiveness	X	X	X	X	X	X	X	X	X
	Impact	X	X	X	.X	X	X	X	X	X
	Scalability or expansion of impact	.	.	.	X	.	.	.	.	.
	Coverage	.	X	X	.	X	.	.	.	.
	Sustainability	X	.	X	X	X	X	X	X	X
	Replicability	.	.	.	X	.	.	.	.	.
	Connectedness	.	X	.	.	X	.	.	.	.
	Linkages	.	.	X	.	.	.	.	.	.
	Coherence/complementarity	.	.	.	.	.	X	.	.	.
	Innovation	.	.	.	X	.	.	.	.	.
	Coherence	.	X	.	.	X	.	.	.	.
	Governance & management	.	.	.	.	.	.	X	.	.
	Resources mobilization & mgmt	.	.	.	.	.	.	X	.	.
	Process[47]	.	X	X	X	.	.	.	.	.

[47] Process may include a broad range of aspects including ethicality (e.g., in service delivery), participation of stakeholders, environmental responsibility, gender equity, attention to HIV/AIDS affected people, respect to people's rights and welfare, quality of project management, etc.

Table 3 – Continued

Group of standards	Standards	Organizations								
		DAC	ALNAP	CPPB	USAID	UNEG	EUAID	I EG	MDB	GEF
for Evaluation processes and products	Consistent/valid/balanced conclusions	X	.	.	.	X	.	X	.	X
	Actionable recommendations/lessons	X	.	.	.	X	.	X	.	X
	Systematic data analysis	X	.	.	.	X	.	.	.	.
	Focused executive summary	X	.	.	.	.	.	.	.	.
	Description program logic	X	.	.	.	.	.	.	.	.
	Discussion of context	X	.	.	.	.	.	.	.	.
	Discussion of methodology	X	.	.	.	.	.	.	.	.
	Reliability of info sources	X	.	.	.	X	.	.	X	X
	Incorporate stakeholders' comments	X	.	.	.	X	.	.	.	X
	Metaevaluation	X	.	.	.	.	.	.	.	.
	Timely & within budget	X	.	.	.	.	.	X	.	X
	Stakeholder participation	.	.	.	X	X	.	X	.	.
	Sound program design	.	.	.	X	.	.	.	.	.
	Transparency of ToR & reports	.	.	.	.	X	.	.	.	X
	Clear reports, appropriate language	.	.	.	.	X	.	.	.	X
for Evaluators	Competence	X	.	.	.	X	.	X	.	X
	Ethicality	X	.	.	.	X	.	X	.	X
	Independence from all stakeholders	X	.	.	.	X	.	X	.	X
	Disclosure of disagreements	X	.	.	.	.	.	.	.	.
	Respect for people	X	.	.	.	X	.	.	.	.
	Conduct evaluability assessment					X				
	Capacity to develop clear/rigorous design					X		X		
	Diversity of evaluation team	X	.	.	.	X	.	.	.	.
for Evaluation commissioners and other stakeholders	Provision of clear direction	X	.	.	.	X	.	.	.	.
	Ensuring free and open process	X	.	.	.	X	.	X	.	X
	Ensuring evaluation use and learning	X	.	.	X	X	.	X	X	X
	Provision of adequate resources	.	.	.	.	X	.	.	.	.
	Hiring capable evaluators	.	.	.	.	X	.	.	.	.
	Ensure partnership w/ other agencies	.	.	.	.	.	.	X	.	X
	Implementing agency (Bank) performance	.	.	.	.	.	.	X	X	.
	Partner (Borrower) performance	.	.	.	.	.	.	X	X	.

CHAPTER IV

EVALUATION STANDARDS FOR INGOS

The history of evaluation in U.S.-based nonprofits dates back to the late 60's. At that time, the U.S. Congress edited the 1969 Tax Reform which included requirements for greater control of the work by nonprofits in order to avoid the observed abuses of tax-exempted charities (Hall 2003). Concomitantly with the discussions of the 1969 Tax Reform, a group of major American philanthropists sponsored an independent high-level commission, led by industrialist Pete Peterson, to make a detailed study of the nonprofit sector (Commission 1970). The main objective for such a commission was to develop long-term policy recommendations to improve accountability and effectiveness of the work done by charitable organizations. Among the recommendations of the commission's study was that "more thoughtful and consistent monitoring and evaluation of foundation grants could not only reduce the possibly of activities by grantees that reflected unfavorably on grant makers, but also for improving the quality and effectiveness of grant-funded projects" (pp. 132-33).

Even though it did not produce significant changes in the sector, the report inspired the Russel Sage Foundation to continue its serious efforts to work with social scientists to apply evaluation methods in assess the work done by foundations. In defining evaluative research in a paper funded and published by Russel Sage, Edward A. Suchman (1967)—quoted by Brim (1973)—provided the first traces of evaluation criteria, i.e., what was seen at the time as the key factors to be considered in assessing social-welfare interventions. Suchman articulated the key factors consisted of systematically determining, the extent to which objectives were achieved and measuring possible negative side-effects (226-27).

Other than the previous account, there is scarce literature on how INGOs started to establish standards, guidelines or criteria to evaluate their efforts. Historically, it is probably fair to say that the establishment of the Impact Evaluation Project at USAID during President Carter's administration (1977-81) has contributed to some extent to discussion on how to do good impact evaluations. However, it was not before the mid- to late-1990's when the INGOs started to develop written rules on how to design, implement and evaluate evaluations.

Rugh (2004) mentioned one of the first discussions of strategies to promote monitoring and evaluation (M&E) in INGOs, during a session at the 1996 conference of the American Evaluation Association (AEA) in Atlanta, GA. He also described probably one of the first efforts pioneering the establishment of evaluation standards for an INGO—the CARE International (CI) Evaluation Policy. CI evaluation standards were issued in 2003 after being devised and improved over more than two years through an intensive and participatory process involving many CI staff from the headquarters of most of the 12 CI members, as well as staff at project and country office levels.

Development and implementation of evaluation standards is relatively new throughout the INGO field. The review of documents describing evaluation standards from 14 U.S.-based INGOs who responded to the survey by Chianca (see details in the next section) indicated that none of those standards were developed prior to 2003. In the same survey, 11 other INGOs indicated that their evaluation standards are still under development. More recently, InterAction has taken the lead in a major effort to establish evaluation standards that should apply to all its 165 plus members (InterAction 2005).

Outside the U.S., there are several movements in the direction of ensuring good practice of INGOs. Some examples of such efforts include the BOND's (British Overseas NGOs for Development) Quality Standards in NGOs[48], Australian Agency for

[48] http://www.bond.org.uk/futures/standards/report.htm

International Development Accreditations for NGOs[49], the International NGOs' Accountability Charter[50], Code of Conduct for The International Red Cross and Red Crescent Movement and NGOs in Disaster Relief[51], and Code of Good Practice for NGOs Responding to HIV/AIDS[52]. None of these organizations, however, have made public so far any articulation of specific standards to guide evaluation practice of the agencies covered by them.

This chapter is an effort to describe the current INGO scene as described by respondents to Chianca's 2007 survey. First, we will first analyze the evaluation standards, guidelines or principles proposed by the 14 INGOs. Those agencies are the ones that have submitted documentation, as part of their response to Chianca's survey, describing the M&E standards they have developed. The second part of the chapter will analyze seven evaluation standards from other agencies. Those standards have been adopted by some of the agencies that responded to Chianca's survey. The last part of the chapter will describe in detail the InterAction M&E standards and guidelines proposed in 2006.

Evaluation standards from 14 InterAction members

From the 50 representatives of INGOs that responded to Chianca's survey, 26 (52 percent) claimed that their agencies have developed their own M&E policies, guidelines or standards, as shown in Table 4. As one would expect, fewer of the smaller INGOs have developed their own M&E standards—only about one-fourth of those agencies have done so. Surprising, only about one-half of the very large agencies (annual expenditures of more than 100 million dollars) indicated in the survey that they had developed their

[49] http://www.ausaid.gov.au/ngos/accreditation.cfm
[50]http://www.greenpeace.org/raw/content/international/press/reports/ingo-charter.pdf#search=%22 International %20Advocacy%20NGO%20Accountability%20Charter%22
[51] http://www.ifrc.org/publicat/conduct/index.asp
[52] http://www.ifrc.org/what/health/hivaids/code/

own M&E standards. Large and medium INGOs seem to encompass the groups with the greatest proportion of organizations with their own M&E policies, guidelines and standards—80 percent and 64 percent, respectively.

Table 4. Distribution of agencies that have developed their own M&E policies, guidelines or standards, according to their size[53]

Agencies that have…	small	medium[54]	large	very large	all
developed their own M&E standards	3	9	8	7	27
not developed their own M&E standards	9	5	2	6	23[55]

Respondents were also asked to provide a copy of any document describing their M&E policies, guidelines or standards, or to indicate a website in case these documents were available online. Among the 27 agencies who answered "Yes" to this question, 14 did not send any supporting documents. The main reason presented by 11 representatives of those agencies was that their M&E policies/guidelines/standards are still under development, and, therefore, not yet ready to be shared with an external audience; two agencies did not present any justification for not sending information, regardless of been contacted at least one time by me, after submitting their completed surveys. Among the 23 respondents who answered "No" to this question, five indicated they are currently in the process of developing their own M&E policies/guidelines/standards.

The 14 documents describing the M&E policies/guidelines/standards submitted by the agencies to support their responses vary extensively. There were handouts with a

[53] Agencies' size was defined based on annual expenses: there were 12 agencies considered small (up to 10 million U.S. dollars per year); 14 medium (between 11 and 50 M/yr); 10 large (between 51 and 100 M/yr); and 13 very large (more than 100 M/yr). We could not obtain information from only one of the agencies.

[54] The response from one organization was changed from "Yes" to "No". According to the information provided by them on a follow-up e-mail message, they claimed to be externally assessed by the Council on Accreditation, but have not developed their own M&E policies, guidelines or standards.

[55] No financial information was available for one of the agencies; therefore it could not be classified in terms of size based on amount of annual expenses. The representative from this agency indicated they had not developed any M&E policies, guidelines or standards.

few pages with bullet points addressing mostly monitoring and evaluation issues. There were also comprehensive documents with several dozens of pages including detailed orientation for program planning/design and descriptions of specific monitoring and evaluation strategies and tools.

For the purpose of this analysis, evaluation standards were considered any guidance or norms provided by the agencies in the documents reviewed aiming at ensuring or improving the quality of evaluations of their agencies' efforts conducted by internal or external evaluators. The way the evaluation policies, guidelines or standards were spelled out by each organization was also quite diverse. Among the 14 agencies, 8 had them under a easily identifiable section such as evaluation "standards", "principles", "policy", "framework", "guidance" or "strategic areas of inquiry", while the remaining 6 had them imbedded within the general description of their M&E systems without being grouped under any suggestive subtitle.

During the analysis of those 14 documents, 21 different evaluation standards were identified. They were classified using the analytical framework discussed in chapter III (p. 86) under four categories: (i) evaluands; (ii) process and product of evaluations; (iii) evaluators, and (iv) evaluation commissioners or other stakeholders. Table 5 shows the standards, organized according to the four groups and the frequency of which they were mentioned by the different INGOs. The agencies are identified by numbers to protect their identity since many of the documents reviewed were for internal use.

Table 5. Evaluation standards mentioned in the supporting documents submitted by representatives of 14 INGOs who responded to the survey

Group of standards	Standards	Agencies													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
for Evaluands	Impact	X	X	X	X	X	X	X	X	X	X	X	X	.	X
	Process	X	.	.	.	.	X	X	X	X	X	.	.	.	.
	Relevance	X	.	X	X	.	.	.	.	.	.	.	.	.	.
	Effectiveness	.	.	X	X	.	X	X	.	X	.	X	X	X	.

Table 5 – Continued

Group of standards	Standards	Agencies													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
for Evaluands (cont.)	Efficiency	.	.	X	X	.	.	.	.	.	.	.	.	.	.
	Cost	.	X	.	.	.	.	.	.	X	.	.	.	.	.
	Sustainability	.	.	X	X	.	.	.	.	.	.	.	X	.	X
for Evaluation processes and products	Participation	X	X	.	.	.	X	X	X	X	X	X	X	X	X
	Transparency	X	.	.	.	.	.	X	.	.	.	.	.	.	.
	Accuracy	X	X	.	.	.	.	X	.	.	.	X	X	X	X
	No unnecessary disruption	.	.	.	.	.	.	.	.	.	.	X	.	.	.
	Thorough reports	X	X	.	X	.	.	X	X	.	X	.	.	X	.
for Evaluators	Team diversity	.	.	.	X	.	.	.	.	.	.	.	.	.	.
	Ethicality	X	X	.	X	.	.	.	.	.	.	X	.	X	.
	Competence	.	X	.	X	.	.	X	.	.	.	.	.	.	.
	Systematic inquiry	.	X	.	.	.	.	.	.	.	.	.	.	.	.
	Respect for people	.	X	.	X	.	.	.	.	.	.	X	X	.	.
	Responsibilities for general welfare	.	X	.	.	.	.	.	.	.	.	.	.	.	.
for Evaluation commissioners	Appropriate resources	.	X	.	.	.	.	X	.	.	.	.	X	X	X
	Use of findings	X	X	.	.	.	.	X	.	.	X	.	X	X	X

As Table 5 indicates, the frequency with which the standards were mentioned by the agencies varied significantly—from 12 times (impact) to only one time (independence and responsibilities for general and public welfare). In terms of groups of standards, it is clear that the ones related to the evaluands and to the evaluations' process and products were the most frequently mentioned in the reviewed documents. There was also some variation in the way the standards were defined in the various documents. The following is a brief analysis of the descriptions presented by the different agencies for each one of the 21 standards identified within the 14 documents reviewed. The definitions are presented according to the four groups of standards defined earlier in this dissertation (p. 86).

Definitions of standards for evaluands

Impact: The need to assess the impact of aid interventions was one of the two most commonly mentioned standards. It was mentioned by 13 of the 14 documents reviewed. Unfortunately, only one agency understood impact on its most comprehensive sense, which includes expected and unexpected as well as the positive and negative effects of the aid intervention. All others confined impact assessments to measuring the level of achievement of objectives (expected positive outcomes). Some of those agencies were very specific in describing the kind of impact an evaluator should search for. The document with the most detailed description of planned impacts divided them into two categories: tangible and intangible. Tangible impacts included: (i) standard of living (basic needs, jobs & income, and assets); (ii) organizational capacity (management, implementation, and resources); and (iii) policy environment (laws, policies, and dissemination & replication). As intangible impacts they included: (i) personal capability (knowledge, skills, and attitudes & values); (ii) organizational culture (vision, participatory practice, and organizational relationships/ alliances/partnerships); and (iii) community norms (values, practice, and relations).

Of the 13 agencies mentioning impact, five indicated that the evaluations should also address the question of attribution. It was not acceptable to indicate that an impact had occurred; it was necessary to establish clear links between the observed changes and the specific actions developed by the efforts supported by the agencies. Three of those agencies brought up the importance of basing the attribution process on ethical and honest judgments. They added that this process should not overwhelm the evaluation and risk its feasibility and utility, by transforming it in an academic research activity, unless, of course, there is a specific requirement by donors who will provide enough resources (financial, time, expertise) for accomplishing the task.

Process: Five agencies indicated in their documents that assessing process (e.g.

quality of interventions) is an important element in an evaluation. According to those agencies, evaluation of an intervention's process should include a systematic documentation of the systems, methods, tools, relationships/partnerships, and accountability to participants so improvements can be introduced to an on-going effort. One of the agencies indicated that process evaluation encompasses testing the hypothesis that the intervention led to goal achievement. Some key points for assessing an aid intervention process were not mentioned in any of the reviewed documents including: (i) ethicality—fairness and/or equity in implementing project activities, (ii) environmental responsibility—aid activities are being developed in a way to preserve and care for the environment; (iii) scientific soundness—alignment of implementation strategies with what is considered best practice in the field, (iv) fidelity—program is being implemented in the way it was promised to recipients.

Relevance: Of the 14 agencies, three mentioned the importance of relevance of the evaluand as one of the aspects to be considered in an evaluation. The definition of relevance for one of the agencies encompassed the assessment of the extent to which the evaluand's strategies are consonant with the agency's and country's overall development goals. The other two agencies connected relevance of the evaluand to its ability to address the needs and rights of the intervention impactees. Given the risk that organizational goals can be misaligned with the actual "real needs" of recipients then it seems reasonable to consider the latter definition for relevance is probably better than the former.

Effectiveness / Efficacy: Those terms were used interchangeably by the agencies to define the determination of the extent to which an evaluand achieved its objectives or goals. Eight documents mentioned effectiveness as an important aspect to be covered in assessing an evaluand. In six of those eight documents, effectiveness referred to meeting objectives set for specific projects or programs. The other two had a broader focus encompassing the strategic goals for the entire organization. The development and

measurement of indicators connected to the expected outcomes were mentioned at least by two of the eight agencies as the main way to determine goal achievement.

Efficiency: Only two agencies made reference to efficiency as an important criterion to assess an evaluand. Efficiency was defined by them as the best possible use of the resources available to achieve results. Its determination would include the assessment of the relationship of project benefits to inputs and their cost and timelines.

Cost: Even though cost was included as one of the components of efficiency by two INGOs, other two other agencies made reference to the evaluand's cost as a stand-alone relevant aspect to be considered in an evaluation. One of the agencies indicated that some projects may choose to conduct cost-benefit studies. However, they warned about the complexity of such studies and argued that they should only be undertaken when the necessary technical assistance and resources are available. No mention was made to a more comprehensive cost analysis that would include key aspects such as monetary and non-monetary costs, direct and indirect costs, and actual and opportunity costs (Scriven 2007, p. 13).

Sustainability: The last standard included in this category of standards for evaluands was sustainability. It was mentioned by four agencies and defined as the likelihood that the positive impacts produced by the aid interventions will be maintained after the original funds run out. The focus of sustainability seemed to be more on economic aspects of the intervention, i.e., the extent to which the program will have the necessary resources to maintain its main activities after the original funding ends. It fails to cover, however, other essential elements of sustainability such as political support, socio-cultural adequacy, technological appropriateness, and institutional capacity.

Standards related to the quality of evaluation processes and products

Participatory: All but three of the reviewed documents referred to participation of

all key stakeholders (recipients, partners, donors) as essential to ensure the quality of an evaluation. Four of those agencies went as far as indicating that such participation should include evaluation design, implementation and analysis. One agency seemed to be more conscious about this issue. This agency indicated that the degree of participation by stakeholders in evaluations should vary depending on the purpose of each evaluation. For instance, greater participation should be expected during mid-term evaluations or annual reviews, while a more objective (external) perspective is desirable in final evaluations.

Transparency: Only two agencies indicated in their documents the importance of the evaluations being as transparent as possible. Both supported the idea of sharing evaluation findings and recommendations with key stakeholders. One of those agencies, however, went further with a more radical conception of transparency by supporting public advertisement of the terms of reference for evaluation and competitive bidding as well as placing all final evaluation reports in the public domain.

No unnecessary disruption: One agency indicated that it is important that local programs are not taxed or eclipsed by evaluation needs. The same agency provided a general estimation of the proportion of a program's budget that should be dedicated to monitoring and evaluation functions: 5 to 10 percent (however, this estimate will vary according to the sophistication of the design, the amount of instrument testing involved and the degree that international consultants are used).

Accuracy: The importance of rigor in the evaluation methods adopted, the quality of the information gathered, and the validity of the conclusions reached were mentioned by seven agencies and were grouped under the standard defined as "accuracy". Four of those agencies mentioned rigorous evaluation methods as important aspects to ensure the quality of an evaluation. They did not provide, however, much detail about what they meant by rigor, except to say that the evaluation approach should be selected based on the evaluation questions posed and the available resources. Three of them indicated that evaluations should, whenever possible, conduct baseline studies and use comparison

groups. The latter can be identified through randomized selection of participants, identification of reference groups, or use of relevant secondary data, such as comparable statistics from the general population. The quality of data collection instruments was also mentioned as an important aspect under accuracy. Specifically, one agency indicated that the data collection instruments should be (i) culturally sensitive and appropriate to respondents, (ii) pilot tested before use, and (iii) participants should be consulted on the best strategies to collect the data.

In terms of comprehensiveness of methods, two agencies stressed that both quantitative and qualitative data should be sought in any evaluation. To enhance accuracy of an evaluation, another agency indicated the need for (i) critical reviews of the design and conduct of the evaluation by members of the evaluation team and external consultants, and (ii) conduct of analyses by the evaluators of information from different perspectives and using different methods. Another agency indicated that they promote the adoption of 'accuracy' as defined by the Joint Committee's, even though in the document reviewed there is just one short mention about the overall definition of the attribute and nothing specific on the 12 standards included under accuracy by the Joint Committee (1994, p. 125).

Thorough evaluation report: The last standard identified in this group relates to the quality of the evaluation reports, mentioned in seven of the 14 reviewed documents. All provided guidance on what should be included in an evaluation report. One main aspect stressed by six agencies was the importance for reports to include recommendations and lessons learned. One agency indicated the importance of independence of the report— program managers should not interfere with the evaluation reports. That agency also defended that reports should include stakeholders' response to the evaluation findings and conclusions. Except for that one agency, none of the others appeared to tackle the issue of creating mechanisms to ensure impartiality and fairness of reports, usually threatened by personal feelings and biases of any party to the evaluation.

Standards related to the evaluators

Team diversity: One agency supported the idea that an evaluation team, to be credible, should have gender balance and geographic diversity, i.e., including people from the countries or regions where the evaluation takes place.

Ethicality: Observing ethical principles was mentioned by five agencies as an important standard to be followed by evaluators. Descriptive aspects included in the reviewed documents included the need for evaluators to be sensitive to beliefs, manners, and customs of people involved in the evaluation process. Two agencies were very specific in terms of the need for careful handling of data collection, analyses and dissemination. They argue that no data collected should be purposefully presented in a deceptive or inaccurate manner. One agency goes as far as indicating that any attempt to falsify data should be considered a fair reason to terminate an employee or grant funding.

Competence: Three agencies have indicated in their documents that competence is another relevant standard to ensure the quality of an evaluator. Two of them just made a general remark saying that evaluators should possess the required qualifications for the evaluation job. One agency was specific in indicating the areas they expect an evaluator should be well versed in: management, planning, monitoring, finance, strategic/global thinking, problem solving, team work, communication, writing, negotiation, and technical knowledge of several evaluation methods.

Systematic Inquiry: Only one agency made a general reference to this standard and connected it to the Guiding Principles for Evaluators (GP) proposed by AEA (2004). The agency, however, did not include three specific aspects included in the GP: (i) “adhere to the highest technical standards appropriate to the methods they use”. (ii) “explore with the client the shortcomings and strengths of evaluation questions and approaches”, and (iii) detailed communication of approaches, methods, and limitations so others can assess the work done (AEA 2004).

Respect for people: Four agencies indicated the importance of respecting the security and dignity of people affected by the evaluation as a way to assess the work of an evaluator. The main aspects describing this standard included (i) protection of the anonymity and confidentiality of recipients, staff and other individuals included in the evaluation process, (ii) respectful contacts with all individuals, avoiding offending the dignity and self-respect of those persons with whom evaluators come in contact—in cases where data collection is of a sensitive nature, procedures for informed consent and data security in research with human beings must be respected.

Responsibility for general and public welfare: This standard was only mentioned by one agency which indicated they follow the Guiding Principles for Evaluators. No further details on the contents of the standard were provided.

Standards related to commissioners of evaluations

Use of findings: One-half of the 14 agencies mentioned in their documents the importance of evaluation findings being used by the intended users of the evaluation. To get used, according to four agencies, evaluations should be designed to answer pressing needs of key stakeholders, especially managers. Three agencies indicated that having evaluation findings communicated in an appropriate language and format to different stakeholders is essential to ensure their use. Having scheduled internal meetings to share lessons learned from the evaluations to inform current and future programming were seen by three agencies also as important for ensuring evaluation use. Three other agencies call for evaluations to make recommendations with clear direction for future action. One of the agencies indicated that to ensure use, evaluation recommendations should be agreed upon by stakeholders, and not written up until later on by the evaluators[56]. Two of them

[56] This is certainly a debatable statement. Even though it is reasonable to assume that stakeholders will be more likely to act upon things they have agreed on, this position has the potential to inhibit the development of creative/ingenious ideas to solve issues that could arise from external perspectives, unfiltered by the people with an active stake in the matter addressed by the evaluand. On the other hand, if

also indicated that such recommendations should be followed-up by implementation plans to be overseen by relevant supervisors.

Ensuring appropriate resources: Evidence that agencies consider viability issues as an essential part of any evaluation was presented in four of the 14 reviewed documents. Ensuring the necessary resources (financial and technical capacity) to implement the scoped evaluation was the main factor mentioned by the agencies.

M&E standards from other agencies adopted by INGOs

Of the 50 respondents of Chianca's survey, 19 (38 percent) indicated that their agencies have adopted, to different degree, M&E policies, guidelines or standards developed by other organizations. As mentioned in Chapter III, USAID seems to be the organization with greatest influence over U.S.-based INGOs on M&E issues—11 respondents said that their agencies need to adopt specific guidance when monitoring and evaluating projects funded by that agency—yet, as discussed on p. 69, USAID does not have clear or consistent standards for evaluation.

The other organizations from which survey respondents indicated they have adopted M&E standards include (i) Building bridges in Planning, Monitoring and Evaluation, (ii) Better Business Bureau Wise Giving Alliance, (iii) Hope for African Children Initiative, and (iv) Focus on Young Adults. The following sections will analyze the evaluation standards proposed by those four agencies. As a side note, two INGOs indicated that their evaluation policies have been influenced by other agencies such as USAID, CARE International, AEA, OCHA, OECD/DAC, DANIDA, SIDA, CIDA, UNICEF, UNDP, and DFID.

---

there is not buy-in on the part of key stakeholders, it is not likely that the recommendations will be accepted and acted upon. This calls for a process that includes external evaluators proposing recommendations yet through a process that includes the perspectives of stakeholders.

Better Business Bureau

The Better Business Bureau Wise Giving Alliance assesses the performance of U.S.-based charities, including many of the INGOs that are members of InterAction. They have adopted 20 accountability standards to orient their evaluations: five are related to governance, two to measuring effectiveness, seven to finances, and six to fundraising and informational materials. The most relevant to program evaluation are the ones dealing with measuring the organization's effectiveness in achieving its mission. These two standards require that organizations have "defined, measurable goals and objectives in place and a defined process in place to evaluate the success and impact of its program(s) in fulfilling the goals and objectives of the organization and that also identifies ways to address any deficiencies" (BBB 2003). Certainly, a positive assessment from the BBB Wise Giving Alliance is a strong indication that an INGO is being responsible and effective in use of its resources. It might also indicate whether the agency is achieving its goals. However, the BBB standards do not provide minimally comprehensive guidelines for sound evaluations.

Hope for African Children Initiative (HACI)

HACI is a partnership involving six agencies, including five U.S.-based INGO members of InterAction (CARE International, Plan International, Save the Children, World Conference on Religions for Peace, and World Vision). The main objective for the partnership is to provide support to communities across Africa so they can offer prevention, care and support to orphans and children affected by HIV/AIDS. HACI developed a document in 2003 describing its monitoring and evaluation framework, with the support from the six agencies comprising the partnership and also the Bill and Melinda Gates Foundation. The 42-page document explains in detail how the M&E function for the whole program should be built around the program's conceptual model

(Circle of Hope) and the specific attributions for each of the four components of the M&E framework: (i) HACI core objective (across countries), (ii) country-specific M&E, (iii) community defined M&E (to assess the work by the implementing agencies), and (iv) operations research (testing of new approaches to support HIV/AIDS affected children by professionals internal to the program and by research organizations).

The program has four core objectives for which specific indicators have been developed to be measured overtime and provide grounds for assessing the program. The aspects proposed to be covered by the evaluations are: (i) implementation process (output and coverage: e.g., number, quality and distribution of services), (ii) outcomes/impacts (e.g., changes in behavior, livelihood security, etc) including positive and negative outcomes and unintended results, (iii) effectiveness (reaching intended outcomes), (iv) cost-effectiveness, and (v) sustainability (of benefits and services).

As far as standards for ensuring the quality of evaluations, the document clearly stresses the importance of (i) participation of beneficiaries, (ii) sound methods (use mixed methods, conduct baseline studies, develop country specific tools for country-specific interventions, and make sure secondary data are trustworthy), and (iii) good reports (quarterly and final) with specific recommendations to improve the program. No specific standard for ensuring the quality of evaluators was presented.

Building bridges in planning, monitoring and evaluation

The reviewed document comprises guidelines for good practice in planning, monitoring and evaluation (PME) of community-based development projects implemented by NGOs in developing or transitional countries with support from European ecumenical agencies (ICCO 2000). This 100-page publication is the result of a collective process spanning 1996-99, involving five ecumenical funding agencies and nine Southern development organizations from Latin American, Africa, the Middle East and Asia. The Logical Framework (Logframe) is the approach adopted throughout the

Building Bridges publication as the basis to orient all three functions—planning, monitoring and evaluation. The publication has quite clear and comprehensive guidance on how to plan a community-based intervention and on how to design and implement a monitoring system that can generate useful data for future evaluations.

Even thought the Building Bridges does not explicitly mention principles or standards for the design and conduct of evaluations, it is possible to infer them from the descriptions presented on how to devise, manage and implement a PME system. In terms of aspects that evaluations of projects should consider, three were raised: (i) effectiveness (expected effects of the program in relation to its objectives), (ii) impact (in relation to the main goals and also including unintended outcomes), and (iii) reach (the extent to which the project reaches the intended beneficiaries and actually produces the desired benefits). Two standards connected to the quality of the evaluation process and products were included in the document. First, they mention organizational provision which refers to specifying responsibilities, procedures, timing, and budget. The second aspect mentioned was quality of data which included ensuring usability, completeness, reliability and validity of data.

FOCUS on young adults

FOCUS is a USAID funded program developed by Pathfinder International (a large U.S.-based INGO) in partnership with The Futures Group International and Tulane University School of Public Health and Tropical Medicine. The program claims to promote the well-being and reproductive health of youth. The publication analyzed in this study is titled “A Guide to Monitoring and Evaluating Adolescent Reproductive Health Programs” and was written by five authors—one from FOCUS, three from Tulane University, and one independent consultant. The more than 450 pages of the document are an extensive and useful resource for establishing M&E systems and to design and conduct epidemiologic research on youth reproductive health programs. The guide, as

many of the other documents we reviewed, does not have a specific section with proposed evaluation standards or principles. When describing the aspects to be assessed in an evaluation of a program they have mentioned several:

- Meeting needs: assessing whether the project strategy is addressing the community's needs.
- Adequacy of resources: assessing whether the necessary resources needed to carry out the program activities are available.
- Quality of program implementation: assessing whether activities developed or services provided are adequate to implement the strategy.
- Program cost-efficiency (no definition for the term was provided)
- Program coverage: proportion of the population with needs that are being positively affected by the program
- Program use: the extent to which a program's services are being used by the intended target population.
- Level of achievement of program objectives
- Program outcome: determining whether outcomes that the program is trying to influence are changing in the target population.
- Program impact: assessing how much of the observed change in outcomes is due to the program's efforts; impact evaluations target long-term outcomes.

No standards connected to the quality of evaluation processes and products or to the behavior and competence of evaluators seem to have been discussed by the authors. Table 6 summarizes the many standards proposed by those four agencies.

Table 6. Summary of standards from other agencies adopted by INGOs

Group of standards	Standards	Agencies BBB	HOPE	I CCO	FOCUS
for Evaluands	Impact/ Outcomes	X	X	X	X
	Effectiveness	X	X	X	X
	Efficiency	.	X	.	X
	Sustainability	.	X	.	.
	Coverage/ Reach	.	X	X	X
	Meeting needs of participants	.	.	.	X
	Quality of program implementation	.	.	.	X
	Use of program services	.	.	.	X
for Evaluation processes and products	Participation	.	X	.	.
	Transparency	.	.	X	.
	Utility	.	.	X	.
	Feasibility	.	.	X	.
	Accuracy	.	X	X	.
	Good reports	.	X	X	.
	Flexibility	.	.	X	.

The InterAction evaluation standards

Since 1994, all InterAction members have had to comply with a set of ethical standards covering governance, financial reporting, fundraising, public relations, management practice, human resources, public policy, and program services (InterAction 2007a). All agencies are required to go through an annual self-assessment to provide evidence to InterAction that they are complying with the standards. Among the more than 150 standards (and sub-standards), there are only ten that make reference to evaluation (InterAction 2007b). Six of those ten standards related to the importance of including the perspectives of gender equity, promotion of diversity, and disability inclusion in the program cycle from design to evaluation[57]. Two of those standards are specific to agencies working with child sponsorship programs. They require that such agencies should have an evaluation policy and also openly communicate to child sponsors what

[57] Standards: 6.4.1.3, 6.4.1.6, 6.4.2.3, 7.2.2, 7.3.2, and 7.4.2.

indicators are being used to evaluate the benefits sponsored children are receiving[58]. The other two standards are more general and refer to the quality of evaluation process of aid interventions:

> 7.1.2 Participants from all groups affected should, to the maximum extent possible, be responsible for the design, implementation, and evaluation of projects and programs. (p. 10)
>
> 7.1.9 A member shall have defined procedures for evaluating, both qualitatively and quantitatively, its programs and projects. These procedures shall address both the efficiency of the use of inputs, and the effectiveness of the outputs, i.e. the impacts on the program participants and the relationship of these impacts to the cost of achieving them. (p. 11)

In September 2005, a position statement on demonstrating NGO effectiveness was approved by the InterAction Board with clear relevance to evaluation practice and principles of InterAction member agencies (EPEWG 2005). The statement was the result of efforts of the Evaluation and Program Effectiveness Working Group (EPEWG) formed by 19 representatives of member agencies and 5 InterAction staff and consultants. The statement indicated that all InterAction members commit to take five actions so they will be able to demonstrate the effectiveness of their work to themselves, their stakeholders and the general public. The actions are:

1. Articulate criteria for success in bringing about meaningful changes in terms of its mission and major program goals.
2. Regularly evaluate progress towards such success.
3. Mainstream monitoring and evaluation in policy, systems and culture.
4. Allocate adequate financial and human resources for strategic evaluation needs.

[58] Standards: 7.11.12 and 7.11.14.

5. Collaborate with partners and stakeholders in developing mutually satisfying goals, methods, and indicators for project and program activities (p. 6).

In order to make the statement concrete to all members, InterAction's EPEWG subsequently assumed the responsibility to revise and add to the current InterAction Standards related to M&E so they would be coherent with and add specificity to the five principles adopted in the NGO Effectiveness statement. A subcommittee comprising InterAction members, staff and consultants was put together to tackle this task. The work of the subcommittee spread through several months until they were able to develop a proposal for a new set M&E standards by the end of the fall 2006.

Before submitting their proposal to InterAction's standards committee, the subcommittee wanted to make a larger consultation among InterAction members about their perceptions of the standards and guidelines they had produced. Specifically they were hoping to: (i) obtain suggestions to improve the standards, (ii) gather ideas on evidence of compliance with standards members would be able to provide, and (iii) identify areas members would like to receive technical assistance. Given Chianca's research interest on issues related to evaluation principles and practice in INGOs, the subcommittee decided to invite him to help with the consultation process. He was charged with designing and implementing a survey with a sample of INGOs (InterAction members) that could shed light on the important questions raised by the subcommittee.

Box 1 presents the set of M&E standards that were sent to InterAction members as part of Chianca's survey.

Box 1. InterAction M&E standards and guidelines included in Chianca's survey.

Standard 2.6. (section on 'Governance'):

"*The board shall ensure that the organization (i) articulates organization-wide criteria for success in meeting the needs of intended beneficiaries in terms of its mission and major program goals; (ii) regularly commissions valid and credible evaluations of the organization's efforts towards such success; (iii) mainstreams and utilizes monitoring and evaluation in the agency's policy, systems and culture; and (iv) allocates adequate financial and human resources for the organization's strategic evaluation needs.*"

Box 1. InterAction new monitoring and evaluation standards and guidelines (cont.)

- Interpretive Guidance associated with standard 2.6..:

"The term regularly *means a pre-determined interval within the organization's strategic planning cycle."*

Standard 4.3. (section on 'Management Practice'):

"To inform its ongoing strategic planning process, a member organization shall incorporate a deliberate and intentional process of monitoring and evaluating the organization's progress toward achievement of its mission and major program goals."

- Interpretive Guidance associated with standard 4.3.:

"- Each agency should have one or more explicit underlying hypothesis(es) or theory(ies) of change about how its activities will lead to desired changes. In other words, it should be able to articulate clear causal links between major program activities, impacts and mission.
- The agency should ensure that valid and credible evaluations of its operations are conducted in accordance with the agency's strategic planning cycle. Such evaluations should be a complete assessment of the quality, value, and significance of the work done by the agency, always including an assessment of the progress made by the agency in achieving its mission and major goals."

Standard 4.4. (section on 'Management Practice'):

"A member organization shall mainstream and utilize monitoring and evaluation in agency policy, systems and culture in terms of the organization-wide criteria for success in bringing about meaningful changes in people's lives, and shall allocate adequate financial and human resources for the organization's planning, evaluation, and institutional learning needs."

- Interpretive Guidance associated with standard 4.4.:

"At both strategic program and project levels, evidence of progress and impacts should be captured through a valid and credible monitoring and evaluation system. While InterAction is not prescribing a common approach to be followed, such a system should provide systematic information about the following key aspects of programs and projects implemented by IA members:

- *Positive changes, e.g. type and scope of benefits, whether material, human/social, organizational, civic, policy, governance, environmental, or other. Evidence of participants' satisfaction with such changes should be included.*
- *Reach, e.g. number of people, communities, organizations, regions, etc.; number of partnerships & alliances; and depth of poverty or marginalization of target populations.*
- *Efficiency of delivery, e.g. timeframe for implementation; costs (monetary and non-monetary—e.g., opportunity, stress, time), compared to results obtained.*
- *Resources for sustainability, e.g. structural changes, commitment by participants to continue activities or benefits, new resources, external stakeholder support, enabling policy environment.*
- *Post-project gains, e.g. replication, expansion, policy change, etc.*
- *Side effects, e.g., documentation of positive and negative unintended outcomes/ impacts connected with the efforts.*
- *Ethical practice, e.g., evidence that the means to produce the results/impacts adhere to relevant ethical standards*
- *Pertinence to needs, e.g., the extent to which the initiative's objectives and implemented strategies are directly connected to existing needs of targeted beneficiaries."*

Standard 4.5. (section on 'Management Practice'):

"An agency's planning, monitoring and evaluation system should draw on commonly accepted professional principles and standards in planning, monitoring and evaluating programs. These systems should take into account not only the defined organization-wide criteria for success toward achievement of its mission and program goals, but also basic components of sound evaluations including, but not limited to, quality of process, intended and unintended outcomes and impact, costs, and sustainability."

Box 1. InterAction new monitoring and evaluation standards and guidelines (cont.)

- Interpretive Guidance associated with standard 4.5.:

" - InterAction does not prescribe particular evaluation methodologies, but it does call upon its members to be aware of the range of methods and approaches, and make informed choices as to which are most appropriate for the various projects and programs they implement.

- See, for example: American Evaluation Association (www.eval.org); the African Evaluation Association (www.afrea.org); ALNAP (www.alnap.org/), Action Aid International's Accountability, Learning and Planning System (ALPS) (www.actionaid.org/index.asp?page_id=472); and Design, Monitoring and Evaluation policies and standards of CARE as well as those of other INGOs (http://www.globaldev.org/m&e/). *These and other relevant links will be provided on the InterAction website."*

Standard 4.6. (section on 'Management Practice'):

"From the outset of program planning, a member organization shall collaborate with partners, clients/intended beneficiaries and other stakeholders in developing mutually satisfying goals, methods, and indicators for project and program activities and results."

- Interpretive Guidance associated with standard 4.6.:

"InterAction members' program theory(ies) of change should involve active participation by communities or other constituencies, and should have clear policies and procedures to engage the active participation of communities and partners in program design, planning, monitoring, evaluation and learning. All InterAction member agencies should regularly assess the satisfaction of those they seek to serve."

Other standards relevant to M&E:

Standard 6.1. (section on 'Communications to the U.S. Public'):

"The member organization shall be committed to full, honest, and accurate disclosure of relevant information concerning its goals, including criteria for objectively measuring progress and success of its programs, finances, and governance in achieving the goals."
Note: Though this standard is within the "Communications to the U.S. Public" section, it points to one of the purposes for an agency's M&E system – to measure and share with its public progress and success of its programs in achieving goals."

Standard 8.1.2. (section on 'Program'):

***"**Participants from all groups affected should, to the maximum extent possible, share responsibilities for the design, implementation, and evaluation of projects and programs."*

Standard 8.1.9. (section on 'Program'):

"A member shall have defined procedures for evaluating, both qualitatively and quantitatively, its programs and projects consistent with the ideas presented in Standards 4.3 to 4.6, and in the Interpretive Guidance section for those Standards."

- Interpretive Guidance associated with standard 8.1.9.:

"Evidence of change needs to be guided by prevailing norms within sectors or sub-field(s) of development or relief. Several sub-fields of development practice have recently defined common standards, e.g., child sponsorship, humanitarian assistance and microfinance (see also interpretive guidance for 4.4 and 4.6)."

Using the analytical framework discussed earlier (p. 86), it is possible to classify the eight standards and guidelines under the four categories of standards: (i) evaluand: 4.5 (and guidelines 4.4); (ii) evaluation processes and products: 4.6; (iii) evaluators: (no

standards were found to pertain directly to this category); and (iv) commissioners or intended user of evaluations: 2.6, 4.3, 4.4, 6.1, 8.1.2 and 8.1.9.

The following chapter will discuss methodological aspects of the survey conducted by Chianca and present his findings based on the feedback provided by representatives of the INGOs that will be potentially affected by the new InterAction M&E standards. It is important to note that after a presentation by Chianca of survey findings to InterAction's EPEWG in August 2007, major changes were made to the proposed standards. Appendix G has the final version of the M&E standards submitted to InterAction's standards committee in October 2007. The last version of the standards will be discussed in Chapter VI of this dissertation.

CHAPTER V

A SURVEY ABOUT THE 2006 VERSION OF THE INTERACTION EVALUATION STANDARDS

The revision of the InterAction evaluation standards included a survey developed by Chianca targeting CEOs and staff responsible for monitoring and evaluation functions in all 167 INGOs identified as members of InterAction as of February, 2007. The survey was designed to serve many objectives. First, it intended to gather ideas from InterAction members (the primary impactees of the new standards) to further improve the initial version of the standards. Second, it aimed at identifying examples of possible evidence members could provide that would indicate their compliance with the standards. Third, the survey planned to collect some basic information about key aspects related to evaluation principles and practice adopted by those agencies to help InterAction identify possible areas to provide support to the agencies. Finally, it had an educational purpose: introduce InterAction members to the new standards.

In this chapter we will discuss the survey's main findings and the specific suggestions to improve the InterAction standards. Before addressing the findings, we will discuss the main methodological features pertinent to the survey.

Survey methodology

The survey instrument and invitation letters to CEOs and staff responsible for M&E were designed primarily by Chianca and Rugh[59]. Suggestions for improving those

[59] Jim Rugh, independent consultant and former Coordinator of Design Monitoring and Evaluation for CARE International and leader of InterAction's EPEWG.

documents were offered by Scriven[60], Clements[61], Giunta[62], Levine[63], Steinke[64], and Wiebe[65].

A letter (Appendix D) signed by InterAction's CEO, Sam Worthington, was used to invite CEOs from member agencies to participate in the study. Potential respondents were offered three options to complete the survey: (i) online, through a web-based instrument; (ii) using an electronic file (MS-Word format); or (iii) participating in a phone interview.

The final version[66] of the survey instrument, presented in Appendix E, had 26 questions: 12 close-ended and 14 open-ended. Questions 1 to 4 were identifiers of the agencies responding to the survey. Questions 5 to 13 inquired respondents about the proposed new InterAction's M&E standards and guidelines in terms of relevance, clarity, examples of compliance, and need for technical assistance. Questions 14 to 25 related to the respondents agencies' M&E structures, policies and practice. Finally, question 26 asked for additional general comments or suggestions to improve the survey instrument.

After approval by the WMU Human Subjects Institutional Review Board[67] on February 9, 2007 (Appendix F), the survey was sent to 41 organizations as a pilot test to assess whether the instrument could work on a real-life setting. E-mail invitations on behalf of InterAction's CEO were sent by Chianca on Feb 13, 2007 to 11 evaluation

---

[60] Michael Scriven, Professor, Claremont Graduate University, chair of Chianca's dissertation committee
[61] Paul Clements, Professor, Western Michigan University, member of Chianca's dissertation committee
[62] Ken Giunta, consultant, former InterAction's Director of Membership and Standards
[63] Carlisle Levine, Senior Technical Advisor, Monitoring and Evaluation, Program Quality Support Department, Catholic Relief Services
[64] Megan Steinke, Planning, Monitoring and Evaluation Specialist, Save the Children
[65] Menno Wiebe, independent consultant working with Church World Service
[66] The survey instrument had 63 questions, including items for each one of the new InterAction M&E standards and guidelines. It was a very comprehensive instrument, however, the time required to complete the survey (between 60 and 90 minutes) was considered too long by the researchers and a therefore a threat to achieving a reasonably good response rate from the agencies.
[67] The study's approval process by HSIRB took more than 2 months (Dec 4, 2006 to Feb 9, 2007)—an unusual amount of time considering the nature of the study. The primary reviewer interpreted the initial versions of the study protocol as having the potential to cause harm to interviewees. It took several conversations to finally convince the reviewer that the study was focused on studying organizations and not individuals.

specialists, members of InterAction's Evaluation Interest Group (EIG). Giunta sent invitations on Feb 20, 2007 to 30 randomly selected CEOs of InterAction member agencies. Two survey reminders (approximately 15 days and one month after the initial invitation) were sent by email to the selected respondents and Giunta made one follow-up phone call to most of the invited CEOs. The response rate for this initial sample was 39 percent (16 of 41). Among the respondents were all 11 members of the EIG group and only five CEOs (or a designated representative). Respondents indicated taking between 15 and 45 minutes to complete the survey, including reading a 2-page handout with InterAction's new M&E standards and guidelines.

Based on preliminary analysis of the 16 responses, the study leader (Chianca) and his committee members (Scriven, Rugh, and Clements) felt that the initial survey instrument was adequate to fulfill the study purposes. Therefore, no major changes were made, except for minor editing and formatting, before sending it to all InterAction members. On May 8, 2007, survey invitations were sent directly from InterAction's central office to the CEOs of the remaining 126 agencies that had not been contacted by the researchers in the first round. The survey was also sent to representatives of the 25 agencies who had not responded to the initial invitation. One survey reminder was sent by Giunta on June 11, 2007 to the CEOs of all agencies that had not responded to the survey by that date. No other survey reminders from InterAction's central office were sent to potential respondents given time constraints.

As an extra effort to increase response rate, two survey reminders (May 4 and June 11, 2007) were sent by Chianca to 27 INGO representatives who participated in a session on M&E facilitated by Rugh and Chianca on April 18, 2007 at the InterAction 2007 Forum in Washington, DC. As it turned out, three of those agencies who responded to the survey were not official members of InterAction. However, they were also U.S.-based nonprofit organizations supporting development and relief projects overseas, or advocating for the poor and most vulnerable people around the globe. The project lead,

Chianca, decided their responses were relevant enough to be included in the analyses of the survey responses.

Out of the 170 agencies invited to participate in the study (from Feb 13 to July 11, 2007), 50 answered the survey (30 percent response rate). Twenty eight agencies completed the survey online, 21 sent their completed survey (MS-Word document) as an attachment of an email message, and one INGO representative was interviewed over the phone by Chianca.

Four agencies provided double responses to the survey. In two cases, the responses came from different people, while in the other two the same individual responded to the survey. These cases might be explained by the fact that the organizations received the survey more than once and different people decided independently to respond to it or the same individual may have assumed that the initial survey was only in "test mode" and that they needed to now respond to the "official" survey. Since the duplicate answers did not contain contradictory information, the double responses were carefully integrated into a single entry for each organization in the survey responses database.

The survey also asked for documents to support agencies' responses to some of the survey questions, such as M&E guidelines and strategic plans. The respondents either sent those documents to Chianca as an attachment to an email message or sent the links for Web pages where those documents could be downloaded.

Several email and phone contacts were established by Chianca with survey respondents between April and July, 2007, to ask follow-up/clarification questions about their answers to the survey. The most common follow-up questions related to apparently contradictory answers; for instance, two agencies indicated that they did not have a regional structure for M&E, but indicated to have regional M&E staff. Another frequent follow-up question to respondents related to the documentation describing M&E standards, guidelines or principles developed by their agencies or by other agencies they

had decided to adopt. A number of agencies had not provided Chianca with such documents and were asked to do so. After gathering the needed information or clarifying pending questions, Chianca updated the survey database with the new information, maintaining the original submitted responses without changes.

Additional information was collected about the 170 agencies (167 members and 3 non-members of InterAction) included in the study. They included agencies' annual expenses, primary foci (development, relief, advocacy and/or technical assistance[68]), number of countries reached by their efforts, time since foundation, and assessment by Charity Navigator[69]. Primary sources for those additional data included INGOs' websites and annual reports, InterAction's website, InterAction Member Profiles 2004-05, and the Charity Navigator website.

All data from survey responses and complementary search were entered in a MS-Excel database. Only Chianca had access to the database which was kept on a password protected folder on his computer. Data analyses included mostly descriptive statistics of numeric data and content analysis of open ended responses and documents received.

Differences between survey respondents and non-respondents

The 50 INGOs whose representatives responded to the survey differ from the other 120 who did not participate in the survey in four main ways:

1. They have larger annual expenses: The median annual expense of the agencies responding to the survey was 31.58 million dollars while the median annual expense of the non-respondents was 12.11 million dollars.
2. They focus their work more on development and less on advocacy and

[68] Some agencies included in this study provide expertise and advice to local NGOs.
[69] Charity Navigator is an independent nonprofit agency which specializes in rating the quality of US-based charitable organizations derived from their financial health status. Their technicians review publicly available financial documents of charities to determine how responsibly those organizations function and the conditions they have to sustain their efforts/programs over time.

technical assistance: 80 percent of the respondent agencies focus their work on development while this proportion among non-respondent agencies is 62 percent; when the comparison is in terms of their work on technical assistance, the situation gets reversed: 14 percent of respondent and 33 percent of non-respondent agencies include technical assistance as part of their work.

3. They cover a larger number of countries: Respondent agencies support projects in 46 countries on average, while non-respondent agencies support projects in an average of 28 countries.
4. They are slightly older on average: The 50 agencies whose representatives responded to the survey have been operating for an average of approximately 42 years while non-respondents have been around for an average of about 38 years.

Table 7 presents in detail the most relevant contrasts between those two groups formed by the agencies that responded to the survey and the ones that did not.

Table 7. Descriptive information for survey respondents and non-respondents

Variables	Respondents (n=50)	Non-Respondents (n=120)	All (n=170)
Annual expenses (in US$ million)			
Mean	181.63	45.10	88.57
Median	31.58	12.11	16.50
Range (min.)	0.97	0.09	0.09
Range (max.)	3,396.79	587.17	3,396.79
sum	8,899.66	5,051.68	13,937.18
Size (based on annual expenses)			
Small: less than $10 million	12 (19%)	53 (81%)	65 (100%)
Medium: $10 million to $49.9 million	23 (40%)	35 (60%)	58 (100%)
Large: $50 million to $99.9 million	14 (50%)	14 (50%)	28 (100%)
Very Large: $100 million or more	9 (47%)	11 (53%)	19 (100%)
Proportion of agencies working on			
Development	80%	62%	67%
Relief	52%	52%	52%
Advocacy	24%	31%	29%
Technical assistance	14%	33%	27%

Table 7 – Continued

Variables	Respondents (n=50)	Non-Respondents (n=120)	All (n=170)
Mean number of countries reached	46	28	34
Mean age of agency (in years)	41.6	37.6	38.9
Proportion accessed by Charity Navigator	56%	57%	57%
Number of stars by Charity Navigator (max: 4 stars)	3.4	3.1	3.2

Sources: INGOs' websites and annual reports, InterAction's website, InterAction's 'Member Profiles 2004-05', and Charity Navigator's website.

There were some limitations in constructing this table, especially in regards to the financial information. We could not find financial data for nine agencies (one respondent and eight non-respondents). Expenses for 16 agencies (four respondents and 12 non-respondents) were based on data older than three years (range: 2002 to 1998), probably causing underestimation of current expenses. Annual expenses for two agencies (one respondent and one non-respondent) were estimated from graphs included in their annual report, since no specific numbers were provided. Therefore there is no precision of the estimations at the thousand dollar level. Due to lack of information, annual expenses figure for one non-respondent agency reflects annual revenue and not expenses, which likely caused an overestimation of its actual expenses.

Finally, two agencies (American Red Cross and Americares) account for slightly more than one-half of the annual expenses of the group of agencies responding to the survey—about 4.47 billions of U.S. dollars. This fact, however, does not invalidate the initial conclusion that the agencies responding to the survey are larger in terms of annual expenses than the non-respondents. If we exclude these two agencies from the calculations, the difference in the mean annual expenses between the two groups will have a sharp drop of almost two-thirds but will still remain quite substantial (from $136.5 M to $49.1 M); while the difference between the medians will be reduced only by less than 14 percent (from $19.5 M to $16.8 M).

A fair question to ask is to what extent the differences between respondent and

non-respondent agencies will affect the possibility of generalizing the survey results. Since the 50 respondents are not a random selection from the InterAction membership, then it would be naïve to claim the results can be statistically generalizable to the whole group of InterAction members. With that said, it is important to recognize the relevance of such a study to the field. This is probably the largest research effort conducted so far to study the scene of evaluation principles and practice in U.S.-based INGOs. Even though unable to claim generalizability to the whole population of INGOs, the results from this carefully designed study involving 50 INGOs do provide an important insights into the current situation.

The known differences between respondents and non-respondents indicated in Table 7 can help us speculate about possible scenarios had we been able to get either 100 percent of response or an appropriate random selection of the InterAction members. The first aspect worth being considered is the difference in size of the organization (larger organizations responded to the survey more often than smaller organizations). Larger agencies probably have more resources available and greater internal and external pressure to have better structured monitoring and evaluation functions within their organizations. With a better structured sector, and, most likely, with more staff charged with M&E responsibilities, those organizations have probably already developed their own monitoring and evaluation standards, guidelines or policies. They might feel they have the capacity to follow the proposed InterAction standards and guidelines, and, therefore, might be less likely to indicate problems with them in the survey.

Differences between respondents and non-respondents agencies' focus may also affect the possibility for generalization of survey findings. The bulk of the InterAction member agencies seem to be involved with development (67 percent) and relief (52 percent) work. Therefore, there is a possibility that the new standards and guidelines have been constructed to address more those functions rather than advocacy and technical assistance which account, respectively, for 29 percent and 27 percent of the agencies'

efforts. If this is the case and since we had a larger proportion of agencies doing development work in the respondents' group, 80 percent vs. 62 percent (non-respondents), there is a possibility that we 'heard' fewer complaints about the new standards and guidelines than we would have received if we had a greater proportion of agencies working with advocacy and technical assistance in the non-respondents' group answering the survey.

It is not clear whether or not differences between the two groups in terms of the number of countries reached by the agencies and their age would affect in any way the level of generalizability of the survey findings, even though number of countries is probably a direct correlate to size of budget.

Relevance of the standards and guidelines

The first five questions of the survey explored the respondents' opinions about the new InterAction M&E standards and guidelines. The first question of this series asked whether they felt there were any irrelevant standards or guidelines among the ones proposed by InterAction. As Table 8 shows, 80 percent of the respondents (40 of 50) found all standards and guidelines relevant to their agencies.

Table 8. Distribution of respondents indicating one or more M&E standards and/or guidelines irrelevant to people in their organizations

Response	Frequency	Percent
Yes, there are irrelevant standards or guidelines	9	18%
No, all standards or guidelines are relevant	40	80%
Blank	1	2%

Comments presented by a minority of the respondents who said 'Yes' suggested that the standards might not work for some of the agencies. The follow up, open-ended question asking respondents to explain their answer brought a diverse array of critiques

from those nine agencies. Some were more general, related to all or most of the standards and guidelines, while others were more specific, connected to individual standards. The following paragraphs describe those critiques.

A survey respondent from a large organization[70] primarily dedicated basically to development indicated they have about 80 percent or more of their programming funded by the U.S. Government. She thinks in general the new standards and guidelines will not be relevant to them since they have to follow the strict rules imposed by the government agencies in terms of design, implementation, monitoring and evaluation. The following are the specific comments she presented to individual standards:

> Standard 4.3: "… implies more self directed programming than we have. We do not have a strategic planning cycle."
>
> Standard 4.4: "Funding for monitoring and evaluation is under the control of our funder, again, largely the US government, with its own agenda. We would not accept money that goes against our goals and objectives, but we are not programming the money as if it were our own."
>
> Standard 4.5: "Each separate program is evaluated according to the funder's criteria, which may vary from program to program."
>
> Standard 4.6: "Program planning takes place with the sponsors of the program, though they may take into consideration the beneficiaries."
>
> Standard 8.1.2: "Design and implementation of programming is almost always in the hands of funders (US govt, primarily)."
>
> Standard 8.1.9: "We adhere to professional guidance on monitoring and evaluation, and have persons trained in those disciplines on our staff. The procedures vary according to the requirements of the funder."

Along the same lines, a representative from a medium-size organization dedicated to development work in commenting on standard 8.1.2, indicated that her agency does

[70] Size of the INGOs was based on their annual expenses (see Table 7 for details)

M&E work on a project by project basis. She also mentioned that her agency is supposed to follow "[s]pecific requirements for monitoring and evaluation … often provided in some detail by the funding organization."

A representative from another large agency with primary focus on development and advocacy was quite critical about the whole set of new M&E standards and guidelines proposed by InterAction. The main issues he raised include: (i) they are founded on "service-delivery" instead of "rights-based" approach to development; (ii) they are exclusively oriented to "projects" and do not encompass "program-based evaluation framework"; and (iii) they do not seem to address M&E issues related to advocacy and institutional change work.

One relatively small agency which provides only capacity building and technical assistance to other NGOs, echoed the critique raised by the previous respondent. She indicated that none of the standards and guidelines would be relevant to her agency since they do not directly operate any development or relief projects. She added however, that "one could make a case, broadly, that we could adopt some of the standards to evaluate our adherence to our mission and to see whether we meet objectives that lead other agencies to better meet their standards".

Respondents from two relatively small agencies that are dedicated to raising funds in the U.S. to support programs designed, implemented and assessed by United Nations' agencies thought many of the standards and guidelines were not applicable to their organizations. Both of them specifically quoted standard 4.4 and its guidelines as irrelevant relative to the work they do. One of them also considered standards 4.3, 4.5, and 4.6, and all the guidelines connected to them, irrelevant.

A representative from a medium-sized agency dedicated to international development and relief services stated that the standards and guidelines are useful only as general goals or suggestions for members to follow. This agency indicates that they are not planning to base their M&E system on 'theories of change' as suggested in the

guidelines. Instead, they claim to have developed agency's goals. Their M&E system is being built to stimulate better reporting from partners on project/program impact, so they can roll up those results to assess how well they are doing to meet those goals. A respondent from a medium-sized agency dedicated to development and relief indicated that they currently have limited M&E capacity. For that reason, it would be quite difficult to integrate the new standards and guidelines in their management operations, and, therefore, they would not be relevant.

Finally, in commenting on standards 6.1 and 8.1.2, one M&E staffer from a large organization dedicated to development work noted that some of his colleagues would not be "very keen to disclose financial information nor cutting edge/competitive advantage type program design."

One respondent, from one of the largest agencies in the group, offered a suggestion to enhance the standards. Ideally, he would like to see spelled out in the text a rubric with the different 'levels' of performance on the standards, e.g., gold, silver, bronze. He argued that agencies could take this up at a level that is appropriate to their circumstances instead of thinking that they are not large or sophisticated enough to comply with the standards.

Clarity of the standards and guidelines

The specific question in the survey asking respondents to assess the level of clarity of the standards was phrased: "Are there any standards or interpretive guidance that, though relevant, may be unclear to people in your organization?" As Table 9 shows, more than three-fourths of the respondents thought the standards and guidelines were clear. However, 11 (22 percent) indicated perceiving some kind of problem with the way they were phrased.

Table 9. Distribution of respondents indicating one or more M&E standards and/or guidelines as being unclear to people in their organizations

Response	Frequency	Percent
Yes, there unclear standards and/or guidelines	11	22%
No, all standards and guidelines are clear enough	38	76%
Blank	1	2%

In the follow-up open-ended question asking respondents to make specific comments about the clarity of the standards, the responses were quite varied as discussed below.

A general comment made by a respondent from one relatively small agency dedicated to advocacy was that the standards are not quite adequate to agencies that are not doing development or relief work, reinforcing major critique presented by other respondents in the previous question. This person generalized her critique to all InterAction standards in the Self-Certification Plus (SCP) guide, and not only to the ones related to M&E. Here is what she had to say:

> The SCP generally has been a little challenging for [*our agency*] as we are an organization that does not provide services; [*our agency*] is a research and advocacy organization. The questions and standards in the SCP are geared more toward service providers.

The need to define undoubtedly the scope of the term "M&E" was the other more general comment made about the clarity of the standards. The respondent making this comment—representing one of the largest INGOs in the US—indicated that it is unclear whether or not the term includes "processes of individual/organizational learning". All the other comments and suggestions were specific to a standard or guideline and are presented in Table 10.

Table 10. Specific critiques and suggestions to make the standards and guidelines clearer

Standard / Guideline	Critiques/Suggestions
2.6.2	Not clear on what you mean in comment #2.6.2 "Board attributions stated in governance documents"
4.3.1	Does InterAction expect that each agency has "underlying theories of change" at the department level or at the agency level? Please clarify.
4.4.1	Please clarify what InterAction means by "strategic program" levels. According to [our] definition of impact, we would rephrase the interpretive guidance as follows: "At both strategic program and project levels, evidence of process and progress should be captured through a valid and credible monitoring and evaluation system."
4.4.1	We are primarily a lobbying organization so these components (4.4.1.A, C, F, G, H) may not be as clear as if we were providing direct service.
4.4.1.B	This should be optional and isn't feasible not pertinent in all situations (for example, survivors of disasters aren't always marginalized but still need help)
4.4.1.C	Hard to see how we'd get comparable evidence given different costs structures in different countries, subjectivity of some of the proposed costs, and different value given to outcomes by different actors. I'd say whether we were able to implement in timely fashion and cost data perhaps not so directly related to benefits/results might be more feasible.
4.4.1.D	Hmmm, for all programs or projects? Applicable for emergencies – all of them?
4.4.1.E	Nice idea but who has the money? We've done one post-final in last five years
4.4.1.F	Ethical practice: refers to the evaluation or to the project/initiative?
4.4.1.G.	Is this a question that asks if we evaluate whether we follow our codes of conduct? the OGAC and other laws? Is it a replication of beneficiary accountability sections?
4.4.1.H.	Monitoring reviews/ evaluations should also result into "programme corrections at appropriate times
4.5.	"Mission" meant strategic goals or organizational mission? Discusses both organization and program goals. The means for monitoring and evaluation the two different levels would be quite different themselves.
4.6.	Negligible difference from 8.1.2
6.1.	I note that it only says evaluations should be conducted and accessible, not that the findings should be summarized and proactively shared with public audiences. I think the standard should go further, though I doubt that many in InterAction would agree.
8.1.2.	Negligible difference from 4.6. Suggest integrating them or stating clearly why they are in two places and whether they are deliberately similar or meant to be different.

Evidence of compliance with the standards and guidelines

The next set of questions in the survey asked respondents to indicate what kind of evidence of compliance with the standards their organizations would be able to present. The main objective of those questions was to provide InterAction with ideas of specific

things they should advise agencies to look for when conducting their Self-Certification Plus annual process.

Respondents were invited to indicate evidence they were ready to present and also evidence they think they would be able to provide in the future. Regarding the former, among the 50 respondents, 39 (78 percent) listed examples of evidence. Eleven of them (22 percent) either left the answer blank (4 respondents) or indicated they currently would have nothing or very little to show in terms of evidence of compliance with the standards (7 respondents). In terms of future evidence, 40 respondents (80 percent) indicated they would be able to provide evidence in the future while seven (14 percent) said they would not be able to do so. Three respondents (6 percent) did not answer this specific question.

The questions were open-ended, and Table 11 presents the categorization of the responses and the corresponding frequency for each category.

Table 11. Frequency of examples of current and future evidence agencies might be able to provide regarding compliance with InterAction standards and guidelines

Sources of Evidence	Currently have		Will have in the future	
	N	%	N	%
Reports of external/internal evaluations and monitoring reviews	22	44%	17	34%
Planning, monitoring, evaluation & learning systems documents	19	38%	17	34%
Organizational plans and policy	16	32%	5	10%
Planning and M&E guidelines, training materials, tools	13	26%	8	16%
Evaluation policies and standards	12	24%	10	20%
Board and senior staff meeting notes, annual report and audits	8	16%	2	4%
Meta-evaluations, synthesis of M&E findings	3	6%	6	12%
Budget showing resources allocated to M&E	3	6%	0	0%

The open-ended answers showed that respondents see two kinds of main evidence of compliance with the standards, both in the present and in the future. First are the actual

products (reports) of evaluation and monitoring activities. Second are the documents explaining the organization's planning, monitoring, evaluation and learning systems. One interesting fact was that only 12 respondents (24 percent) mentioned their M&E standards or policies as an example of evidence of compliance. This was somehow surprising since, as we saw in the previous Chapter IV (p. 92), slightly more than one-half of the respondents (26 respondents; 52 percent) indicated their agencies have developed their own M&E standards.

Few respondents indicated that their agencies would be able to provide meta-evaluations, synthesis/aggregation of findings from different evaluation studies, or budget allocations to M&E as evidence for compliance. One of the largest INGOs responding to our survey indicated that they would have difficulties to provide budget resource allocation to M&E as evidence of compliance with the standards. This respondent indicated that M&E resources are often incorporated into other line items along with other activities, especially at the country program or project levels.

We would probably obtain different results if we had made those questions closed-ended, i.e., if options for responses were provided. Since the questions were open-ended, the responses do not necessarily include all possible evidence the agencies might be able to present. With this possible limitation in mind, the frequencies for each category of evidence presented in Table 11 should be seen just as a preliminary indication of importance and not as a solid ranking.

Need for technical assistance with aspects of the standards and guidelines

Almost two-thirds of the respondents indicated that they would like to receive external support in some of the areas related to the InterAction standards, as shown in Table 12.

Table 12. Need for technical assistance in areas related to InterAction's M&E standards

Response	Frequency	Percent
Yes, my agency would benefit from technical assistance	32	64%
No, my agency does not need technical assistance	17	34%
Blank	1	2%

Respondents indicating interest in receiving technical assistance identified several areas their agencies could benefit from that support. Table 13 presents the categories mentioned by the respondents and their frequency.

Table 13. Type of technical assistance needed in areas related to the standards

Needed assistance	Frequency	Percent
Methodology: measuring impact, development of indicators, "hard to measure" initiatives, and qualitative & quantitative methods in general	11	22%
Setting up, improving and managing organizational M&E systems integrated with programming	7	14%
Evaluation standards/guidelines (operationalization, best practices, including InterAction's M&E standards/guidelines)	7	14%
Need for more sharing of experience re M&E among INGOs	5	10%
Special evaluation topics (cluster evaluation; meta-evaluation; cost/efficiency; participant satisfaction)	3	6%
General learning / any free training necessary for good M&E practice	2	4%
Use of external evaluators	1	2%
E-learning resources in M&E	1	2%
Evaluating bilateral assistance	1	2%
Improve communication of evaluation findings with U.S. public	1	2%
Strategic plan review questions	1	2%

Among the three most frequently mentioned areas, two are related to M&E practice and one to M&E principles. On the "practice" side, technical support on methods for monitoring and evaluation was clearly the most often mentioned. Agencies indicated an interest in getting external support or support from fellow agencies with greater

experience on several methodological aspects. Those aspects would include (i) impact measurement in diverse areas especially the ones that tend to produce “hard to measure” outcomes such as peacebuilding and advocacy, (ii) definition of process and success indicators/variables, (iii) measuring unintended impacts, and (iv) support on qualitative and quantitative methods in general.

Orientation on how to create, manage and improve internal M&E systems was the second most important aspect in which agencies indicated they could benefit from technical assistance. Based on the general responses to the survey, many organizations seem to be in the process of getting their M&E systems established or having to make changes and improvements in their existing systems due to expansion. Respondents mentioned that they would like help in making their M&E system integrated, useful, credible and independent from the other sectors of the agencies, especially planning and design.

In regards to M&E principles, seven respondents indicated that they could benefit from training on M&E standards and guidelines, especially the ones being proposed by InterAction. The most comprehensive response provided in response to the survey regarding this issue was from a representative working for one of the largest U.S.-based INGOs. She indicated they could benefit from support on: (i) raising awareness regarding M&E and its potential added value, especially strategies to achieve senior management buy-in; (ii) helping to establish parameters of performance for the standards, from "minimum acceptable" to "gold" standard; and (iii) explaining the standards in more detail, and explaining how they might apply at an operational level.

Finally, an interesting aspect raised by five respondents was that INGOs are learning too little from one another. There is an opportunity for InterAction to play an important role in promoting greater exchange of experiences among member agencies.

Discussing the survey findings

Undoubtedly, the great majority of the survey respondents found the M&E standards and guidelines proposed by InterAction important to their agencies. The few agencies[71] which presented issues about the standards and guidelines may be roughly divided into four groups: (i) agencies heavily funded by the U.S. Government and obligated to follow their M&E guidelines, (ii) agencies working primarily with technical support, advocacy or fundraising for other agencies that felt the standards are focused too much on development issues, (iii) agencies that think the standards have serious conceptual flaws since they do not take into account a rights-based approach to development, and (iv) agencies that think they do not have the necessary M&E structure to comply with the standards or that would have problems sharing some information they consider sensitive.

Addressing the situation of the first group (agencies heavily funded by the U.S. Government) seems quite challenging, and might require intense political negotiations by InterAction M&E leaders with U.S. Government agencies. Those agencies are required by contract to closely follow the regulations imposed by funders, which are usually quite different from the ones InterAction would like its agencies to adopt. For instance, USAID—the largest U.S. agency funding U.S.-based INGOs—holds grantees accountable for hundreds of process monitoring indicators, usually connected to outputs. Tracking those indicators is extremely time-consuming for project managers, leaving little space for the implementation of other evaluative activities. Also, funders might not provide specific resources for evaluation beyond what they require. InterAction's M&E leadership team has made efforts to establish a dialogue with representatives from USAID responsible for M&E functions. At least two meetings, open to all InterAction members, have occurred as a way to discuss USAID and InterAction's M&E policies.

[71] Nine agencies (20 percent of respondents)

However, it is naïve to think this process will have substantial and concrete results in a short timeframe.

The second group represented 10 percent of our survey sample and comprised the member agencies that are dedicated exclusively to advocacy, technical support to local NGOs or fundraising to UN agencies. A few of those agencies feel the standards and guidelines were mostly coined to organizations working with development or relief initiatives with responsibilities for delivering or supporting projects or programs reaching direct beneficiaries. The way the InterAction standards and guidelines are currently phrased, they intend to hold all members accountable to produce positive impacts in the lives of beneficiaries of aid assistance. Since those agencies do not have direct services targeting such beneficiaries, they feel the standards do not apply to what they do. This view is challenged by some leaders within InterAction who believe all actions developed by any INGO (or aid agency) should have clear connections with impacting the most vulnerable and poor people around the globe. According to their view, even agencies that do not provide or fund direct services to these populations should be able to articulate in some way how their work will plausibly have an impact in those people's lives and have an M&E system to hold them accountable for that impact. In Chapter VI we will provide a thorough discussion on how embracing the idea of different categories of standards (for evaluands, for evaluation processes and products, for evaluators, and for evaluation commissioners) will help solve this issue.

Dealing with the few agencies in the third group (the ones that claim the standards and guidelines are 'needs'-based and not 'rights'-based) is another challenge. At the core of the rights-based approach (RBA) is the relationship between rights-holders—the citizens—and duty-bearers—usually the state, but also institutions and individuals (Windfuhr 2000). RBA pushes for holding duty bearers to account for their obligations, empowers citizens to demand their rights, promotes equity and challenges discrimination (Theis 2003). Intensive advocacy work is clearly in the heart of the approach, and,

apparently, the perception that the InterAction standards do not adequately address this area is what seems to be troubling most of the organizations in this group. Of course the terminology used in the standards such as “addressing needs” (instead of “promoting rights”) seems also to be a concern for this group.

In looking into some of the literature specifically related to monitoring and evaluation within RBA (Theis 2003; Adams 2003; Chapman and Wameyo 2001; Patel 2001; Marcelino 2005), the authors include as some of the specificities of the approach (i) the use of qualitative inquiry as the primary method used to assess changes in attitude and practice, (ii) the need for M&E to be integrated and help promote rights issues (e.g., monitoring and reporting on rights violations), (iii) the need for strong participation of all stakeholders, especially the program recipients, in the evaluation design, implementation and reporting, and (iv) use of specific indicators to measure changes in people’s lives, in policies and practices, in equity, and in participation and empowerment. One could easily argue, however, that these aspects are not so unique and are actually part of most evaluations of the more “traditional” development or relief efforts. We think the list of standards proposed in chapter VI will contribute to address this issue.

Finally, the fourth group presents a different challenge for InterAction. They are the small and medium agencies with little M&E structure, if any, that find it hard to follow the standards and guidelines. Within this scenario, it seems that the actual contents of the standards would not matter too much, since their great challenge is lack of M&E capacity. InterAction will have to think of creative ways to address the needs of such group of smaller agencies which could potentially comprise one-third or more of the affiliated agencies. InterAction’s EPEWG is currently working on proposals to constitute a M&E supporting system for member agencies that will hopefully help address the issue of evaluation capacity building among smaller INGOs (EPEWG 2007).

The survey results brought to light the perspective of the primary impactees of the InterAction standards (member INGOs) on some important aspects that we need to take

into consideration when devising suggestions to improve such standards. In the next, and closing, chapter we will combine the lessons from this survey with the concepts and conclusions from our extensive review of the existing evaluation standards for aid evaluation (chapters III and IV). Our main objective is to propose important improvements to the evaluation standards from InterAction. The suggestions might also apply to improving sets of evaluation standards proposed by other aid agencies.

CHAPTER VI

DISSERTATION CONCLUSIONS AND A PROPOSAL FOR TAKING INTERACTION EVALUATION STANDARDS TO THE NEXT LEVEL

As its utmost objective, this dissertation contributes to the knowledge and practice of development aid evaluation. This paper has provided: (i) a review of the main movements for improving aid evaluations and their different strategies (Chapter II), (ii) a thorough analysis of the current evaluation standards proposed by aid agencies (Chapters II and III), and (iii) an empirical study of key aspects related to the structure, practice and principles of evaluation among U.S.-based INGOs (Chapters VI and V as well as Appendix B). The final contribution, which is delineated and discussed in this final chapter, comprises ideas for improving evaluation standards that should be seriously considered for adoption by InterAction and other aid agencies.

In this chapter, we will start with a summary of the main conclusions emerging from this dissertation. These conclusions refer to the current movements aimed at improving international aid evaluation, the level of institutionalization of evaluation in INGOs, and the quality of the different set of evaluation standards proposed by international development organizations. Then we will present a framework to assess the many standards identified in this dissertation and apply this framework to assess all of these standards under each one of the four categories devised from the outset of this dissertation—standards for (i) evaluands, (ii) evaluation processes and products, (iii) evaluators, and (iv) evaluation commissioners. We will select (and justify) the standards that should be considered for adoption by InterAction or other aid agencies.

We will also discuss implications for the proposed adoption of the standards by some groups of agencies within InterAction working primarily on advocacy, technical assistance and fundraising. A discussion of the limitations of this dissertation and closing comments will conclude the chapter and the dissertation.

Central findings

There are a number of movements to improve international aid evaluation involving most of the influential actors in the sector among donors (e.g., World Bank, UK's Department for International Development, Gates Foundation, etc), UN agencies, INGOs, professional associations, and research groups. In this dissertation, we identified and assessed 16 of the most prominent movements currently in place. The OECD/DAC development evaluation network seems to be the most influential to the field. This is due to the several significant contributions OECD/DAC has made to the field, its longstanding work (since 1970's), and the level of influence of its members (virtually all bilateral and multilateral agencies participate in this network). Two other movements led by different consortia of influential aid organizations, the Network of Networks on Impact Evaluation (NONIE) and the International Institute for Impact Evaluation, (3IE) have been gaining broader visibility over the past two years and are likely to have great influence in the field.

Ten of the reviewed movements propose more holistic approaches for improving international aid evaluation, including the development of evaluation standards. The remaining six are solely focused on impact evaluations. This group includes 3IE and NONIE. The main problem with these latter groups is the prevailing view that restricts evaluation functions entirely to the measurement of expected outcomes. To conduct a sound evaluation of an aid intervention, a competent and

thorough evaluator needs to rely on a number of criteria including side-effects, sustainability, exportability, ethicality, environmental responsibility, cost, and comparisons to possible alternatives.

An additional difficulty within the groups that push for "impact-only evaluations" is the few agencies that support the use of Randomized Control Trials (RCTs) as the method of choice for conducting evaluations. Even though RCTs are one of the powerful tools available to determine causal effects, they have limited use and applicability in aid evaluation. Such design is applicable to interventions that are discrete and homogeneous. In reality, however, the great majority of aid interventions are complex entities with a heterogeneous delivery of services influenced by several unpredictable factors and must constantly change to adapt to evolving contexts. Those factors make them ill-suited to RCTs.

The survey conducted by Chianca as part of this dissertation revealed important aspects of the current status of evaluation principles and practice among a sample of U.S.-based INGOs. The first relevant finding is that INGOs have a low level of institutionalization of evaluation. Approximately one-third of the agencies surveyed (34 percent) do not have any formal internal structure to manage and support their evaluation functions; 32 percent of them do not have any staff at headquarters with more than 50 percent of their time dedicated to M&E work.

The second important finding is that INGOs have limited capacity to learn from the evaluations they commission or conduct. Only 45 percent of the surveyed agencies reported having a system in place to collect reports from the evaluations they perform or commission. Furthermore, only 29 percent indicated that they synthesize, on a regular basis, the findings from such reports and share those syntheses within and/or outside the agency to promote learning. Therefore, it is fair to assume that most

INGOs lack a sound feedback loop to inform their decisions (e.g., to improve, expand, discontinue programs) and do not fully utilize the learning from their evaluations.

Limited knowledge about the quality of the evaluations of the INGOs' work is the third important finding emerging from the survey. Only 9 percent of the surveyed agencies indicated that they have conducted meta-evaluations of the evaluations conducted or commissioned by them. Additionally, a little more than one-half of the agencies reported having one-third or less of their programs, projects, or other major efforts evaluated by external professionals with evaluation expertise. This also reaffirms the existence of a lack of uniform quality in the evaluations and periodic assurances of objectivity.

We have reason to suspect that our sample of survey respondents was positively biased. Among the 50 agencies responding to the survey, we received answers from all INGOs that were part of InterAction's Evaluation Interest Group (EIG). Since those INGOs have staff dedicated to the M&E functions, they are likely to have more sophisticated evaluation systems than the other agencies. Furthermore, our sample had a large proportion of wealthiest agencies (annual expenses of 50 million U.S. dollars or more) which usually have more resources available for structuring M&E functions within their agencies. Therefore, we believe that the institutionalization and quality of evaluation, as well as the organizations' capacity for learning from evaluations are likely even more limited in INGOs that did not respond to the survey.

On the more positive side, the survey indicated that there is a sincere desire (or recognition of need) among INGOs to make improvements to their evaluation systems and practice. In many open-ended responses to our survey questions, respondents representing agencies that did not have a good evaluation structure indicated that they

were conscious about the situation and were taking steps to improve it. Almost one-forth of the respondents indicated that their agencies are taking the necessary measures to make M&E functions better structured or are developing their own evaluation standards.

This dissertation also revealed important findings related to the existing evaluation standards in the international development arena. A more blanketing finding was that agencies do not make distinctions among the standards in terms of the type of entity to which they refer. After reviewing the almost 40 different standards proposed by various agencies or coalitions, it became clear that evaluation standards can refer to (i) evaluands (e.g., aid interventions), (ii) evaluation processes and products, (iii) evaluators, and (iv) evaluation commissioners or major stakeholders. Awareness of those different groups of standards can help agencies consider more carefully all key aspects important to the quality of evaluations as they devise or adopt evaluation standards for their agencies.

Our study indicated that the most comprehensive set of evaluation standards proposed so far seem to be the ones from the OECD/DAC network on development evaluation (including the ones proposed specifically for humanitarian assistance, and the ones for conflict prevention and peacebuilding) and from United Nations Evaluation Group. However, both these sets of standards still come up short in adequately addressing some key issues such as the appropriate criteria to assess an aid intervention. For instance, these standards focus on donors' or countries' priorities when determining the relevance of an intervention instead of focusing on the needs of the intervention impactees. While donor or national needs are important, they are successful only when they compliment and address the needs of the impacted population. The standards also fail to take into account the quality of process and

exportability as an important part of the evaluation of an aid intervention.

The U.S. Agency for International Development (USAID), given its position as the major funder of U.S.-based INGOs, appears to be the donor agency with the greatest influence over the evaluation practice of InterAction members. Our thorough review of USAID's website and several key documents indicated that there is no common general set of evaluation standards supported by that agency. Reports from members of InterAction's EIG indicated that USAID usually contractually requires grantees to measure a large set of output indicators for M&E. According to the same source, this emphasis on tracking output indicators, besides overwhelming program managers and internal M&E experts, does not generate useful data for impact evaluations.

The M&E standards proposed by InterAction are, by and large, well accepted by the member INGOs. The overwhelming majority of survey respondents (80 percent) indicated that the proposed standards seemed relevant to their agencies. This seems a good indicator that InterAction is well connected with the interests and priorities of its member agencies. However, the applicability of the standards to INGOs working primarily on advocacy, technical assistance and fundraising was put into question. A few representatives from those INGOs indicated that the standards seem to be targeted more toward the agencies that provide development or relief services directly to intended beneficiaries.

A framework to assess the evaluation standards

Throughout this dissertation we have been using the term standards as defined by the Merriam Webster's Collegiate Dictionary (1995), i.e., referring to both (i) principles that others should conform to and (ii) criteria by which the accuracy and/or

quality of others is judged. A reasonable question to ask is why evaluation standards are important to the practice of evaluation. There are at least three very good reasons.

First, standards are extremely helpful to evaluators when performing the complex tasks related to practicing evaluations. Evaluations are (or should be) systematic processes to determine the quality, value and importance of complex entities. Standards make explicit the aspects evaluators need to consider to plan, perform, and evaluate evaluations with the highest possible degree of quality. Second, standards may help evaluators gain professional credibility if they indicate to clients and other evaluation stakeholders that they follow a set of accepted standards in their work. Third, standards can provide evaluation commissioners or other intended users (e.g., program managers) with some assurance that the evaluator they are hiring or evaluation they are commissioning will have better quality since they have followed recognized set of standards. Such standards will also provide them with some reference to consider whether the evaluation reports they are reading or listening to can be considered credible.

Standards, however, need to be carefully thought out since "[i]nappropriate standards can cause substantial harm by providing unwarranted assurances" (Picciotto 2006, p. 38). In defining good evaluation criteria (or standards)[72], Scriven (2000) indicated seven main attributes that should be taken into account when devising or assessing a criterion:

1. Valid (criterial status): They are directly connected to the quality of the evaluand, and not mere indicators—e.g., knowledge gain by participants is a criterion of merit (direct measure of quality) of a training course while level of attendance or students' opinion about the course are indicators (an

[72] In his original article Scriven (2000) refers to criteria of merit checklist which are the aspects that need to be considered by an evaluator when determining merit, worth and significance of an evaluand.

empirical approximation/correlate) of the quality of the program.

2. Complete: All relevant aspects related to its definition are included (no significant omissions).
3. Nonoverlapping: They are fairly independent from other standards (no significant overlap).
4. Commensurable: Have similar importance as the ones in the same level
5. Clear: Easily understood by users.
6. Concise: Can be easily remembered by users.
7. Confirmable: Can be measured or reliably inferred by the evaluators.

We will use these seven attributes as a reference to aid our decision on which standards should make it to the final list in each one of the four categories of standards we have identified which target evaluands, evaluations, evaluators, and evaluation commissioners. Establishing these four mega-categories of evaluation standards has the same mnemonic objective as having evaluation criteria or standards in the first place—a reminder of the important things that should be taken into consideration in an evaluation.

It is interesting to note that most of the existing set of evaluation standards or criteria for aid evaluation (e.g., OECD 2006) or even for evaluations in general (e.g., the Key Evaluation Checklist—KEC and the Program Evaluation Standards—PES) make no distinction between mega-categories. Rather, the existing standards treat each in similar venues, creating questions of ambiguity and applicability. Also, many of them miss some key aspects related to those categories, if they address them at all. UNEG (2005) is probably the one set of evaluation standards that came closer to establish mega-categories similar to the ones we are proposing. They have classified their 50 standards under four mega-categories: (1) institutional framework and

management of the evaluation functions, (2) competencies and ethics, (3) conducting evaluations, and (4) evaluation reports. UNEG's mega-categories 1 and 2 resemble quite closely our mega-categories "evaluation commissioners" and "evaluators", respectively. Their mega-categories 3 and 4 are both related directly to our mega-category "evaluations", even though their mega-category 3 has one or two standards related to the quality of the "evaluands". We believe the new framework we propose here provides more clarity to evaluators in identifying the relevant groups of criteria they should consider when conducting evaluations of aid interventions.

Our analytical process will first include an assessment of all standards proposed by the different agencies[73] based on the seven attributes of merit by Scriven (2000) described previously. Then, we will contrast the standards that were considered most relevant by Chianca with the ones currently proposed by InterAction. The result will be a final list of candidates for inclusion in the next revision of the standards by InterAction.

It is important to indicate that Chianca presented a preliminary report to the group within InterAction responsible for proposing new M&E standards, the Evaluation and Program Effectiveness Working Group (EPEWG) with the results of the survey from the 50 members. The survey report generated further discussion among the EPEWG members and major changes were made to their initial proposal. This revised version of the M&E standards and guidelines, included as Appendix G, was sent to InterAction's standards committee in October of 2007. This committee will make the final decision on what will be submitted to InterAction Board for final approval. Due to this dialogue with and the resulting actions of the EPEWG, the considerations and final suggestions for improvement in this chapter are made based

[73] Included in the summary tables at the end of chapters II and III.

on the most recent version of the InterAction M&E standards and not the one used in the survey.

Evaluation standards for evaluands

Our revision of the many different evaluation standards, criteria, policies, and guidelines from dozens of aid agencies generated a list of 19 standards focusing specifically on the quality, value and importance of evaluands, i.e., aid interventions.

Our analysis is synthesized in Table 14. In the first column, we present the best definition for each standard. Those definitions are a compound developed by Chianca of the definitions presented by the different agencies. In a few cases, Chianca expanded the definition of a criterion to make it more complete (these instances are explained in the footnotes). The second column provides an assessment of the standards based on the seven attributes described earlier. The last column has comments from Chianca about his rating for each standard. When needed, justifications are presented to the ratings provided, especially for the ones that were not considered in full compliance with one or more of the seven attributes.

When assessing the extent to which each standard meets Scriven's attributes of quality, we will provide a rating of 'FA' (for fully addressed), 'PA' (for partially addressed) and 'NA' (for not addressed). The assessment will be based on logical argumentation about the evidence available on each standard. In case of lack of evidence, Chianca will use his best judgment to provide the ratings.

Table 14. Assessment of standards for evaluands

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Impact: Positive or negative, expected or unexpected, direct and indirect, short-, mid-, and long-term effects attributable (beyond reasonable doubt) to an intervention	FA	FA	PA	FA	FA	FA	FA	Some overlap with effectiveness (positive expected outcomes)
Quality of process: Observation of key aspects of good practice during intervention implementation such as ethics, environmental responsibility, scientific soundness, proper coverage, stakeholder participation, etc	FA	PA	PA	FA	FA	PA	FA	Hard to make complete and concise given diversity of sub-criteria; may overlap with others (e.g., coverage, governance & mgmt)
Relevance[74]: Intervention's design and activities suited to meet important needs and underlying causes of priority problems faced by the impacted population, and appropriately tailored to local contexts and needs	FA	FA	FA	FA	FA	FA	FA	
Effectiveness: Planned objectives of the intervention being successfully achieved	NA	FA	NA	PA	FA	FA	FA	Achieving goals does not necessarily mean that needs are being met, hence closer to being an indicator of impact rather than a criterion; not in the same level of importance than the others; complete overlap with expected positive impact
Efficiency[75]: Comparison of intervention's immediate results with input considering costs and possible alternatives to determine if intervention is producing the best value for the investment compared to alternative interventions.	FA	FA	PA	FA	FA	PA	FA	Overlaps with costs; combines two very important criteria (cost and comparison) that might need different headings

[74] This criterion combines the concepts of relevance and appropriateness proposed by ALNAP (2006), since both are directly connected to meeting people's important needs. Appropriateness was originally only connected to emergency interventions; however, it seems very relevant to any other aid intervention.

[75] The new criterion proposed by UNEG, "value-for-money" was subsumed under efficiency, since they relate exactly to the same idea: comparison between resources invested and results achieved.

Table 14 – Continued

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Cost: Consideration of monetary, non-monetary, direct, indirect, actual and opportunity costs of an intervention	FA	FA	NA	FA	FA	FA	FA	Important component to determine efficiency.
Sustainability: Likelihood or evidence that intervention's positive impacts will continue after donor funding is withdrawn (financially, technically, environmentally and culturally sustainable)	FA	FA	PA	FA	FA	FA	FA	Overlaps with connectedness, linkages,
Connectedness: Measures to ensure that activities of an emergency nature are carried out in a context that takes longer-term and interconnected problems into account	PA	FA	PA	PA	PA	FA	FA	Closer to be an indicator of impact and sustainability than a criterion; overlaps with impact and sustainability; probably not in the same level of importance than the other criteria; need to read description in order to fully understand criterion.
Linkages: Establishment of connections between key actors and efforts at different levels to prevent conflict and build peace.	PA	FA	PA	PA	FA	FA	FA	Closer to be an indicator of impact and sustainability rather than a criterion; overlaps with sustainability and impact; probably not in the same level of importance than the other criteria
Coverage: Reaching all target population in need wherever they are	PA	FA	PA	PA	FA	FA	FA	Could be argued as having criterial status or as being an indicator of impact or a sub-criterion of quality of process; overlaps with impact and quality of process; possibly not in the same level of importance
Coherence: Level of alignment of security, developmental, trade, military and humanitarian policies, especially in terms of humanitarian and human-rights considerations	NA	FA	PA	NA	FA	FA	FA	Maybe an indicator of impact, but definitely not a criterion; not in the same level of importance than the others; overlaps with coherence/complementarity

Table 14 – Continued

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Coherence/complementarity: Coherence of intervention with other interventions and policies of donors, governments, and communities.	PA	FA	PA	PA	FA	FA	FA	Closer to be an indicator of impact, overlaps with coherence; not as important as others
Replicability/Exportability: Likelihood that part or the whole of an intervention could produce important contributions to other interventions	FA	FA	FA	PA	FA	FA	PA	Probably not as important as others; always based on inferences that sometimes might be hard to find good supporting evidence
Innovation: Level of creativity and innovation in addressing enduring problems or needs of target population.	PA	FA	PA	PA	FA	FA	FA	Closer to be a sub-criterion of exportability rather than a criterion
Scalability or expansion of impact: Potential for an intervention to be replicated on a larger scale to expand its impact	PA	FA	NA	PA	FA	FA	PA	Closer to be a sub-criterion of exportability rather than a criterion; overlaps with exportability and possibly with impact and; less important than others
Client satisfaction: Level of satisfaction of recipients with the intervention	NA	FA	NA	PA	FA	FA	FA	Clearly an indicator of relevance; not in the same level of importance than others
Community value added: Value added to recipients due to involvement of multilateral agency instead of bilateral	NA	FA	NA	NA	PA	FA	PA	Redundant, sub-criterion of comparisons under efficiency; very specific to EuropeAid; not as important as the others; not fully explained; unclear how to measure
Governance & management: Structures and processes created to govern complex interventions	PA	FA	PA	PA	FA	FA	FA	Closer to be a sub-criterion of quality of process or an indicator of future positive impact; overlaps with quality of process
Resources mobilization & management: Processes of soliciting funds by managers, provision of funds by donors, and allocation and control of funds in complex interventions	NA	FA	NA	PA	FA	FA	FA	Clearly a sub-criterion of quality of process or indicator of future impact; overlaps with quality of process.

From this analysis, we can conclude that there are six standards that seem to be clear candidates to be included in any list of standards related to the quality, value and importance of an aid intervention. Those standards were considered to have criteria status and no major flaw in relation to the other six assessed attributes. They are:

(i) Impact,

(ii) quality of process,

(iii) relevance,

(iv) efficiency,

(v) sustainability, and

(vi) exportability.

Impact is the only standard that was mentioned by virtually all agencies. Making it encompassing enough to capture all dimensions of impact (positive, negative, expected and unexpected, direct and indirect, short-, mid-, and long-term) is essential to ensure the standard's completeness. Other proposed standards that were not determined to be on the final list could be subsumed under impact, as indicators or correlates to this criterion. For instance, level of goal achievement (effectiveness) might be considered an indicator of expected positive impact (only if we assume the goals are relevant to existing needs). Another example is coherence; the level of alignment of policies and existing interventions of donors, implementing agencies, governments, and community can be an indicator of future impact. Linkages[76] can also be considered a third example—the level of success in bringing together the key actors and efforts relevant to the intervention can directly affect an intervention's impact and future sustainability. Coverage is also a good candidate to be listed under impact.

[76] Even though initially conceived as being specific for peacebuilding initiatives by their original developers, it can also be considered relevant to other aid interventions. Establishing the right connections among key actors interested or already working in the region or on the issues addressed by an aid intervention is very important to ensure impact and future sustainability.

Having a comprehensive definition of impact allows for all important aspects to be comparatively analyzed and will give a clear sense of whether or not the overall real impact was relevant, not only those desired. Furthermore, there are distinguished scholars that make the argument that impact can actually subsume additional aspects that are listed in this dissertation as stand-alone standards, such as relevance, exportability and sustainability (Clements 2007, November). The main issue with this argument is that expanding the definition of impact to encompass all those aspects might cause evaluators to overlook some of them and, therefore, compromise the quality of the evaluation.

Quality of process is probably the most complex of the standards given the many important sub-criteria related to it such as ethicality, environmental responsibility, scientific soundness, etc. Governance & management and resource mobilization & management, which are seen as especially relevant to complex, transnational interventions, can also be seen as sub-criterion of quality of process. The quality of the governance system and the way resources are mobilized and managed are directly related to the quality of that intervention's process. These two aspects might also be considered as indicators of impact for the same reasons argued for connectedness and linkages in the previous paragraph.

Relevance seems to be an undisputable and solid standard. It is the first in the list of all evaluation criterion proposed by multilateral and bilateral agencies to assess aid interventions. Attention needs to be paid, however, to the focus of relevance. Some of the agencies, instead of focusing it primarily on the needs of the target population, put too much weight on the intervention's alignment with donor priorities or recipient countries' policies which weakens the standard. Client satisfaction is an important indicator for relevance. It cannot, however, have the status of a criterion since the level

of recipients' satisfaction can be connected to different factors and does not necessarily mean that an intervention is good or bad.

Efficiency carries within its definition three very important components: costs, benefits and comparisons. The terms have been mostly used in development evaluation as proposed by the OECD/DAC as a compound to determine how well the resources are being used in comparison with alternatives. However, given the importance of each one of those criteria, there are some who would argue that cost and comparisons should become separate criteria (benefits are already subsumed under impact). The strongest argument for making them stand-alone standards is that evaluative conclusions can be influenced independently by an evaluand's cost (e.g., this intervention is too expensive) and by the comparison with alternatives (e.g., intervention X is superior to intervention Y). Impact alone does not necessarily lead to questions of viable alternatives or the efficient allocation of resources. Another good argument to separate those components is the fact that by doing so, it will be possible to call the evaluators' attention to consider a more comprehensive cost analysis (including non-monetary, indirect, and opportunity costs). Either way (maintaining or separating the terms), it is important that these aspects be looked very carefully in an evaluation.

Sustainability is clearly another important standard to determine the quality of an aid intervention. Considerations of aspects beyond continued financial support, such as cultural appropriateness, environmental responsibility and self-reliance of individuals, groups, and institutions are essential for sustainability. At least two other criteria mentioned in the reviewed standards (connectedness and linkages) can be considered indicators of (or contributors to) sustainability. Both deal with important conditions to ensure that the flow of positive impact will continue after major external

funding for an intervention is withdrawn.

Exportability (or replicability) is connected to determining the importance of an intervention. The greater the possibilities of applying new ideas, technologies, and processes developed by an intervention in other contexts, the greater the importance of an intervention. Innovation in addressing social problems is an important sub-criterion of exportability. Exportability is an umbrella for lessons learned and lessons applied. This does not necessarily mean by implication that exportability equals success in other environments. Rather, it includes best practices that are appropriate but helps to ensure that the organization and the programs are learning from experience and demonstrates the positive attributes that can be further utilized, expanded, or adopted in other areas.

Implications for the InterAction standards related to evaluands

In the current (proposed) version of the InterAction M&E Standards (Appendix G) there is one standard that briefly mentions the standards related to the evaluands. Standard 7.X.4 reads: “An agency’s planning, monitoring and evaluation system should draw on commonly accepted professional principles and standards in planning, monitoring and evaluating programs. These systems should take into account not only the defined organization-wide criteria for success toward achievement of its mission and program goals, but also basic components of sound evaluations including, but not limited to, quality of process, intended and unintended outcomes and impact, costs, and sustainability” (InterAction 2007).

The guidelines for standard 7.X.1 clearly address most of the criteria listed in the previous section (e.g., impact, quality of process, sustainability, etc). However, the standard itself refers to the importance of mainstreaming evaluation within the

agencies and ensuring adequate resources for evaluations. The first suggestion for a future revision of InterAction's M&E standards is to make them into standards for evaluating the quality of interventions supported by the member agencies in what is now included as guidelines for standard 7.X.1. The list should be revised taking into consideration the list of six standards presented on page 149 with the proper definitions from Table 14.

A remaining question is how to deal with the issues raised by member agencies that are primarily dedicated to advocacy, technical assistance, and fundraising. Most of them do not have programs or projects providing direct services to people where those services would be considered the evaluand in applying the proposed standards. Our study was not designed to address those important questions and further studies should be conducted in that direction. However, based on discussions with members of this dissertation's advisory committee and a cursory review of existing practices, some ideas were expressed that are worth being considered to help address some of these questions.

First, it is important to recognize that there are specificities in the practice and evaluation of advocacy initiatives. Similar to what happened in the areas of humanitarian action and peace-building, it is worth considering the establishment of efforts aimed at developing specific evaluation criteria for assessing advocacy interventions. One, however, can make the case that most of the six standards we propose here might be applicable to advocacy interventions. When assessing advocacy efforts, a thorough evaluator will probably have to consider the quality of their implementation process (e.g., ethicality, scientific soundness, etc.), the relevance of their design and activities to address the needs of the impactees, their efficiency (best use of resources for the results obtained), their possibilities for exportability (are there

components or aspects of the advocacy effort that can be useful to other initiatives?), and their impact (including the ones affecting the people whose rights and needs they advocate for). Sustainability, at a first glance, might not seem universally applicable to all advocacy interventions. Nonetheless, a number of advocacy efforts seek long-term changes of the current situation and a sustainable stream of benefits should be expected and assessed.

The agencies working exclusively to provide technical assistance to local NGOs and the ones dedicated to fundraising activities, even though clearly working as intermediary entities, are still playing an important role in supporting direct efforts to reduce poverty and oppression throughout the world. Those agencies at a minimum should stimulate the organizations they work with to conduct sound and thorough evaluations of their aid interventions using the standards proposed in this section. Furthermore, as general good practice, they should also evaluate their work as a way to improve their strategies and to be accountable to the main stakeholders of their organizations. If the objective of such INGOs is to strengthen the capacity of local NGO partners, the process and outcomes of such capacity-strengthening should be subject to evaluation, as well as the effectiveness and sustainability of their NGO partners. After all, the idea of setting standards is to help improve evaluations conducted in the sector so the quality of work by INGOs can improve and, hopefully, contribute to improvements in the conditions of many impoverished and oppressed people in developing and transitional countries.

Standards for evaluation processes and products

Our in-depth analysis of the many evaluation standards, guidelines, criteria or principles proposed by international development agencies resulted in the

identification of 15 standards related to the quality of the evaluation's processes and products. Table 15 presents our assessment of the quality of those standards based on the seven quality attributes proposed by Scriven. For clarity and to ensure proper weight to the originally mentioned standards, some that were initially grouped under one general heading in some of the previous analysis were unpacked. For instance, accuracy was used in the analysis of the standards shared by the 14 INGOs to refer to rigor of methods and instruments for data collection and analysis, quality of the information sources, and validity of the conclusions reached. In the following table, accuracy was unpacked into the three mentioned criteria. On the other hand, some standards that were overlapping extensively were collapsed into a single standard. For instance, systematic data analysis was subsumed under rigor in methods for data collection and analysis.

Table 15. Assessment of standards for evaluations

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Timely and within budget	FA	FA	PA	FA	FA	FA	FA	May overlap with evaluation use
Minimal disruption of program operations	FA	FA	FA	FA	FA	FA	FA	
Rigor and appropriateness of methods for data collection and analysis	FA	FA	PA	FA	FA	FA	FA	May overlap with stakeholder participation
Valid and balanced conclusions	FA	FA	PA	FA	FA	FA	FA	May overlap with stakeholder participation
Reliability of information sources	FA	FA	PA	FA	FA	FA	FA	May overlap with stakeholder participation
Actionable recommendations/lessons	FA	FA	PA	FA	FA	FA	FA	May overlap with stakeholder participation
Metaevaluation	FA	FA	FA	FA	FA	FA	FA	
Clear reports (easily understood and in appropriate language)	FA	FA	PA	FA	FA	FA	FA	May overlap with use of findings
Focused executive summary	FA	FA	PA	FA	FA	FA	FA	May overlap with clear report and use of findings
Discussion of program context	FA	FA	FA	FA	FA	FA	FA	

Table 15 – Continued

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Clear presentation of evaluation methodology and limitations	FA	FA	FA	FA	FA	FA	FA	
Inclusion of stakeholders' comments in report	NA	FA	NA	FA	FA	FA	FA	A sub-criterion of balanced conclusions; overlaps also with stakeholder participation
Description of program logic	PA	NA	FA	NA	FA	FA	FA	Describing program logic is not sufficient; what is really important is to describe the program as it actually was and not how it was intended to be; unless redefined, not as important as the others
Transparency of evaluations' terms of reference and reports	NA	FA	PA	NA	FA	FA	FA	Unclear connection between transparency and quality of evaluation process and products; may be an indicator for use of findings; not in same level of importance than others
Stakeholder participation (in planning, data collection and analysis, interpretation of findings, and development of recommendations)	NA	FA	NA	NA	FA	FA	FA	A non-participatory evaluation can't be considered poor evaluation (depends on evaluation purposes—capacity building vs. accountability); overlaps with balanced conclusions, reliability of info sources, appropriate methods, and actionable recommendations; not in the same level of importance as others

The analysis presented in table 15, allows us to draw some important conclusions. First, there are at least 11 standards related to the evaluation process and products that clearly meet all criteria for sound standards proposed by Scriven (2000). They are:

(i) Valid and balanced conclusions

(ii) Reliable sources of information

(iii) Use of rigorous methods for data collection and analysis

(iv) Timely & within budget

(v) Minimal disruption of program

(vi) Metaevaluation

(vii) Clear reports (easily understood and in appropriate language)

(viii) Focused executive summary

(ix) Discussion of context

(x) Discussion of methodology and limitations

(xi) Inclusion of actionable recommendations and lessons learned

One of the standards needs to be refocused in order to fully meet the validity criterion (criteria status) set by Scriven. Description of program logic should be expanded to include a description of what actually the initiative being evaluated did and not only what it intended to do. The standard should be renamed as:

(xii) Description of program (including its logic and actual implementation)

Three standards did not meet the main criterion of validity and therefore were not included in the final list for further adoption. They are: (a) inclusion of stakeholders' comments in report; (b) stakeholder participation; and (c) transparency of evaluations' terms of reference and reports. There are important reasons to justify our conclusion. First, incorporating stakeholders' perspectives about the evaluation conclusions is clearly part of the standard related to 'valid and balanced conclusions' and, therefore, should not be considered a stand-alone standard.

Second, stakeholder participation in all aspects related to the evaluation seems to be too broad to be considered a valid standard for good evaluation process and

products. There are certainly some instances when the participation of primary stakeholders will be important to ensure the quality of the evaluation. Those instances include suggestions of reliable sources of information and of culturally appropriate data collection strategies, provision of their perspectives on the evaluation findings, and feedback on evaluation recommendations so they can be more grounded in reality and increase the possibility of adoption. Since those aspects are already contemplated in other standards, it would seem redundant to have a general standard for stakeholder participation.

Finally, evaluation terms of reference and reports might not have been broadly disseminated for several reasons. Their disclosure, similarly to the use of evaluation findings, will depend directly on decisions of the evaluation commissioners. An evaluation should not be considered of lower quality due to possible limitations in disclosing those aspects.

A comparison of the final list of 12 standards related to the evaluation processes and products suggested to be adopted by InterAction with the Program Evaluation Standards (PES)[77] helped us identify three important aspects that were not tackled by any of the other sets of evaluation standards reviewed for this dissertation. The following are the most relevant ones that can directly influence the quality of both the process and products of an evaluation, and that should also be considered by InterAction:

(xiii) Identifying the different stakeholders and their information needs and political agendas

[77] Even though the PES have been developed originally for educational programs, their most recent revision has made their definitions more comprehensive to be relevant programs in general. Also, they have been extensively used as inspiration for other sets of evaluation standards (e.g., AfrEA). The comparisons with the PES in this dissertation are made as a strategy to generate ideas for possible missing standards that may help improve the set of standards proposed by InterAction.

(xiv) Making explicit the values used to interpret the evaluation findings, so the basis for value judgments are clear.

(xv) Being cost-efficient, so the value of the information generated will justify the resources invested in the evaluation.

Implications for the InterAction standards related to evaluation processes and products

InterAction has only one standard that is somehow related to the quality of evaluation process and products. Standard 7.X.2 is generally connected to stakeholder participation and reads: “From the outset of program planning, a member organization shall collaborate with partners, clients/intended beneficiaries and other stakeholders in developing mutually satisfying goals, methods, and indicators for project and program activities and results” (see Appendix G). It is clear that there are several important standards missing in the current list proposed by InterAction and that should be carefully considered by that agency in future revisions of their evaluation standards.

In terms of agencies working with advocacy, technical assistance and fundraising, the 15 standards discussed in this section seem relevant to them. Different from the standards related to the aid interventions discussed in the previous section, the standards discussed here can be applied to any evaluation any InterAction member may conduct or commission.

Standards for evaluators

We have identified among the many sets of evaluation standards proposed by aid agencies, 10 that are related to the work of evaluators. Table 16 presents our assessment of the quality of those standards based on the seven attributes by Scriven.

Table 16. Assessment of standards for evaluators

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Develop clear/rigorous design: to ensure the quality of evaluations	NA	FA	FA	FA	FA	FA	FA	Clear indicator for competence or systematic inquiry
Ethicality: evaluators should demonstrate honesty and integrity in all aspects of the conduct of an evaluation (e.g., negotiating contracts, presenting findings, etc)	FA	FA	PA	FA	FA	FA	FA	May overlap with independence and with respect for people
Competence: evaluators should have the needed skills, education, and cultural competence to perform the required functions of an evaluation	FA	FA	PA	FA	FA	FA	FA	May subsume conduct of evaluability assessment and development of rigorous evaluation design
Systematic inquiry: evaluators ensure accuracy and credibility of the evaluations they conduct by implementing thorough and consistent studies	FA	FA	PA	FA	PA	FA	FA	May subsume development of rigorous evaluation design; needs further clarification
Respect for people: evaluators should make sure sensitive individual information is not released and that all evaluation stakeholders are treated with respect ands dignity	FA	FA	PA	FA	FA	FA	FA	May overlap with ethicality
Responsibility for general welfare: evaluators should ensure evaluations will help organizations meet the needs of diverse groups of impactees	FA	FA	FA	FA	PA	FA	FA	Needs further clarification
Independence: evaluators should not have any vested interest in the initiative being evaluated	FA	FA	PA	FA	FA	FA	FA	May overlap with ethicality
Diversity: evaluation team should be diverse (gender, ethnicity, etc.) to increase cultural sensitivity	NA	FA	FA	FA	FA	FA	FA	Indicator for competence
Disclosure of disagreements: among members of the evaluation team or between evaluators and stakeholders	FA	FA	PA	FA	FA	FA	FA	May overlap with ethicality

Table 16 – Continued

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Conduct evaluability assessment: to determine whether there are the necessary political conditions and resources for an evaluation to take place; it is also used for framing evaluations to ensure reasonable evaluation design.	NA	FA	FA	FA	FA	FA	FA	Clear indicator for competence

Based on our analysis, seven standards have adequately met Scriven's criteria and should be considered by InterAction for adoption. They are:

(i) Ethicality

(ii) Competence

(iii) Systematic inquiry

(iv) Respect for people

(v) Responsibility for general welfare

(vi) Independence

(vii) Disclosure of disagreements

Even though independence has been defined by the agencies as the absence of any vested interest of the selected evaluators in the initiative being evaluated, it is important to acknowledge that there is an underlying structural issue that can affect evaluator's independence beyond what have been indicated here. Clements (2005b) has identified the many incentives influencing evaluators' independency and that have led to positive bias and analytic compromise in aid evaluations. Those incentives include "political incentives for donor and recipient governments, organizational incentives for development agencies, and personal incentives for managers" (p.13). To

address such threat, Clements has suggested a structural solution that encompasses the creation of an independent association, paralleling the ones created by accountants and auditors. The members of such an association need to achieve specific qualifications and must follow a set of rules that will ensure their independence from aid management (pp.30-33).

From the 10 standards listed in Table 16, three should be subsumed under other standards. Conduct of evaluability assessment and diversity should be embedded under competence—even though it is important to recognize that diversity of the evaluation team is not always required for competence; this fact however does not make this criterion eligible for meeting all Scriven's attributes and become a stand alone criterion in our short list. Development of clear/rigorous design could be either subsumed under competence or systematic inquiry. All other standards were considered strong enough to meet Scriven's seven criteria of merit and should be seriously considered by InterAction in future revisions of its evaluation standards.

Implications for the InterAction standards related to evaluators

The current version of the InterAction M&E standards has no reference to any standard related to the behavior and competence of evaluators. The standards discussed in this session can clearly apply to all InterAction members regardless of their primary focus of work (development, relief, advocacy, technical assistance or fundraising).

Standards for evaluation commissioners

We identified seven standards related to the commissioners of evaluations. Table 17 presents our assessment of the quality of those standards based on the seven

quality attributes proposed by Scriven.

Table 17. Assessment of standards for evaluation commissioners

Standard	Attributes of quality							Comments
	Valid	Complete	Non-over-lapping	Commen-surable	Clear	Concise	Confirm-able	
Realistic scope: Provision of clear direction and realistic scope to the evaluation	F A	F A	F A	F A	F A	F A	F A	
Open access: Ensuring free and open access to needed information	F A	F A	F A	F A	F A	F A	F A	
Protect evaluation from external pressures: Protect evaluators from pressures from managers or other stakeholders	F A	F A	F A	F A	F A	F A	F A	
Commitment to use: Committing to use evaluation findings and promote learning from evaluations	F A	F A	F A	F A	F A	F A	F A	
Proper staffing: Hiring capable evaluators	F A	F A	F A	F A	F A	F A	F A	
Ensure resources: Provision of adequate resources to the design and conduct of sound evaluations	F A	F A	F A	F A	F A	F A	F A	
Promote joint evaluations: Ensure partnership among agencies in conducting evaluations of joint efforts	F A	F A	F A	F A	F A	F A	F A	
Use of findings	NA	FA	PA	FA	FA	FA	FA	Sometimes evaluations do not get used for different reasons, but that does not mean an evaluation was poorly conducted or produced a bad report; however, an evaluation that is used will have greater value and importance; overlaps with clear report

All but one standard were considered to have adequately met Scriven's criteria for quality of standards and should be considered by InterAction. Many factors can influence use of evaluation findings by evaluation commissioners or other

stakeholders including several of the ones already included in the final list of standards for evaluation processes and products (e.g., timely and clear reports; valid conclusions; actionable recommendations). However, using the findings relies significantly on the decisions of evaluation commissioners and that is why the standards on Table 17 are important. The following is the list of standards that were identified as important to be considered by InterAction:

(i) Realistic evaluation scope

(ii) Open access to information

(iii) Protect evaluation from external pressures

(iv) Commitment to use evaluation findings

(v) Ensure adequate resources

(vi) Proper staffing of evaluations

(vii) Promotion of joint evaluations

Evaluation commissioners can, with no doubt, contribute to ensure the quality of aid evaluations by including safeguards to protect the process from external influences. However, the structural incentives in development practice pointed out by Clements (2005b) might be a larger force influencing the level of independence of such evaluations by increasing opportunities for positive bias (see pages 164-65 for more details).

'Promotion of joint evaluations' is especially relevant for joint initiatives involving multiple agencies. Avoiding the waste of resources is the first reason to support such evaluations. Another very important reason for promoting joint evaluations is to prevent overwhelming program participants or managers with too many evaluation requests from different agencies.

Implications for the InterAction standards related to evaluation commissioners

This is clearly the category where InterAction has developed the greatest number of standards. The following five standards were considered as having connections to the bodies within INGOs responsible for commissioning evaluations (Appendix G):

> 2.6.4 The agency's Board shall ensure that the organization (i) articulates organization-wide criteria for success as defined by its vision, mission and major program goals; (ii) incorporates and practices regular, deliberate evaluative activities to determine achievements of program goals and mission fulfillment; (iii) mainstreams and utilizes monitoring and evaluation in the agency's policy, systems and culture; and (iv) allocates adequate financial and human resources for the organization's strategic evaluation needs.
>
> 3.8 The member organization shall be committed to full, honest, and accurate disclosure of relevant information concerning its goals, including criteria for objectively measuring progress and success of its programs, finances, and governance in achieving the goals.
>
> 3.9 To inform its ongoing strategic planning process, a member organization shall incorporate a deliberate and intentional process of monitoring and evaluating the organization's progress toward achievement of its mission and major program goals.
>
> 7.X.1 A member organization shall have a policy (or similar operative document) that defines how monitoring and evaluation are integrated within program / project management, as well as evidence that the policy is being adhered to.
>
> 7.X.3 A member organization shall assure that program and project budgets allocate adequate resources for monitoring, evaluation and

institutional learning

Contrasting the list of seven standards in Table 17 that met Scriven's criteria and the standards currently proposed by InterAction reveals an interesting scenario. InterAction standards 7.X.1 and item (i) of standard 2.6.4 clearly address the 'realistic evaluation scope' standard on Table 17. InterAction's standard 7.X.3 addresses standard 'ensure adequate resources' on Table 17. The other standards on Table 17 do not seem to be clearly addressed by the InterAction standards. This analysis also reveals that two new standards related to the evaluation commissioners proposed by InterAction are not included in Table 17. Those standards are:

(viii) Mainstreaming evaluation thinking and practice among INGOs[78], and

(ix) Disclosure of evaluation criteria for assessing the organization[79]

Regardless of the primary focus of an INGO (e.g., development, advocacy, etc) these standards will be equally applicable.

Limitations of the dissertation

The sample of 50 INGOs that participated in the survey included in this dissertation was not randomly selected. Possibilities to generalize the findings to all 167 members of InterAction are limited. However, the large number of wealthier INGOs in the sample possibly indicates that the situation of INGOs in general is even worse than what was identified. Agencies with less money will probably have fewer resources to invest in evaluation. Regardless of the limited number of survey respondents, this is the largest empirical study conducted about evaluation principles and practice in U.S.-based INGOs.

The study of the different standards proposed by aid agencies conducted as part

[78] Compound of InterAction standards 3.9 and 2.6.4 (items (ii) and (iii))
[79] InterAction standard 3.8

of this dissertation is comprehensive. Sources for identifying the existing standards included not only the available literature (conventional and electronic), but also an extensive informal query with some of the most prominent evaluators working on the international development field. Nonetheless, there is still the possibility that there might be still other standards out there, especially from INGOs not based in the U.S., which were not included in our review.

The suggestions presented for improving the InterAction M&E standards were made as complete as possible. However, they were not written in a way that allows them to be directly incorporated in the existing set of InterAction general standards. InterAction will have to consider them carefully and find the best way to put them on a narrative form that will best fit their organizational language.

The assessment of the different standards using Scriven's framework to determine the quality of criteria of merit can certainly be improved. Only Chianca made the assessment decisions which bring in the risk for individual bias. An expanded panel including other experienced judges to prove the ratings for the standards will probably increase the accuracy of the process and could constitute subject of further research.

Closing comments

This dissertation has brought unique contributions to the field including: (i) assessment of current movements to improve aid evaluation; (ii) analyses of the existing evaluation standards proposed by aid agencies; (iii) first large-scale empirical study of evaluation principles and practice among U.S.-based INGOs; and (iv) proposal of specific improvements to evaluation standards of InterAction (and other aid agencies).

Having a good set of sound evaluation standards is a huge step forward in the direction of improving evaluation practice and, hopefully, the work done by aid agencies. However, simply having good standards on paper does not mean that they will be applied and that improvements in practice will follow. Strategies to provide support to aid agencies to incorporate the right evaluation standards into their daily operations are essential. Further research on how evaluation standards are being implemented and with what results should be pursued in the future.

One aspect that was not addressed in this dissertation is the idea of creating consortia of agencies on a specific sector to think evaluation more thoroughly within that sector, such as ALNAP is doing for humanitarian action. This can be a key to help push the field forward and improve evaluation practice in those specific sectors. The possibility of commissioning joint evaluations across agencies, instead of only project-level evaluations, might contribute for more significant learning, and should be object for further studies.

The development of standards for aid evaluation has come a long ways. In the past two decades we have witnessed important improvements and sophistication of the sets of evaluation standards proposed by different aid agencies. Regardless of such advances, there is still room for significant improvement. It is hoped that this dissertation has gone some way toward making a contribution in this direction.

APPENDIX A

Acronyms

The following is a list of the acronyms used in this dissertation. Throughout the text the first appearance of a compound term is followed by its acronym in parentheses (e.g., Development Committee of the Organisation for Economic Co-Operation and Development (OECD/DAC). After their first appearance only the acronym is used in text.

3IE	International Institute for Impact Evaluation
AEA	American Evaluation Association
AfDB	African Development Bank
AFrEA	African Evaluation Association
ALNAP	Active Learning Network for Accountability and Performance in Humanitarian Action
AsDB	Asian Development Bank
CDA	Collaborative for Development Action Inc.
CGD	Center for Global Development
CI	CARE International
CIDA	Canadian International Development Agency
CPDC	Conflict, Peace and Development Co-operation
CPPB	Conflict Prevention and Peacebuilding

DANIDA	Danish International Development Agency
DIME	Development Impact Evaluation
DFID	Department for International Development (UK)
EBRD	European Bank for Reconstruction and Development
EC	European Community
ECG	Evaluation Cooperation Group (MDB)
EDEPO	Centre for the Evaluation of Development Policies
EHA	Evaluation of Humanitarian Action
EIG	European Investment Bank
EPEWG	Evaluation and Program Effectiveness Working Group (InterAction)
EuropeAid	European Commission Agency for External Cooperation
GEF	Global Environmental Facility
GP	Guiding Principles for Evaluators (AEA)
IDB	Inter-American Development Bank
IDEAS	International Development Evaluation Association
ILO	International Labor Organization
IMF	International Monetary Fund
INGO	International Non-Governmental Organizations
InterAction	American Council for Volunteer International Action
IOCE	International Organization for Cooperation in Evaluation
J-PAL	Abdul Latif Jameel Poverty Action Lab
KEC	Key Evaluation Checklist

LAC	Latin America and the Caribbean
MDB	Multilateral Development Banks
MDRC	Manpower Demonstration Research Corporation
M&E	Monitoring and Evaluation
NGO	Non-Governmental Organization
NIS	Newly Independent States
NONIE	Network of Networks of Impact Evaluation
OCHA	United Nations Office for Coordination of Humanitarian Affairs
ODI	Overseas Development Institute
OECD/DAC	Development Assistance Committee of the Organisation for Economic Co-operation and Development
OED	Operations Evaluation Department (World Bank)
PES	Program Evaluation Standards
RBA	Rights-Based Approach
RBM	Results-Based Management
RCT	Randomized Controlled Trial
SCP	Self-Certification Plus
SEGA	Scientific Evaluation for Global Action
SIDA	Swedish International Development Cooperation Agency
TEC	Tsunami Evaluation Coalition
UN	United Nations
UNDP	United Nations Development Programme

UNEG	United Nations Evaluation Group
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development
WB	The World Bank

APPENDIX B

Survey results on INGOs' M&E structure and practice

The second part of the survey explored some key issues related to the INGOs' M&E structure and practice including (i) organization of M&E functions, (ii) M&E staffing, (iii) strategies to learn from evaluations and ensure their quality, and (iv) level of independence of evaluations.

M&E structure

In terms of structure, the survey asked respondents to indicate how the M&E functions were organized and how many M&E professionals their agencies had. The responses presented a quite diverse picture of the way agencies organize their M&E functions and also provided some sense of the level of sophistication and quality of those structures. Figure 1 presents the way respondents classified how their agencies organize their evaluation functions.

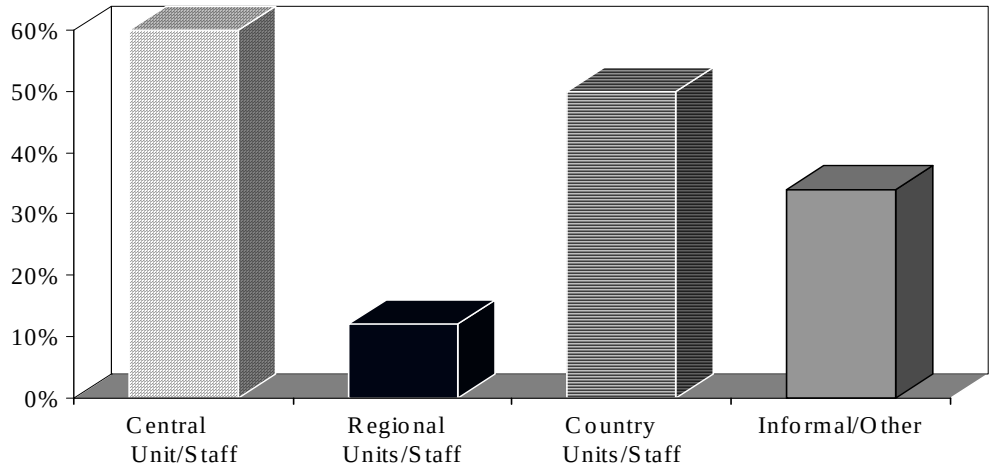

Figure 1. Reported monitoring and evaluation structure in INGOs

As shown in Figure 1, 60 percent of the agencies reported having M&E structure at a central office, usually located in the U.S., 12 percent reported having it at regional offices (e.g., Latin America, Africa, etc), and 50 percent at country-level offices. Of course many agencies, especially the larger ones, reported having staff in two or even in all three levels. There were 13 agencies (26 percent) reporting M&E structure at central and country offices; one (2 percent) at central and regional offices; and five (10 percent) at all three levels—central, regional and country. There were five respondents who mentioned that their agencies, besides having specialized M&E staff allocated to offices at different levels, also have other important elements complementing their M&E structure including program managers of specific areas (e.g., HIV/AIDS, education, environment, etc.) with part of their time dedicated to monitoring and/or evaluation functions and external evaluation consultants hired for specific projects.

Of the 50 respondents, 17 (34 percent) indicated that their agencies had informal or other types of M&E structure. In general, those agencies do not have specific staff dedicated to M&E functions or, if they do, they are considered by the respondents not enough to adequately address the agencies' existing M&E needs. Another common thread among those agencies was the fact that none of them seemed to have any general policy integrating the different M&E functions taking place within their different organizational areas. Furthermore, they are not always small agencies, as one might have expected; there were seven of those agencies classified as small[80], three as medium, four as large, and three as very large.

Four of those 17 respondents indicated that systematic monitoring and

[80] Agencies' size was defined based on annual expenses: there were 12 agencies considered small (up to 10 million U.S. dollars per year); 14 medium (between 11 and 50 M/yr); 10 large (between 51 and 100 M/yr); and 13 very large (more than 100 M/yr). We could not obtain financial information about 1 of the 50 agencies.

evaluation activities only take place on an "ad hoc" basis, sometimes using external consultants. In four responses, the monitoring and evaluation of existing projects are claimed to be conducted by headquarters' staff during site visits, or by project staff or even volunteers—who report their findings to officers at headquarters. In two agencies, it was reported that program managers analyzed success based on compilation of project reports and other data from the field. Three agencies indicated that their M&E structure is restricted to headquarters staff preparing reports to funders based on specific indicators determined by the funders, usually U.S. government agencies, and tracked by project/program staff.

Two respondents from agencies dedicated to advocacy and fundraising summarized their agency's M&E structure as an effort by headquarters to keep close financial control of operations, including regular auditing and assessments by independent watchdog agencies, such as Better Business Bureau Wise Giving Alliance[81] and Charity Navigator[82]. One respondent connected his agency's lack of formal M&E structure to the fact that they are part of a world federation of organizations that have quite independent functioning and operational structures; he points out that even though member agencies should follow some basic principles, none are specifically related to M&E functions. Another way of structuring M&E functions within an INGO was presented by one respondent from a medium-size agency dedicated to development work. She informed that her agency has a person

[81] The Better Business Bureau Wise Giving Alliance is a nonprofit organization dedicated to help potential "donors make informed giving decisions and advances high standards of conduct among organizations that solicit contributions from the public." They evaluate charities based on 20 standards related to governance, oversight, finances, measuring effectiveness, fundraising and informational materials. The volume of public inquiries and self-request help identify which charities they will evaluate. The participation by charities in the evaluation process is volunteer; evaluation reports are freely available in their website; BBB includes a disclaim in their website about agencies that do not provide the needed information for the evaluation; BBB has a voluntary seal of quality program—agencies who meet the 20 standards and want to display the seal in their adds need to pay an annual fee from $1,000 to $15,000 depending on agency's annual expenses.

[82] See footnote 4 for info about Charity Navigator

responsible for “learning” at headquarters. This person leads a task group of program staff from different departments charged with monitoring the use of M&E guidelines, providing support in the development of M&E strategies and identifying learning opportunities across programs.

Explanations presented by two respondents from agencies with annual expenditures of more than 100 million dollars indicated that they have informal M&E structure was that each program/division in their agencies has M&E systems tailored to their particular kind of programs. For the third very large agency in this group, its representative noted that the M&E function in his agency was created and they are still trying to get things organized.

Even though there were more organizations considered small and medium-sized reporting less structured M&E systems in their agencies, these results indicate that more resources in general in an organization will not necessarily mean that it will have a stronger M&E structure. Responses to this survey question also shed light into another interesting aspect: having a M&E structure does not necessarily ensure that the M&E systems will be comprehensive. For instance, the representative of a very large agency affiliated to InterAction and dedicated to relief and emergency assistance indicated having M&E staff at central and country level, but their work is basically concerned to monitor whether the inputs are received by local agencies and used according to what was stipulated in the original plan. No strategies are yet in place to assess outcomes of the programs they fund, even though some efforts in this direction are currently underway.

M&E staff

The survey asked respondents to report on the approximate number of staff in

their agencies who had at least 50 percent of their time dedicated to monitoring and evaluation functions. There were a few problems with the answers provided since a considerable number of respondents did not know the answer to the question, especially for the staff located outside the agencies' headquarters office. Among the 50 respondents, four (8 percent) did not know the number of M&E staff at headquarters; 17 (34 percent) did not know how many M&E staff were located in their regional office(s); and 16 (32 percent) didn't know this information for their country offices. Also, six respondents (12 percent) indicated they did not have regional offices and five (10 percent) said they did not have any country offices. As a result, the numbers provided in Table 1 reflect the answers by a subset of respondents (46 regarding central office, 27 on regional offices, and 29 on country offices), and should be interpreted with caution. The results are presented as medians to diminish distortion. In order to facilitate interpretation, the results are also presented stratified by the size of the organizations. The range of the answers indicated in the number between parentheses below the medians provides an idea of how wide the variation across agencies was.

Table 18. Median number of INGO staff dedicating at least 50 percent of their time to M&E functions by size of the agencies

Location	Median number (and range) of M&E staff by agency size				
	small	medium	large	very large	all
Central Office	0 (0 to 2)	1.25 (0 to 4)	2 (0 to 6)	2 (0 to 10)	1 (0 to 10)
Regional Offices	0 (0 to 0)	0 (0 to 2)	2 (0 to 5)	2 (0 to 6)	0 (0 to 6)
Country Offices	0 (0 to 8)	2 (0 to 15)	8 (0 to 30)	40 (2 to 57)	5 (0 to 57)

Table 1 suggests, as one would expect, that in general, the number of people with at least 50 percent of time dedicated to M&E is directly related to the size of the

organization; larger agencies will have more staff dedicated to M&E activities than smaller agencies.

Systems to learn from evaluations and ensure their quality

To tackle these two very important issues, survey respondents were asked to indicate whether their agencies (i) had system to collect evaluation reports about any of their efforts, (ii) conducted syntheses of findings from multiple evaluations, and (iii) commissioned or developed meta-evaluations of their evaluations.

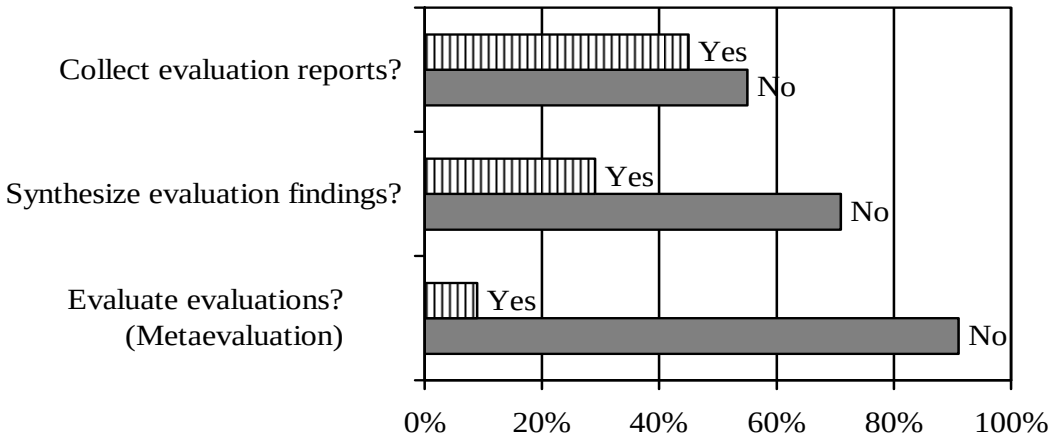

Figure 2. Proportion of INGOs with systems in place for collecting evaluation reports, synthesizing evaluation findings, and evaluating their own evaluations

As indicated in Figure 2, the general situation is quite worrisome. Less than one half of the agencies (44 percent) appear to have a basic system to collect evaluation reports global of programs, projects or other efforts sponsored or implemented by them. Eleven of those agencies stressed that their systems have been in place for sometime and are working quite well. They indicated that the reports are accessible within the agency and, in a few cases, also to the general public such as in

the USAID's electronic clearing house. One of those agencies indicated that their system even allows tracking certain indicators across projects although they are not using it for that yet. Two respondents mentioned that their agencies have just started such a system and one of them is not sure if the submission of the evaluation reports conducted in the different areas of the organization to the central office is mandatory. Four respondents indicated that even thought their agencies do not have such a system at the moment, they have started discussions to start one.

Only 14 respondents (28 percent) indicated that they periodically synthesize and share findings from multiple evaluation reports within their agencies. One of the respondents said that those syntheses are sectorial (e.g., health, education, etc,) and are shared not only internally, but also with the external public. Another agency indicated that syntheses of groups of projects are only made if there is a request by donors. In explaining how this process is conducted in his agency, a respondent provided the following account:

> We give this a qualified "yes" because it is not periodic but rather we synthesize our findings when we believe we have sufficient evaluation information to cull and produce a meta-study or review. This is typically done for a particular program and not across programs although we do not discount that for the future. For example we are currently considering a [organization]-wide initiative to synthesize what we have learned from various programs regarding the integration of life skills into youth development programs. An example of a program-specific meta-study is the one published by [agency's name] in 2006 on job placement programs in Latin America. This meta-study encompassed the findings from 6 external evaluations of [agency's name] projects in 6 countries. It was published in English and Spanish and disseminated through a direct mailing, through several regional websites and 2 international events.

Three agencies indicated they are synthesizing and sharing evaluation findings but not on a systematic way. One of them indicated that even though they do not have a standard procedure or process for synthesizing evaluation reports, on a case-by-case basis, they will share evaluations, lessons learned and best practices with varying sector specific agencies, working groups, and collaborating partners. The second agency mentioned that they have only done it once. The third agency reported that they present evaluation findings from their programs at conferences and workshops.

Conducting meta-evaluations seems to be clearly a virtue of a few agencies. Only four indicated that they have done any formal meta-evaluation of their evaluations. One way of conducting meta-evaluation mentioned by two agencies was to assess the extent to which the evaluations conducted have adhered to the M&E standards proposed by the agencies. The respondent from one of the agencies answering no to this question indicated that, even though they would like to conduct such meta-evaluations, the donor agencies they work with (e.g., USAID) usually does not provide funding for such activities.

Level of independence of INGOs' evaluations

External evaluations by trained evaluators of aid interventions supported and developed by INGOs are not common. As Figure 3 shows, more than one half of the respondents (54 percent) said that less than one-third of their programs or other major efforts completed during the recent past were evaluated by external professionals with evaluation expertise. Only 16 percent of respondents indicated more than two-thirds of their efforts are evaluated by external evaluators.

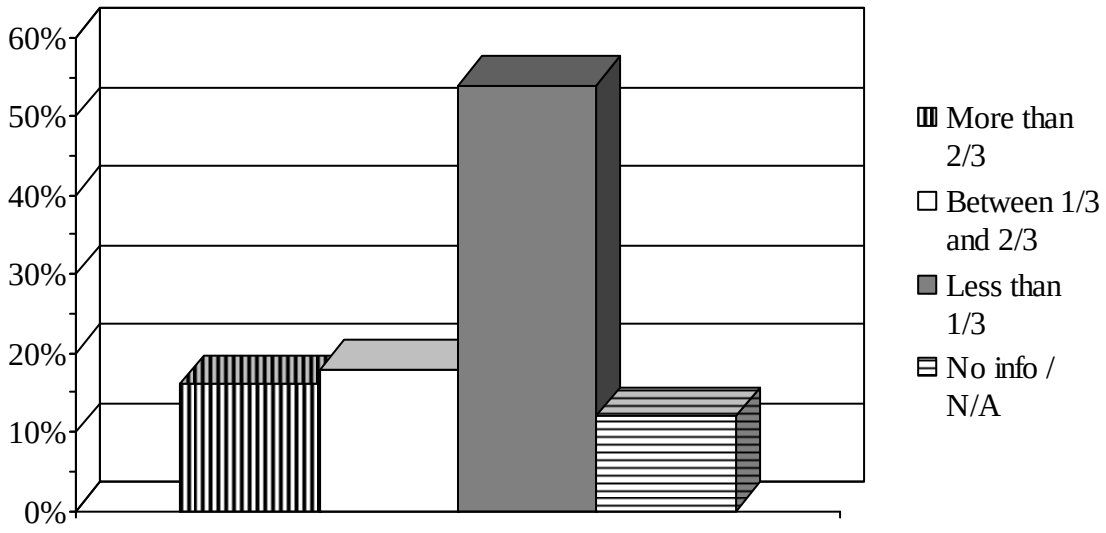

Figure 3. Proportion of aid interventions supported by INGOs that have been reported to be externally evaluated in the recent past

At least two agencies mentioned that the greatest challenge to have more external evaluations conducted in their agencies is the limitation of funding; one of them stated that her agency do not possess the resources to support external evaluations without donor support.

It is worth noting that in the analysis of the 14 documents from the INGOs that responded to the survey explaining their M&E standards, guidelines, or policies, only three explicitly mentioned the importance of including, as part of their M&E systems, external evaluation of the agencies' programs.

APPENDIX C

List of regional and national evaluation associations, networks or societies[83]

1. African Evaluation Association – www.afrea.org/
2. American Evaluation Association – www.eval.org/
3. Aotearoa New Zealand Evaluation Association (ANZEA) – www.anzea.org.nz/
4. Australasian Evaluation Society – www.aes.asn.au/
5. Bangladesh Evaluation Forum – Syed Tamjid ur Rahman, tamjidr@bangla.net
6. Benin – Maxime Dahoun, mdahoun@yahoo.fr, or francois-corneille.kedowide@iucn.org
7. Botswana Evaluation Association – Kathleen Letshabo, letshabo@mopipi.ub.bw
8. Brazilian Evaluation Network – www.avaliabrasil.org.br
9. Burkina Faso M&E Network – Marie-Michelle Ouedraogo, mmouedraogo@unicef.org
10. Burundi Evaluation Network – Deogration Buzingo, buzingdeo@yahoo.com
11. Cameroon Development Evaluation Association (CaDEA) – Debazou Y. Yantio, yantio@hotmail.com
12. Canadian Evaluation Society – www.evaluationcanada.ca/
13. Cape Verdi – Francisco Fernandes Tavares, Francisco.Tavares@ine.gov.cv or chicotavares@yahoo.com.br
14. Central American Evaluation Association – Johanna Fernandez, johannaf@cariari.ucr.ac.cr
15. China – Chaoying Chen, chenzhaoying@ncste.org
16. Columbian Network for Monitoring and Evaluation – Consuelo Ballesteros, consocds@colomsat.net.co or Daniel Gomez, dgomez@uniandes.edu.co
17. Danish Evaluation Society – www.danskevalueringsselskab.dk
18. Dutch Evaluation Society – www.videnet.nl/
19. Egyptian Evaluation Society – Ashraf Bakr, picardm@care.org
20. Eritrean National Evaluation Association – Bissrat Ghebru, bissratgk@asmara.uoa.edu.er or Woldeyesus Elisa, dolab@eol.com.er
21. Ethiopian Evaluation Association – Gizachew Bizayehu, medac2@telecom.net.et
22. European Evaluation Society – www.europeanevaluation.org/
23. Finnish Evaluation Society – www.finnishevaluationsociety.net/
24. French Evaluation Society – www.sfe.asso.fr/
25. German Evaluation Society – www.degeval.de/
26. Ghana Evaluation Network (GEN) – Charles Nornoo, cnornoo@internetghana.com or bds@africanus.com
27. Ghana Evaluators Association – isodec@ghana.com
28. Indian Evaluation Network – Suresh Balakrishnan, sbalakrishnan@vsnl.net
29. International Program Evaluation Network (Russia & Newly Independent States) – http://www.eval-net.org/
30. Israeli Association for Program Evaluation – www.iape.org.il

[83] Source: IOCE website (http:ioce.net), November 2006

31. Italian Evaluation Society – www.valutazioneitaliana.it/
32. Japan Evaluation Society – www.idcj.or.jp/jes/index_english.htm
33. Kenya Evaluation Association – Gitonga Mburugu Nkanata, gitonga35@avu.org or Karen Odhiambo, karenodhiamboo@hotmail.com
34. Korean Evaluation Association – Sung Sam Oh, edulove@kkucc.konkuk.ac.kr
35. Latin American and Caribbean Programme for Strengthening the Regional Capacity for Evaluation of Rural Poverty Alleviation Projects (PREVAL) – www.preval.org/
36. Madagascar – Barbara Rakotoniaina, Barbara.Rakotoniaina@caramail.com or Dominique Wendling, Aea.evaluation@netcourrier.com or aea.evaluation@yahoo.fr
37. Malawi Network of Evaluators – John Kadzandira, csrbasis@malawi.net or csr@malawi.net
38. Malaysian Evaluation Society – www.mes.org.my
39. Mauritanian M&E Network – Ba Tall Oumoul, oktconsult@yahoo.fr or Mohammeden Fall, mfall@unicef.org
40. Namibia Monitoring Evaluation and Research Network – Bob Hochobeb, bhochobeb@unam.na
41. Nepal M&E Forum – Suman Rai, srai@icimod.org.np
42. Niger Network of Monitoring and Evaluation (ReNSE) – www.pnud.ne/rense/
43. Nigeria – Adam Suleiman, adamsuleiman@yahoo.com (interested in establishing a network)
44. Pakistan Evaluation Network (PEN) – pen.dmne@yahoo.com
45. Perú Network for Monitoring and Evaluation – Emma Rotondo, erotondo@terra.com.pe
46. Polish Evaluation Society – www.pte.org.pl/obszary/enginfo.htm
47. Quebec Society for Program Evaluation – www.sqep.ca
48. Red de evaluacion de America Latina y el Caribe (ReLAC) – www.relacweb.org
49. Rwanda Network for Monitoring and Evaluation – James Mugaju, imungaju@unicef.org or Philippe Ngango Gafishi, pgafishi@yahoo.fr
50. Senegalese Network of M&E – Eric d Muynck, eric.de.muynck@undp.org
51. South African Evaluation Network (SAENet) – www.afrea.org/webs/southafrica/
52. Spanish Evaluation Society – Carmen Vélez Méndez, carmenvelez@idr.es or Carlos Román del Río, carlosroman@idr.es
53. Spanish Evaluation Society – www.sociedadevaluacion.org
54. Sri Lanka Evaluation Association (SLEvA) – www.nsf.ac.lk/sleva/
55. Swedish Evaluation Society – www.svuf.nu
56. Swiss Evaluation Society – www.seval.ch/de/index.cfm
57. Thailand Evaluation Network – Rangsun Wiboonuppatum, rangsun@hotmail.com
58. Uganda Evaluation Association (UEA) – www.ueas.org
59. United Kingdom Evaluation Society – www.evaluation.org.uk
60. Wallonian Society for Evaluation (Belgium) – www.prospeval.org
61. Zambia Evaluation Association (ZEA) – Greenwell Mukwavi, gmukwavi@zamtel.zm or twizamtc@zamnet.zm
62. Zimbabwe Evaluation Society – Mufunani Tungu Khosa, mkhosa@mandel.co.zw or emkhosa@ecoweb.co.zw

APPENDIX D

Survey Invitation Letter

Dear [CEO's name],

I am contacting you to invite your organization to participate in an important study sponsored by InterAction on monitoring and evaluation principles and practice within InterAction member agencies. As you may know, in September 2005 the InterAction Board approved the "Position Statement on Demonstrating NGO Effectiveness."[84] Based on that statement, each InterAction member agency commits to:

1. Articulate its own criteria for success in bringing about meaningful changes in people's lives, in terms of its mission and major program goals.
2. Regularly evaluate its progress towards such success.
3. Mainstream relevant monitoring and evaluation in agency policy, systems and culture.
4. Allocate adequate financial and human resources for its strategic evaluation needs.
5. Collaborate with partners and stakeholders in developing mutually satisfying goals, methods, and indicators for project and program activities.

In 2006 InterAction's Evaluation and Program Effectiveness Working Group (EPEWG) reviewed the InterAction Standards and is proposing the inclusion of several new standards, each with interpretive guidance, specifically related to monitoring and evaluation (M&E). The intention is to help each agency discover ways to strengthen its own evaluation policies and practices in order to promote not only program quality, but also accountability for results and institutional learning. If adopted by the Board, these revised standards will also be included as part of the Self-Certification Plus[85] process in the future.

InterAction is currently conducting a study to help introduce members to the new InterAction Standards related to M&E, gather ideas for future improvement of the standards, and identify consistent, defensible, and practical ways to gather evidence of member compliance with the standards. I am inviting your organization, along with nineteen others representing a range of InterAction members, to participate in the first round of the study. Your responses and feedback will be essential to improve the survey instrument before we try to reach all 165 InterAction members. This study is separate from the recently sent survey to update member profiles

The study is being developed free of charge to InterAction by Thomaz Chianca, Doctoral Associate at the Evaluation Center, Western Michigan University (WMU). Data from the study will also be used for his PhD dissertation: "The Practice and Principles of Evaluation in International Non-Governmental Organizations (INGOs)," supervised by Michael Scriven (Associate Director, WMU Evaluation Center), Jim Rugh (DME Coordinator, CARE International), and Paul Clements (Associate Professor, WMU Dept. of Political Sciences).

The survey has 26 questions (many are multiple-choice) and should not take more then 40 minutes

[84] www.interaction.org/files.cgi/5031_Position_Statement_on_demonstrating_NGO_effectiveness.pdf
[85] www.interaction.org/files.cgi/4981_SCP_Guidelines.pdf

to be completed. There are three optional ways to respond to the survey:

(i) Online: Here is the link to the web-based version of the survey: http://www.zoomerang.com/survey.zgi?p=WEB2265YR5B4M9. The online survey permits saving the answers and returning on a later time to finish it. Before responding to the online survey, please print and review the attached 4-page handout with the new M&E standards, which will be necessary to answer questions 5 to 13.

(ii) MS Word file: The survey can be completed using the attached MS Word document, and submitted either by e-mail attachment to thomaz.chianca@wmich.edu or by regular mail to Thomaz Chianca, 4405 Ellsworth Hall, Kalamazoo, MI, 49008-5237.

(iii) Phone: Please contact Thomaz Chianca by phone (269 387-3207) or by email (above) to schedule a phone interview.

While Dr. Chianca's final report will be made available to the InterAction community, your answers will be kept confidential—InterAction will only have access to aggregated results (i.e., only the researchers will see individual responses.) If you have any questions or concerns about the study please contact Thomaz Chianca (see contact info above).

Please complete the survey by April 5, 2007. If this deadline is not feasible for your organization, please contact Thomaz so he can make different arrangements.

Thank you very much for your support!

Sincerely,

Sam Worthington, CEO

InterAction

APPENDIX E

Survey on Evaluation Principles and Practice in INGOs

Identification of Agency Responding to the Survey

1. Agency:

2. Contact person:

3. Email:

4. Phone:

Interaction Standards and Evidence of Members' Compliance

The 2-page handout sent to you with the invitation letter to participate in this survey presents the proposed new InterAction Standards related to monitoring and evaluation (M&E), their interpretive guidance, and gives examples of what constitutes evidence of compliance to be included in future versions of the Self-Certification Plus process. *Please print and read that handout before answering questions 5 to 13*! Use as much space as necessary to answer the descriptive questions below—the lines will automatically expand as the text is typed in.

5. Do any of the proposed new M&E standards or interpretive guidance appear NOT relevant to people in your organization?

( ) Yes ( ) No

6. If yes, please indicate which one(s) (use the numbers at the beginning of each paragraph) and explain why:

---

7. Are there any standards or interpretive guidance that, though relevant, may be unclear to people in your organization?

( ) Yes ( ) No

8. If yes, please identify them by including the number at the beginning of each paragraph, and include any suggestions people in your organization might have to improve them:

---

9. If your organization was asked by InterAction to present evidence of the level of compliance with these standards today, what documents or other evidence would your organization be able to provide? (If you are willing to share any of them with the researcher, please send them by email to thomaz.chianca@wmich.edu or mail them to Thomaz Chianca, 4405 Ellsworth Hall, Kalamazoo, MI, 49008-5237.)

---

10. Though it might be difficult to provide such evidence currently, as your organization invests more time and energy in building capacities, taking action to demonstrate effectiveness, and developing reporting systems, do people in your organization think it will be possible to provide more evidence of compliance to InterAction's standards in the near future?

( ) Yes ( ) No

11. If yes, what would this evidence consist of?

---

12. Would it be beneficial for people in your organization to receive technical assistance in any of the areas related to the InterAction standards?

( ) Yes ( ) No

13. If yes, what form of assistance might your organization need?

Basic Information about Monitoring and Evaluation (M&E) in your Agency

Use as much space as necessary to answer the descriptive questions below—the lines will automatically expand as the text is typed in.

14. What is your agency's M&E structure? (CHECK ALL that apply)

( ) Central M&E unit at U.S. or other global headquarters
( ) Regional M&E staff (e.g., for Latin America, Africa, Asia)
( ) Country M&E staff
( ) No formal M&E structure
( ) Other

15. If there is no formal structure in place or if you chose "other" in the above options, please briefly explain how the M&E functions in your agency are managed.

---

16. Approximately how many employees in your agency have at least 50% or more of their time dedicated to M&E functions at these levels:

___ Central Office ___ Regional Offices ___ Country Offices ___ I don't know

17. Has your agency developed its own M&E policies, guidelines and/or standards?

( ) Yes ( ) No

18. If yes, please send a copy of the document(s) describing such policies/guidelines/standards to Thomaz Chianca by e-mail or regular mail (see contact info on question 9) or, if document(s) is(are) available online, provide the website(s): ***http://***

19. Has your agency adopted specific M&E policies, guidelines and/or standards developed by other agencies (e.g., USAID, DAC, Joint Committee for Ed. Evaluation, AEA, AfrEA, etc)?

( ) Yes ( ) No

20. If yes, please attach a copy of the document(s) describing such policies/guidelines/standards, or, if document(s) is(are) available online, provide the website(s): ***http://***

21. Does your agency have a global system for collecting reports of evaluations of programs, projects or other efforts supported/implemented by your agency?

( ) Yes ( ) No

22. Does your agency periodically synthesize and share the findings of evaluation reports? (See for example CARE's MEGA reports at http://pqdl.care.org/pv_obj_cache/pv_obj_id_3F0964E46D34E15DD78EB2D03DF10200)

( ) Yes ( ) No

23. Does your agency have a system for meta-evaluating the quality of the evaluations?

( ) Yes ( ) No

24. If you answered yes to any of the past three questions (21, 22 or 23), please describe the system and/or send a copy of document(s) describing it to Thomaz Chianca by e-mail or regular mail (see contact info on question 9), or, if document(s) is (are) available online, provide the website(s): ***http://***

25. Roughly, what percentage of projects, programs or other major efforts supported/implemented by your agency that were completed during the recent past were evaluated by external professionals with evaluation expertise?

( ) Less than one-third
( ) Between one-third and two-thirds
( ) More than two-thirds
( ) I don't know

26. Please include below any additional comments about the work in your agency related to M&E or suggestions to improve this survey.

APPENDIX F

Study Protocol Approval by the WMU Human Subjects Institutional Review Board

Date: February 8, 2007

To: Michael Scriven, Principal Investigator
Thomaz Chianca, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: Approval not needed for protocol 06-12-17

This letter will serve as confirmation that your project "The Practice and Principles of Evaluation in International Non-Governmental Organizations" has been reviewed by the Human Subjects Institutional Review Board (HSIRB). Based on that review, the HSIRB has determined that approval is not required for you to conduct this project because as revised on February 7, 2007, you will study organizations and not individuals. Thank you for your concerns about protecting the rights and welfare of human subjects.

A copy of your protocol and a copy of this letter will be maintained in the HSIRB files.

APPENDIX G

New/changed InterAction Monitoring and Evaluation (M&E) standards proposed by the Evaluation and Program Effectiveness Working Group to InterAction's Standards Committee[86]

Note: The numbering used here indicates where these should be inserted in the present version of the InterAction INGO Standards on the InterAction website.

Standard related to M&E	**Comments / Interpretive Guidance / Examples of evidence[87]**
Insert the following standard under the Governance section	
2.6.4 The agency's Board shall ensure that the organization (i) articulates organization-wide criteria for success as defined by its vision, mission and major program goals; (ii) incorporates and practices regular, deliberate evaluative activities to determine achievements of program goals and mission fulfillment; (iii) mainstreams and utilizes monitoring and evaluation in the agency's policy, systems and culture; and (iv) allocates adequate financial and human resources for the organization's strategic evaluation needs.	*Note: This standard deals with policies for which an agency's board should be responsible. More details, including examples of evidence, are included with the more specific M&E standards, below.* The term 'mission' refers to an articulation of the agency's over-all purpose, related to how it will work towards its vision. 'Program goals' could include major approaches used by the agency to achieve its mission. The term 'regular' means a pre-determined interval, e.g. within the organization's strategic planning cycle, or any other decision-making timeframe adopted by the organization. The term 'mainstream' would involve the establishment of sound and comprehensive monitoring and evaluation systems, and their use by at least a majority of the agency's program units. The term 'strategic' here refers to corporate and agency-wide programs. *Proposed Evidence: Reports or other documents describing Board responsibilities and policies.*

[86] Chianca's note: This is the final version of the M&E standards and guidelines submitted by the EPEWG to InterAction's standards committee in October 2007.

[87] Examples of suggested evidence are intended to indicate types of data to be collected are not exhaustive and may not be applicable in all cases to each InterAction member.

Add the following standards to the Organizational Integrity section.	
3.8 The member organization shall be committed to full, honest, and accurate disclosure of relevant information concerning its goals, including criteria for objectively measuring progress and success of its programs, finances, and governance in achieving the goals.	*Proposed Evidence: In conjunction with M&E standards within the Program section, provide evidence that objective evaluations, including evaluations by external experts, have been conducted, and are assessable to relevant stakeholders. Note: This standard points to one of the purposes for an agency's M&E system – to objectively assess, and share with its public, progress of its programs in achieving goals.*
3.9 To inform its ongoing strategic planning process, a member organization shall incorporate a deliberate and intentional process of monitoring and evaluating the organization's progress toward achievement of its mission and major program goals.	• Each agency should have one or more explicit underlying hypothesis(es) or theory(ies) of change about how its activities will lead to desired changes. In other words, it should be able to articulate clear causal links between major program activities, impacts and mission. • The agency should ensure that valid and credible evaluations of its operations are conducted in accordance with the agency's strategic planning cycle. Such evaluations should be a complete assessment of the quality, value, and significance of the work done by the agency, always including an assessment of the progress made by the agency in achieving its mission and major goals. *Proposed Evidence: Documents outlining the process the organization went through to prepare its strategic plan, including a statement of its monitoring and evaluation requirements. Documentation that outlines the organization's monitoring and evaluation of its programs against its strategic plan, including the organization's established criteria for assessing progress against the strategic plan.*

Move existing Standard 7.1.9 to new section under Program, preferably labeling it 7.2.n, moving other standards down. *Until that is done this set of M&E standards are numbered 7.X.n.*	
7.X Monitoring and Evaluation 7.X.1 A member organization shall have a policy (or similar operative document) that defines how monitoring and evaluation are integrated within program / project management, as well as evidence that the policy is being adhered to.	The M&E procedures should address: (i) efficiency of the use of inputs, (ii) quality of processes, (iii) outcomes and impacts (positive, negative, intended, non-intended), (iv) the relationship of the positive impacts to the costs of achieving them, (v) reach, (vi) pertinence to the needs of the participants, (vii) post-project sustainability, and (viii) ethical practice. At both strategic program and project levels, evidence of progress and impacts should be captured through a valid and credible monitoring and evaluation system. While InterAction is not prescribing a standardized approach to be followed by all members, an agency's system should provide systematic information about the following key

Move existing Standard 7.1.9 to new section under Program, preferably labeling it 7.2.n, moving other standards down. *Until that is done this set of M&E standards are numbered 7.X.n.*	
	aspects of programs and projects implemented by IA members: • Positive changes, e.g. type and scope of benefits, whether material, human/social, organizational, civic, policy, governance, environmental, or other. Evidence of participants' satisfaction with such changes should be included. • Side effects, e.g., evaluation and documentation of positive and negative unintended outcomes/ impacts connected with the efforts. • Efficiency of delivery, e.g. timeframe for implementation; costs (monetary and non-monetary—e.g., opportunity, stress, time), compared to results obtained. • Reach, e.g. number of people, communities, organizations, regions, etc.; number of partnerships & alliances; and depth of poverty and/or marginalization of target populations. • Pertinence to needs, e.g., the extent to which the initiative's objectives and implemented strategies are directly connected to existing needs of targeted beneficiaries • Resources for sustainability, e.g. structural changes, commitment by participants to continue activities or benefits, new resources, external stakeholder support, enabling policy environment. • Post-project gains, e.g. sustainability, replication, expansion, policy change, etc. • Ethical practice, e.g., evidence that the means to produce the results/impacts adhere to relevant ethical standards *Proposed Evidence:* Agency's stated monitoring and evaluation policies, standards and guidelines. They could further include evidence of evaluations being conducted, a system for reviewing the quality of such evaluations, and use of the lessons learned from monitoring and evaluations to promote institutional learning and decision-making.

Move existing Standard 7.1.9 to new section under Program, preferably labeling it 7.2.n, moving other standards down. *Until that is done this set of M&E standards are numbered 7.X.n.*	
7.X.2 From the outset of program planning, a member organization shall collaborate with partners, clients/intended beneficiaries and other stakeholders in developing mutually satisfying goals, methods, and indicators for project and program activities and results.	InterAction members' program theory(ies) of change should involve active participation by communities or other constituencies, and should have clear policies and procedures to engage the active participation of communities and partners in program design, planning, monitoring, evaluation and learning. All InterAction member agencies should regularly assess the satisfaction of those they seek to serve. *Proposed Evidence: Relevant internal documents that demonstrate written protocols, policies and/or procedures that guide the organization's engagement of and dialog with partners, clients/intended beneficiaries and communities at point-of-service related to the development and design of project proposals, as well as to the evaluation of program impacts.*
7.X.3 A member organization shall assure that program and project budgets allocate adequate resources for monitoring, evaluation and institutional learning.	Though some donors have formulas calling for 5%-10% of a project's budget to be allocated for M&E, the amount required depends on the purpose of the project. If it is a pilot project that is testing a new intervention that will be multiplied at a larger scale if proven to be successful, its M&E plan should have more of a rigorous research focus and commensurate budget. If, on the other hand, all that is needed is to assess compliance with the project's planned objectives, the M&E system can be relatively less expensive.

Move existing Standard 7.1.9 to new section under Program, preferably labeling it 7.2.n, moving other standards down. *Until that is done this set of M&E standards are numbered 7.X.n.*	
7.X.4 An agency's planning, monitoring and evaluation system should draw on commonly accepted professional principles and standards in planning, monitoring and evaluating programs. These systems should take into account not only the defined organization-wide criteria for success toward achievement of its mission and program goals, but also basic components of sound, objective evaluations including, but not limited to, quality of process, intended and unintended outcomes and impact, costs, and sustainability.	InterAction does not prescribe particular evaluation methodologies, but it does call upon its members to be aware of the range of methods and approaches, and make informed choices as to which are most appropriate for the various projects and programs they implement. See, for example: American Evaluation Association (www.eval.org); the African Evaluation Association (www.afrea.org); ALNAP (www.alnap.org/), Action Aid International's Accountability, Learning and Planning System (ALPS) (http://www.actionaid.org/main.aspx?PageId=261); and CARE International's Design, Monitoring and Evaluation policies and standards (http://pqdl.care.org/). These and other relevant links will be provided on IA's website. In addition to generic M&E standards, IA members need to be guided by prevailing norms within sectors or sub-field(s) of development or relief. For example. several sectors have defined common standards, e.g., child sponsorship, humanitarian assistance and microfinance. *Proposed evidence: Documented reference to commonly accepted professional principles, standards and good practice used by the organization to guide monitoring and evaluation of its programs..* *Additional evidence could include how an agency gathers, reviews and synthesizes project monitoring and evaluation reports. Look for meta-evaluations that assess the quality of evaluation reports. And look for examples of post-project evaluations that summarize (1) lessons learned based on the findings from monitoring and evaluation, (2) how those lessons are being applied in subsequent programming, and (3) the processes for making these lessons accessible to all relevant stakeholders.*

It is proposed that the following summary M&E standards be included in Self-Certification Plus	
A member organization shall have a policy (or similar operative document) that defines how monitoring and evaluation are integrated within program / project management, and evidence that the policy is being adhered to.	*Proposed evidence: As noted in the guidance accompanying the M&E standards, during the SCP process gather and review materials summarizing the organization's guidelines and procedures for monitoring and evaluating the effective use of inputs, as well as material summarizing the organization's procedures for monitoring and evaluating the impact on program participants and measuring the effectiveness of these outcomes by factors relevant to the organization's work, and criteria for measuring it against the organization's strategic plan.*

REFERENCES

Adamchak, S. at al. (2000). *A Guide to Monitoring and Evaluating Adolescent Reproductive Health Programs*. Washington, DC : FOCUS on Young Adults.

Adams, J. (2003). *Monitoring and evaluation of rights based approaches*' The Newsletter of the International NGO Training and Research Centre. (23) 4. Retrieved on 08/21/07 at: http://www.intrac.org/docs.php/343/ontrac23.pdf.

AEA – American Evaluation Association (2003). *American Evaluation Association response to U. S. Department of Education notice of proposed priority, Federal Register RIN 1890-ZA00, November 4, 2003 "Scientifically based evaluation methods"*. Retrieved on 10/26/07 at: http://www.eval.org/doestatement.htm.

AEA – American Evaluation Association (2004). *American Evaluation Association Guiding Principles for Evaluators*. Retrieved on September 7, 2007 at: http://www.eval.org/Publications/GuidingPrinciplesPrintable.asp.

ALNAP – Active Learning Network for Accountability and Performance in Humanitarian Action (2005). *Assessing the quality of humanitarian evaluations: The ALNAP Quality Proforma 2005 (v. 02/03/05)*. London : ALNAP.

ALNAP – Active Learning Network for Accountability and Performance in Humanitarian Action (2006). *Evaluating humanitarian action using the OECD-DAC criteria: An ALNAP guide for humanitarian agencies*. London : Overseas Development Institute

ALNAP – Active Learning Network for Accountability and Performance in Humanitarian Action (2007). *ALNAP's Website*. Retrieved on 10/20/07 at: http://www.alnap.org/.

BBB – Better Business Bureau Wise Giving Alliance (2003). *Standards for Charity Accountability*. Better Business Bureau Wise Giving Alliance Website. Retrieved on September 20, 2007 at: http://www.give.org/standards/.

Binnendijk, A. (2001). *Results based management in the development co-operation agencies: A review of experience*. Paris : OECD/DAC Working Party on Aid Evaluation.

Brim, O.G., Jr. (1973). *Do We Know What We Are Doing*? In F.F. Heimann (ed.), *The Future of Foundations*, pp. 216-258. Englewood Cliffs, NJ : Prentice-Hall, Inc

Business dictionary (2007). *Definition of Economic Rate of Return*. Retrieved on 10/04/07 at: http://www.businessdictionary.com/definition/economic-rate-of-return-ERR.html.

Chapman, J. and Wameyo, A. (2001). *Monitoring and Evaluating Advocacy: A Scoping Study*. Retrieved on 08/21/07 at: http://www.preval.org/documentos/00545.pdf.

Chianca, T. (2006). *A critical view of interaction's position statement on demonstrating NGO effectiveness*. Retrieved on 10/24/07 at: http://interaction.org/library/detail.php?id=5009.

Clements, P., Chianca, T. and Sasaki, R. (2006) *Applying a Cost-Effectiveness Framework to Assess International Development Projects: A Case Study of the First Uttar Pradesh Sodic Lands Reclamation Project*. Manuscript submitted for publication.

Clements, P. (2005a) *Inventory of Evaluation Quality Assurance Systems*. Unpublished manuscript prepared for the United Nations Development Program, November 7, 2005.

Clements, P. (2005b) *Monitoring and evaluation for cost-effectiveness in development management*. Journal of MultiDisplinary Evaluation, (2)11-38. Retrieved on 02/21/08 at: http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/118/133.

Clements, P. (2007, November). *Reaction to paper: Taking the OECD/DAC evaluation criteria to the next level*. Presentation at the 2007 American Evaluation Association conference, Baltimore, Maryland.

Cracknell, B.E. (2000). *Evaluating Development Aid – Issues, Problems and Solutions*. London : Sage.

Commission on Foundations and Private Philanthropy (1970). *Foundations, Private Giving, and Public Policy: Report and Recommendations of the Commission on Foundations and Private Philanthropy*. Chicago, IL: University of Chicago Press.

Davidson, E.J. (2003). *Linking organizational learning to the bottom line: Methodological issues, challenges, and suggestions.* The *Psychologist-Manager Journal*, 6(1), 54-67.

Davidson, E.J. (2005). *Evaluation Methodology Basics: The nuts and Bolts of Sound Evaluation*. Thousand Oaks, CA : Sage.

Davidson, E.J. (2006). *The RCTs-only doctrine: Brakes on the acquisition of knowledge?* Journal of MultiDisplinary Evaluation, (5)iii-iv. Retrieved on 10/26/07 at: http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/35/45

Davies, R. & Dart, J. (2005). *The 'Most Significant Changes' (MSC) Technique: A guide to its use*. Retrieved on 10/06/07 at: http://www.mande.co.uk/docs/ccdb.htm.

USDE – United States Department of Education (2003). Scientifically Based Evaluation Methods: RIN 1890–ZA00. Federal Register, Vol. 68, No. 213, Tuesday, November 4, 2003, Notices, pp. 62445-47.

Donaldson, S. & Christie, C. (2005). *The 2004 Claremont debate: Lipsey vs. Scriven; Determining causality in program evaluation and applied research: Should experimental evidence be the gold standard?* Journal of MultiDisciplinary Evaluation, (3)60-77. Retrieved on 10/26/07 at: http://evaluation.wmich.edu/jmde/content/JMDE003content/PDFs%20JMDE%20003/5_The_2004_Claremont_Debate_Lipsey_vs_Scriven.pdf.

ECG – The Evaluation Cooperation Group (2007). ECGNet website. Retrieved on

10/06/07 at: https://wpqp1.adb.org/QuickPlace/ecg/Main.nsf/h_Toc/73ffb29010478ff348257290000f43a6.

EDEPO - Centre for the Evaluation of Development Policies (2007). Research Project: Income expectations, income risk. Centre for the Evaluation of Development Policies website. Retrieved on 09/28/07 at: http://www.ifs.org.uk/edepo/projects_ research.php?project_id=242.

EPEWG - Evaluation and Program Effectiveness Working Group (2005). *Position Statement on Demonstrating NGO Effectiveness*. Washington DC : InterAction Evaluation and Program Effectiveness Working Group.

EPEWG - Evaluation and Program Effectiveness Working Group (2007). *Terms of Reference for InterAction Evaluation and Program Effectiveness Working Group —draft as of July 20, 2007*. EPEWG internal document.

EuropeAid (2007). *Definition of Evaluation*. Retrieved on 10/13/07 at: http://ec.europa.eu/europeaid/how/evaluation/introduction/introduction_en.htm.

EuropeAid Cooperation Office (2005). Project / programme evaluations Guidelines for the evaluation team. Retrieved on 10/13/07 at: http://ec.europa.eu/europeaid/evaluation/methodology/guidelines/gbb_det_en.htm#02_06#02_06.

GEF – Global Environment Facility (2006). *The GEF Monitoring and Evaluation Policy*. Washington DC : Global Environment Facility.

Goldenberg, D.A. (2001). *Meta-Evaluation of Goal Achievement in CARE Projects: A Review of Findings and Methodological Lessons from CARE Final Evaluations, 1994-2000*. CARE USA Program Division. Retrieved on 01/31/07 at http://www.care.ca/libraries/dme/CARE%20Documents%20PDF/CARE%20MEGA%20Evaluation%20Synthesis%20Report.pdf.

Goldenberg, D.A. (2003). *Meta-Evaluation of Goal Achievement in CARE Projects: A Review of Findings and Methodological Lessons from CARE Final Evaluations, 2001-2002*. CARE USA Program Division. Retrieved on 01/31/07 at http://www.kcenter.com/phls/MEGA%202002.pdf

HACI – Hope for African Children Initiative (2003). *Monitoring and Evaluation Framework*. Nairobi, Kenya : HACI.

Hall, P.D. (2003). *A solution is a product in search of a problem: A History of Foundations and Evaluation Research*. Retrieved on 10/21/07 at: http://ksghome.harvard.edu/~phall/EVALUATION%20ESSAY.pdf.

IBRD (1972). *Bank Operations in Colombia—An Evaluation.* International Bank for Reconstruction and Development, Report No. Z-18, dated May 25, 1972.

IBRD (1972). *Operations Evaluation Report: Electric Power*. International Bank for Reconstruction and Development, Report No. Z-17, dated March 10, 1972.

ICCO – Interchurch Organization for Development Co-operation (2000). *Building Bridges in PME*. Zeist, The Netherlands : ICCO.

IDEAS – International Development Evaluation Association (2005). *President's Report 2004-2005 presented by Sulley Gariba, IDEAS President*. April 12, 2005. Ottawa, Canada : IDEAS.

IDEAS – International Development Evaluation Association (2007). *President's Report 2006-2007 presented by Dr Marie-Hélène Adrien, IDEAS President*. July 3rd, 2007. Ottawa, Canada : IDEAS.

InterAction – American Council for International Voluntary Action (2005). *Position Statement on Demonstrating NGO Effectiveness*. Washington, DC : The Working Group on Evaluation and Program Effectiveness. Retrieved on 10/30/07 at: http://interaction.org/files.cgi/5031_Position_Statement_on_demonstrating_NGO_effectiveness.pdf.

InterAction – American Council for International Voluntary Action (2007a). *InterAction website*. Retrieved on 08/16/07 at http://www.interaction.org/about/index.html.

InterAction – American Council for International Voluntary Action (2007b). *InterAction standards*. Revised on March 2007. Retrieved on 10/31/07 at: http://interaction.org/files.cgi/6014_PVOStandardsMarch2007.pdf.

Investopedia (2007). *Definition of Internal Rate of Return*. Retrieved on 10/04/07 at: http://www.investopedia.com/terms/i/irr.asp

IOCE – International Organisation for Cooperation in Evaluation (2007). IOCE website. Retrieved on 10/06/07 at: http://ioce.net/.

Joint Committee on Standards for Educational Evaluation (1994). *The Standards for Program Evaluation*. 2nd Edition. Thousand Oaks, CA : Sage.

J-PAL Abdul Latif Jameel Poverty Action Lab (2007). Abdul Latif Jameel Poverty Action Lab website. Retrieved on 09/29/07 at http://www.povertyactionlab.org/.

Kremer, M. (n.d.). *Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons*. Retrieved on 10/27/07 at: http://post.economics.harvard.edu/faculty/kremer/papers/Randomized_Evaluations.pdf.

Kruse et al. (1997). *Searching For Impact And Methods: NGO Evaluation Synthesis Study. A Report prepared for the OECD/DAC Expert Group on Evaluation.* Retrieved on 01/28/07 at http://www.valt.helsinki.fi/ids/ngo/.

Leading Edge Group (2007). *Evaluation Gap Update April 2007*. Center for Global Development website. Retrieved on 09/28/07 at: http://www.cgdev.org/section/initiatives/_active/evalgap/eupdate.

Marcelino, E. (2005). *Towards a Human Rights Based Approach to Evaluation: UNIFEM's Initial Experience*. Retrieved on 08/21/07 at: http://cfapp1-docs-public.undp.org/eo/evaldocs1/workshop/uneg/Human%20Rights%20Based%20Approach%20to%20Evaluation%20in%20UNIFEM,%20April%202005.doc

MDB – Multilateral Development Bank (n.d.). *Good practice standards for evaluation of MDB supported public sector operations*. Retrieved on 10/19/07 at: http://www.adb.org/Evaluation/wgec.pdf.

MDRC (2007). *MDRC website*. Retrieved on 09/28/07 at http://www.mdrc.org/.

NONIE – Network of Networks on Impact Evaluation (2007). *NONIE website*.

Retrieved on 09/29/07 at: http://www.worldbank.org/ieg/.

OECD – Organization for Economic Cooperation and Development (1992). *Development Assistance Manual: DAC Principles for Effective Aid*. Paris : OECD.

OECD – Organization for Economic Cooperation and Development (1998). *Review of the DAC Principles of Development Assistance*. Paris : OECD/DAC Working Party on Aid Evaluation.

OECD – Organization for Economic Cooperation and Development (1999). *Guidance for Evaluating Humanitarian Assistance in Complex Emergencies*. Paris : OECD/DAC Working Party on Aid Evaluation.

OEDC – Organization for Economic Cooperation and Development (2002). Glossary of Key Terms in Evaluation and Results Based Management. Paris : OECD/DAC Working Party on Aid Evaluation.

OECD – Organization for Economic Cooperation and Development (2006a). *DAC Criteria for Evaluating Development Assistance*. Retrieved on 09/16/07 at: http://www.oecd.org/dataoecd/15/21/39119068.pdf

OECD – Organization for Economic Cooperation and Development (2006b). *About OECD – History*. Retrieved on 10/05/07 at: http://www.oecd.org/pages/0,3417,en_36734052_36761863_1_1_1_1_1,00.html.

OECD – Organization for Economic Cooperation and Development (2006c). DAC Evaluation Quality Standards (for test phase application). Paris : DAC Network on Development Evaluation.

OECD – Organization for Economic Cooperation and Development (2007). *An approach to dac guidance for evaluating conflict prevention and peacebuilding activities*. Paris : DAC Network on Conflict, Peace and Development Co-operation & DAC Network on Development Evaluation.

OED – Operations Evaluation Department (2002). *Institutional Changes for Independent Evaluation at the World Bank - A Chronology (1970-2002)*. The

World Bank Operations Evaluation Department. Washington D.C.: The World Bank.

Ofir, Z. (2007, July 27). *Seeking Impact Evaluation case studies for a Very Important Purpose*. Message posted to the American Evaluation Association EVALTALK electronic mailing list, archived at http://bama.ua.edu/cgi-bin/wa?A1=ind0707d&L=evaltalk.

OIOS (2006). *Evaluation*. United Nations Office of Internal Oversight Services. Retrieved on 10/06/06 at http://www.un.org/depts/oios/evaluation.htm.

Paris Declaration (2005). *Paris Declaration on Aid Effectiveness: ownership, harmonisation, alignment, results and mutual accountability*. Paris : Organization for Economic Cooperation and Development.

Patel, M. (2001). *Human Rights as an Emerging Development Paradigm and some implications for Programme Planning, Monitoring and Evaluation*. Retrieved on 08/21/07 at: http://www.vpu.lt/socpedagogika/unicef/crrtm/Resource%20Documents/CRC_21%20HRAP%20ME%20~%20Mahesh%20Patel.doc.

Picciotto, R. (2006). *The Value of Evaluation Standards: A Comparative Assessment.* Journal of MultiDisciplinary Evaluation (3) pp. 30-59. Retrieved 10/01/07 at: http://evaluation.wmich.edu/jmde/content/JMDE003content/PDFs%20JMDE%20003/4_%20The_Value_of_Evaluation_Standards_A_Comparative_Assessment.pdf

Rockefeller Foundation, The (2007). *Designing a New Entity for Impact Evaluation: Meeting Report*. Bellagio, Italy : Rockefeller Foundation Bellagio Study and Conference Center.

Rugh, J. (2004). The CARE International Evaluation Standards. *New Directions for Evaluation*, 104, 79-88.

Russon, C. (2005). *Meta-Evaluation of Goal Achievement in CARE Projects: A Review of Findings and Methodological Lessons from CARE Final Evaluations, 2003-2004*. CARE USA Program Division. Retrieved on 01/31/07 at

http://pqdl.care.org/pv_obj_cache/pv_obj_id_3F0964E46D34E15DD78EB2D03DF1DFEFE1FC0200

Sasaki, R. (2006). *A Review of the History and the Current Practice of Aid Evaluation*. Journal of MultiDisciplinary Evaluation, (5) 55-88. Retrieved on 01/15/07 at http://evaluation.wmich.edu/jmde/content/JMDE005content/PDFs_JMDE_005/Review_of_Aid_Evaluation.pdf.

Savedoff, W. D. et al. (2006). *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Washington, D.C. : Center for Global Development. Retrieved on 01/31/07 at http://www.cgdev.org/content/publications/detail/7973.

Scriven, M. (1991). *Evaluation Thesaurus*. (4th ed.) Newbury Park, CA : Sage.

Scriven, M. (2000). *The Logic and Methodology of Checklists*. Retrieved on 11/02/07 at: http://www.wmich.edu/evalctr/checklists/papers/logic_methodology.pdf.

Scriven, M. (2007). *The Key Evaluation Checklist*. Retrieved on 09/07/07 at: http://www.wmich.edu/evalctr/checklists/kec_feb07.pdf.

SEGA – Scientific Evaluation for Global Action (2007). *Scientific Evaluation for Global Action website*. University of California, Berkley. Retrieved on 09/28/07 at: http://cider.berkeley.edu/sega/.

Suchman, E.A. (1967). *Evaluative Research: Principles and Practice in Public Service & Social Action Programs*. New York, NY: Russell Sage Foundation.

TEC – Tsunami Evaluation Coalition (2007). TEC's Website. Retrieved on 10/21/07 at: http://www.tsunami-evaluation.org/home.

Theis, J. (2003). *Rights-based Monitoring and Evaluation: A Discussion Paper*. Save the Children. Retrieved on 08/21/07 at: http://www.crin.org/docs/resources/publications/hrbap/RBA_monitoring_evaluation.pdf.

UNFPA – United Nations Population Fund (2007). UNFPA website. Retrieved on 10/24/07 at: http://www.unfpa.org/results/index.htm.

UNEG – United Nations Evaluation Group (2005a). *Standards for Evaluation in the UN System*. New York : United Nations.

UNEG – United Nations Evaluation Group (2005b). *Norms for Evaluation in the UN System*. New York : United Nations.

UNEG – United Nations Evaluation Group (2007). *The UN Evaluation Group Website*. Retrieved on 10/04/07 at: http://www.uneval.org/.

United Nations (2006). *The Millennium Development Goals Report 2006*. New York : United Nations.

United Nations Evaluation Group (2005). *Norms for Evaluation in the UN System*. Retrieved December 28, 2005 from http://www.uneval.org/docs/ACFFC9F.pdf.

USAID – United States Agency for International Development (2003). PVO Guidelines for Title II Emergency Food Proposals and Reporting. Retrieved on September 18, 2007 at: http://www.usaid.gov/our_work/humanitarian_assistance/ffp/emerg.htm.

USAID – United States Agency for International Development (2004a). *Functional Series 200 – Programming Policy; ADS 200 – Introduction to Programming Policy; 03/19/2004 Revision*. Retrieved on February 21, 2008 at: http://www.usaid.gov/policy/ads/200/200.pdf.

USAID – United States Agency for International Development (2004b). *Functional Series 200 – Programming Policy; ADS 203 – Assessing and Learning; 03/19/2004 Revision*. Retrieved on September 17, 2007 at: http://www.usaid.gov/policy/ads/200/203.pdf.

USAID – United States Agency for International Development (2005). *USAID Mechanism for Conducting Evaluations*. EvalWeb collaborative research on USAID programs and performance. Retrieved on September 17, 2007 at: http://evalweb.usaid.gov/resources/USAIDMECHCONDUCTEVAL10_05.pdf.

USAID – United States Agency for International Development (2006 a.). *Guidelines for Unsolicited Proposals and Reporting*. Office of U.S. Foreign Disaster

Assistance (OFDA). Retrieved on September 17, 2007 at: http://www.usaid.gov/our_work/humanitarian_assistance/disaster_assistance/resources/pdf/OFDA_Guidelines_Unsolicited_Proposals_Reporting.pdf.

USAID – United States Agency for International Development (2006 b.). *Final Evaluation Guidelines*. USAID/GH/HIDN/NUT Child Survival and Health Grants Projects. Retrieved on September 17, 2007 at: http://www.usaid.gov/our_work/global_health/home/Funding/cs_grants/cs_index.html.

USAID – United States Agency for International Development (2007). *Definitions of 'Evaluation'*. EvalWeb collaborative research on USAID programs and performance. Retrieved September 17, 2007 at: http://evalweb.usaid.gov/resources/definitions.cfm.

WGEC – Working Group on Evaluation Criteria and Ratings for Public Sector Evaluation (n.d.). *Good Practice Standards for Evaluation of MDB Supported Public Sector Operations*. Washington DC : Multilateral Development Bank (MDB), Evaluation Cooperation Group (ECG).

Willoughby, C. (2003). *First Experiments in Operations Evaluation: Roots, Hopes, and Gaps*. In: Grasso, P. G., Wasty, S. S. & Weaving, R. V. (2003). *World Bank Operations Evaluation Department: The First 30 Years*. Washington DC : The World Bank.

Windfuhr M. (2000). *Economic, Social and Cultural Rights and Development Cooperation*. in Frankovits, A. and Earle, P. (2000) *Working Together: The Human Rights Based Approach to Development Cooperation - Report of the NGO Workshop*. Part 1, p. 25. Retrieved on 08/21/07 at: http://www.humanrights.se/upload/files/2/Rapporter%20och%20seminariedok/sv-hr%20based%20devcoop.pdf.

World Bank, The (2006). *Website of the Independent Evaluation Group*. Retrieved on 12/28/06 at: http://www.worldbank.org/ieg/oed_approach.html.

World Bank, The (2007a). *PovertyNet website*. Retrieved on 10/05/07 at:

http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/0,,menuPK:336998~pagePK:149018~piPK:149093~theSitePK:336992,00.html.

World Bank, The (2007b). *Africa Impact Evaluation Initiative website.* Retrieved on 10/05/07 at: http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/AFRICAEXT/EXTIMPEVA/0,,menuPK:2620040~pagePK:64168427~piPK:64168435~theSitePK:2620018,00.html.

World Bank, The (2007c). *Sourcebook for evaluating global and regional partnership programs*. Washington, DC : Independent Evaluation Group—World Bank.