UNEG
United Nations Evaluation Group

Compendium of Evaluation Methods Reviewed - Volume 1

UNEG Working Group on Evaluation Methods

December 2020

Table of Contents

Foreword

The United Nations Evaluation Group (UNEG) aims to support the professionalization and capacities of its membership and safeguard and influence the quality of evaluation practice in the United Nations (UN) system. Strengthening the professional capacity of UN evaluators to design and manage evaluations is a central objective of UNEG's Strategy 2020 – 2024 which will help ensure that evaluations within the UN system produce knowledge and evidence that can be used to inform relevant, effective, sustainable UN support to Member States to achieve the Sustainable Development Goals.

I welcome the work of the UNEG Working Group on Evaluation Methods, which emerged organically over the course of monthly exchanges on good practice in evaluation design and conduct between evaluators from eleven UN evaluation entities. Presented here, their review of seven distinct evaluation methods and their applications to concrete UN evaluation contexts is a useful resource for UN evaluators seeking to expand the range and diversity of evaluation design in their evaluation work.

I would like to thank in particular Christine Fu and Henri van den Insert (UNHCR); Andrew Fyfe and Pietro Tornese (UNCDF); Carlos Tarazona (FAO); and Tina Tordjman-Nebe (UNDP) for setting up and managing the group and I look forward to continued work by UNEG in the coming months on this important topic.

Masahiro Igarashi
UNEG Chair
Rome, December 2020

Introduction

For UN evaluators, choosing appropriate approaches and methods are pre-requisites for credible, independent and useful evaluation exercises. Yet it is not always straightforward to know which method and approach best fits the evaluation questions stakeholders want answered or which take into consideration the time, data and budget constraints with which many of us are faced.

With this in mind, and recognising the broader objectives of the United Nations Evaluation Group's Strategy (UNEG) 2020 – 2024[1] to support continued professionalization of UN evaluators, a UNEG Working Group on Evaluation Methods was formed in 2019 which served as a forum for exchange of good practice on this topic. The Group focused both on the practicalities of applying the method in question to a concrete evaluation, but also the reasons why the method was selected in the first place, drawing on relevant literature from social science research as well as the current practice of international development evaluation.

Comprised of staff from eleven UNEG member agencies[2], the Group reviewed seven different evaluation methods, as well as a tool developed by BOND, a UK civil society network, which supports the choice of evaluation method depending on available resources and the evaluation question posed.[3] The Group also worked with a number of experts from established evaluation organisations whose role was to complement, challenge or bring additional information to the group. We acknowledge their contributions and are grateful for their support.

This document presents key highlights from our discussions, as well as the conceptual underpinnings of the choice of evaluation design. Recognizing the breadth and depth of guidance on evaluation design already available in both academic and professional literature, the purpose of this guidance document is more modest, bringing examples of interesting UN evaluation practices to a wider audience, with an additional focus on the underlying principles that led evaluation managers to choose the method they did. Conceived as the first in a series of written resources, the document will be updated as the work of the group continues. Hopefully, UNEG colleagues will be inspired to join the Group as its focus and content expand going forward.

The UNEG Methods Group coordinators

---

[1] www.unevaluation.org/UNEG_Strategy_2020-2024
[2] FAO, IOM, JIU, UN Environment Programme, UNCDF, UNFPA, UNHCR, UNICEF, UNITAR, UNODC and WIPO.
[3] For more information, see the guidance document by Dr. Barbara Befani:
https://www.bond.org.uk/sites/default/files/caem_narrative_final_14oct16.pdf

Definitions

Defining Evaluation Methods

An important initial step in this work was to define what is meant by evaluation methods and how these might differ from other terms such as evaluation approaches, paradigms, techniques etc. The group considered how to organize the system of evaluation methods it was interested in covering recognizing that, with eight sessions planned, it would only scratch the surface of methodological approaches and techniques currently being applied by evaluators across the full array of the UN's work.

The *UNEG Norms and Standards*[4] is the foundational document for evaluation in the UN and provided a starting point in defining evaluation methods. Specifically, they state:

- 'Evaluation must be credible. Credibility is grounded in independence, impartiality and a rigorous methodology' (Norm 3 "Credibility")

- 'Evaluation methodologies must be sufficiently rigorous such that the evaluation responds to the scope and objectives, is designed to answer evaluation questions and leads to a complete, fair and unbiased assessment' (Standard 4.5 "Methodology")

- 'Methodologies should be chosen with a clear intent to provide credible answers to the evaluation questions' (Standard 4.5 "Methodology")

- 'Methodologies provide what information should be collected, from which source(s) it should be collected, for what purpose it should be collected and how the collected data will be analyzed in order to answer the evaluation questions. Importantly, the methodology should not be confused with the data collection strategy' (Standard 4.5 "Methodology")

To guide its work, the Group followed a rather loose distinction between evaluation approaches, methods and techniques, as these definitions are subject to debate and are used rather inter-changeably[5] by evaluation practitioners:

- An **evaluation approach** relates to the type of question that needs answering and what type of evaluation study needs to be put in place overall to answer that question. Examples of evaluation approaches could include experimental, quasi-experimental or non-experimental evaluation, or evaluations with a focus on stakeholder participation, on being culturally responsive, or with a focus on the UN's normative work (recognising that these categories are not always mutually exclusive).

- An **evaluation method** is a more specific term, largely following UNEG Standard 4.5, and relates to the process of how data will be gathered overall, and how it will be analyzed in order to answer evaluation questions.

- An **evaluation technique** relates more to the specific contents of the data collection strategy and could include qualitative or quantitative techniques, such as key informant interview, document review or Likert rankings of qualitative responses.

---

[4] www.unevaluation.org/2016-Norms-and-Standards.
[5] The Better Evaluation Website, for example, describes contribution analysis, outcome mapping etc. as approaches rather than methods. The Bond UK tool, on the other hand, draws on concepts in social science research to define an evaluation approach describing, for example, the ontological, epistemological or normative foundations of an evaluation method.

Setting Our Universe of Methods

The primary objective for this work was for it to be useful to other UN evaluation practitioners working across the full range of UN system work: development, human rights and peace and security, humanitarian work. It also needed to support Member States working to introduce the norms of the UN system into their countries' public policy evaluation systems. Therefore, in defining a focus for the year, the group identified the following categories – again not mutually exclusive - of evaluation methods to cover:

- Methods helping to assess the causal relationship between what agencies deliver and the results that they claim at different levels of their impact pathways;

- Methods that allow partners and beneficiaries to fully participate in evaluation exercises, bearing in mind the SDG central objective to *Leave No One Behind*;

- Methods that ensure that evaluations take account of, and are sensitive to, the cultural contexts in which UN interventions take place; and

- Methods that allow evaluation of the UN's policy and normative work helping its Member States achieve relevant SDGs in the areas of public policy.

Methods Discussed in 2019-2020

The Group organised learning events on the application of six evaluation methods and a session which looked more broadly into culturally-responsive evaluation (see Table 1 below).

Table 1: List of Methods and Evaluations Reviewed by the Working Group (2019-2020)

#	Method	Agency	Evaluations Reviewed	Approach	Technique
1	Evaluability assessment	UNFPA, FAO	Thematic Evaluations of UNFPA and FAO work on gender	Assessing evaluation readiness and support to evaluation design	Storyline techniques; critical moment reflection ; ToC approaches
2	Synthesis and meta-analysis	FAO, IFAD	Annual Report on Results and Impact of IFAD Operations (IFAD) Synthesis of biodiversity-related project evaluations in Ecuador (FAO) Synthesis of pesticide-related project evaluations in Africa (FAO)	Evaluation and evidence synthesis	Qualitative (Synthesis), Quantitative (meta-analysis)
3	Contribution analysis	UNCDF	Mobile Money for the Poor programme evaluation	Non-experimental looking at contribution to change in highly dynamic and complex market and policy systems in LDCs	Mixed Methods including country contextual studies, case studies, quantitative analysis
4	Qualitative comparative analysis	GEF/UNDP; UNICEF	Joint Impact Evaluation of GEF/UNDP Support to Protected Areas	Non-experimental, systematic cross-case comparison, focusing on	Various qualitative and quantitative data collection and analysis techniques (necessity

				intersections rather than correlations of factors	analysis, subset sufficiency analysis, Boolean minimisations)
			Evaluation of UNICEF support to Child Protection Systems Strengthening		
5	Randomised controlled trials	UNHCR, UNICEF	Impact evaluation of an education programme in a humanitarian setting	Counterfactual Impact Evaluation using treatment, control and comparison groups	Various techniques related to data collection and analysis within the treatment and control groups
6	Outcome harvesting/ outcome evidencing	FAO	Evaluation of a FAO Food Security Intervention	Non-experimental, participatory evaluation with action learning at its core	Participatory Rural Appraisal techniques involving programme 'Change Makers' as central interlocutors
7	Culturally-responsive evaluation methods such as 'most significant change', or 'outcome harvesting'	Various	Various examples mentioned in the discussion	An evaluation approach that emphasizes the centrality of culture and the 'lived experience' of (often-marginalized) communities in evaluation design and conduct.	Storytelling, relational interviews, community history techniques in support of 'co-produced' evaluation findings

1. Methods supporting evaluation design: Evaluability assessment, theory of change and storyline approaches

Defined as the extent to which an intervention 'can be evaluated in a credible and reliable fashion' (OECD DAC, 2012), **evaluability assessment** can be used at various stages of programme and evaluation design. At the:

- Design and approval stage of an intervention with a view to ensuring well – articulated impact pathways as part of the intervention's broader theory of change;

- Start of an intervention when monitoring and evaluation systems are being set up to include appropriate performance indicators, data collection methods and tools to generate information on progress on performance at different levels of the results chain; and

- Design or inception stage of an evaluation, when evaluators are seeking to describe very clearly the intervention to be evaluated and how the intervention will be evaluated concretely.

A **critical moment reflection** methodology is the concept underpinning the use of **storyline approaches** along with the reconstruction of the "intervention logic" in the scoping of complex evaluations. This approach is particularly useful in thematic evaluations that include in their scope several components of an organization's work: with different units of analysis, and two or more results frameworks (covering different programmatic periods in time).

Critical moment reflection methodology can be applied as part of evaluation reference group discussions to understand the strategic storyline under the scope of the evaluation. These discussions are organized with evaluation reference groups at global, regional and country levels. Furthermore, this exercise offers an opportunity to visualize the organization's milestones, gaps and potential shifts in positioning, including the identification of internal and external drivers of change which will help to understand the decision-making process throughout the time period covered by the evaluation.

Advantages of these methods

Evaluability assessment

- A participatory evaluability assessment approach can be a useful tool in the initial stages of evaluation as it provides clarity to stakeholders on what will be evaluated and in what way, before the data collection phase of the evaluation begins.

- Embedding evaluability assessment into the design of an intervention can ensure appropriate data collection systems are embedded at the start of the intervention. This can support more sophisticated evaluation design when the evaluation is due, such as more rigorous pre- and post-analysis of key indicators, or changes in indicators for control and treatment groups in evaluation designs with counterfactuals.

Storyline

- Allowing members of evaluation reference groups to identify timeline turning points that are meaningful to them can help articulate key questions for the attention of evaluators. These can relate to shifts in strategic direction, the introduction of new policies, or new implementation

instruments, for example. This can support and contribute to the setting up of an appropriate strategy for data collection and analysis.

Disadvantages of these methods

Evaluability assessment

There is risk in thinking too rigidly that evaluability assessment as a one-size fits all tool. Programme designers and evaluators should think of evaluability assessment as a set of tools that can be deployed at different points in the programme and evaluation management cycles, when clarity is needed on what evaluation questions should be posed and what data is necessary to answer these questions.

### *Critical moment reflection methodology*

Critical moment reflection methodology takes more time than an interview. It has the advantage of offering an immediate picture of the storyline to the organization and its key stakeholders which can be useful both in evaluation and programmatic settings.

Storyline approaches

While the critical moment reflection methodology applied as part of discussions on reconstructing the 'intervention' storyline usually leads to a consensus, in some cases differing opinions may emerge regarding critical moments by participants.

Applications in the UN

Evaluation of FAO's work on gender
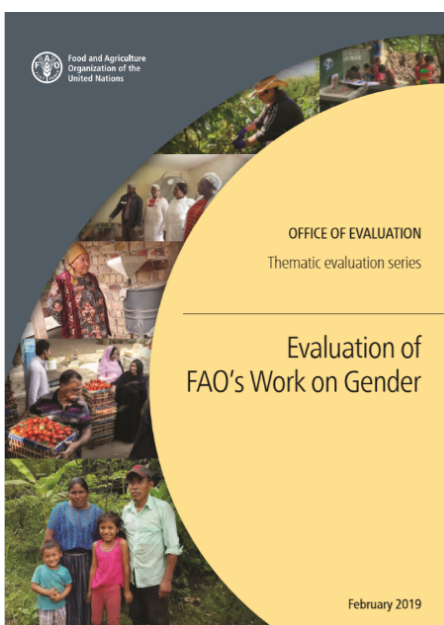
The FAO Office of Evaluation conducted an evaluability assessment (EA) of FAO's work on gender responsiveness, six months before a planned evaluation in early 2018. The EA reviewed available data and stakeholder expectations to propose a set of evaluation questions that would meet the needs of the Organization in its planning of future work.

Carried out by a team of five, the exercise relied on extensive document analysis and key informant interviews conducted with key stakeholders at FAO Headquarters and five regional offices.

The EA confirmed that evaluating FAO's work on gender across the time period of two Strategic Frameworks was feasible, and it contributed to the formulation of a theory of change for the evaluation.

In addition to the gender evaluation, the method has been used in at least two major evaluations in FAO:

- [Evaluation of FAO's contribution to the reduction of rural poverty through Strategic Programme 3](#)
- [Evaluation of FAO's contribution to inclusive and efficient agricultural and food systems (SO4)](#)

Thematic evaluation of UNFPA support to gender equality and women's empowerment (2012 – 2020)
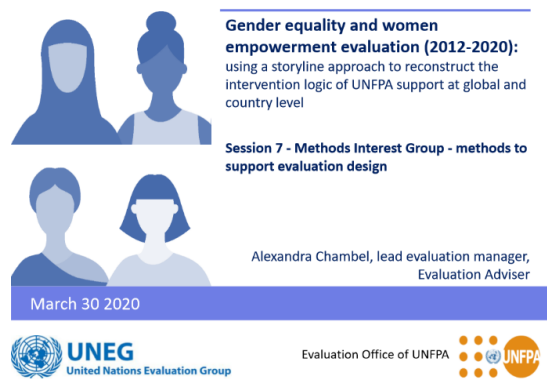
The storyline approach was used in the thematic evaluation of UNFPA support to gender equality and women's empowerment across development and humanitarian settings (2012-2020). The purpose of the evaluation was to inform the strategic direction, gaps and opportunities for UNFPA programmatic work in the area of gender; provide learning inputs to inform the strategic positioning of UNFPA in this area; and reflect the changing development environment, including UN reform, and alignment with the 2030 development agenda.

The storyline approach was used to help reconstruct the theory of change of UNFPA's support to gender equality and women's empowerment across both the development and humanitarian spheres over two corporate programming periods, mapping key moments, strategic milestones and turning points at global, regional and country levels. The analysis not only focused on the results pathway - the 'what' - but also explored the 'why?' by asking 'how did changes happen?'. This required making explicit key assumptions underpinning the theory about how change happened, including which actors UNFPA was collaborating with to advance change.

The analysis was conducted over three strategic planning periods at three different levels: an evaluation reference group at global level; a national reference group for each in-country case study; and an evaluation reference group for the regional case study (see example below).

UNFPA LACRO STORYLINE: CRITICAL MOMENTS, 2010 - 2020
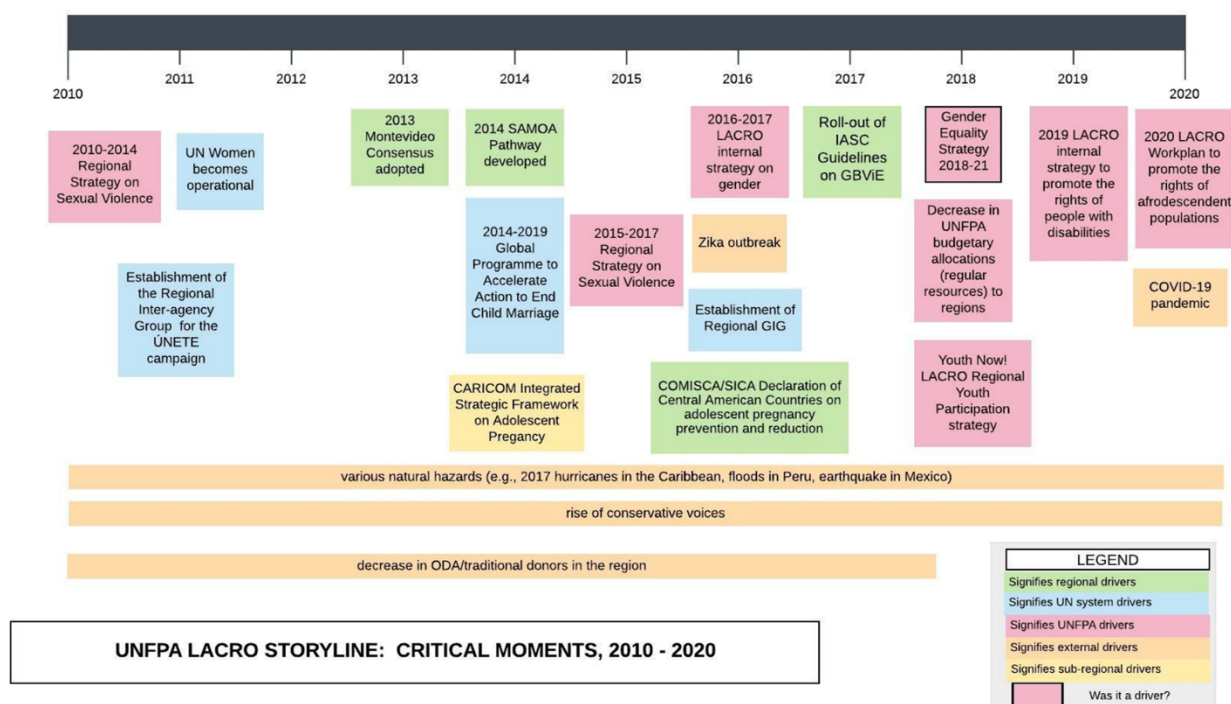
While the primary contours of UNFPA support to gender equality and women's empowerment are broadly similar over time, the Strategic Plans also demonstrate changes in nature of the results and programming approaches. The reconstruction of the intervention logic and the storyline approaches by the evaluation have allowed for a comprehensive and systematic overview of these changes. Other evaluations conducted by UNFPA Evaluation Office using similar storyline approaches include:

- Joint Evaluation of the UNFPA-UNICEF Joint Programme on the Abandonment of Female Genital Mutilation: Accelerating Change Phase I and II (2008–2017)
- Thematic Evaluation of UNFPA support to the prevention, response to and elimination of gender-based violence and harmful practices (2012-2017)

2. Synthesis and Meta-analysis
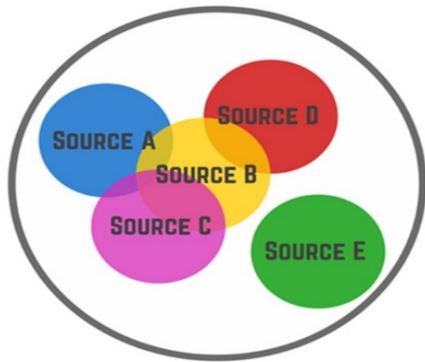
Evaluation synthesis[6] attempts to combine existing evaluation knowledge and findings relevant to a topic in the form of new knowledge products to support better dissemination of the results of evaluations and help in the design of new evaluations. It addresses the challenge of "information overload", delivering products that distil relevant evidence for decision-making, and are conducted following a well-structured process. Meta-analysis[7] is a statistical method allowing systematic analysis of quantitative data to answer evaluation questions.

Advantages

- Savings in direct and indirect costs (both for the evaluator and for the groups being evaluated).
- Focuses on analyzing structural/common factors of success/failure.
- Increases perception of robustness as it builds on previous findings, instead of starting from scratch.

Disadvantages

- Restricted access to/availability of suitable evaluations (i.e. those that have assessed the same subject, in similar contexts and with the same objectives) may affect the feasibility/quality of a synthesis.
- Resources and time available for synthesis work.
- Systematic review methods can be complex (such as comprehensive search methods and meta-analysis).

Suitability

This method has several possible applications.

- Can be applied to inform the design of a new evaluation: A synthesis of previous evaluations' findings on the theme and/or geographic region subject to evaluation can inform discussions on the objectives and scope of a new evaluation.
- Can supplement the evidence gathered by an evaluation: A synthesis of previous evaluations' findings on the theme and/or geographic region subject to evaluation can be used for triangulation purposes and/or fill in gaps in evaluative evidence.
- Can avoid the need for a new evaluation: When sufficient evidence exits, a synthesis can satisfy the information needs of decision-makers, without the need for additional primary data collection.

The method should not be used in the following cases:

- When there is not enough evidence available for synthesis.
- When users expect primary data as main source of evaluative evidence.

---

[6] https://www.betterevaluation.org/en/rainbow_framework/synthesise/synthesise_across_evaluations
[7] https://www.betterevaluation.org/en/evaluation-options/meta-analysis

Applications in the UN

There are several examples of application in the UN.

- FAO used synthesis to distil lessons common to a set of projects dealing with similar objectives & themes and undertaken in similar contexts and periods. This includes biodiversity programmes in Ecuador and pesticide programmes in Africa, among others.
- IFAD produced annual synthesis reports based on an analysis of past evaluations and their ratings.
- UNESCO and partners (NORAD, World Bank and UNICEF) produced a synthesis report to explore evidence gaps and summarize what works for advancing equality and inclusion in education.
- WFP has produced a synthesis of WFP's country portfolio evaluations in Africa (2016-2018) to inform new policies and country strategic plans; and a synthesis of four evaluations of the impact of WFP Programmes on Nutrition in humanitarian contexts in the Sahel, to identify lessons to improve effectiveness.
- UNODC produces bi-annual meta-synthesis reports focused on the recommendations formulated by evaluations of UNODC's projects and programmes.

Challenges

The evaluations collected for synthesis were often diverse asking different types of evaluation questions, and reporting on projects, programmes and activities using varied methodologies and indicators. Furthermore, evaluations rarely reported on impact or effect sizes using a standard impact evaluation methodology, and when they did, they used a variety of measures to report on outcomes. This variation makes the synthesis of evaluation findings on outcomes and impact complex and challenging. In addition, the timeframe and resources for synthesis were sometimes limited, which led to the decision to select a smaller sample and/or exclude evidence published by researchers and other units of UN organizations and elsewhere.

Lessons Learned

Decision makers in UN agencies and at country level appreciated the perceived robustness of the methodology and the strategic messages of synthesis reports.

Evaluation Offices in UN agencies have started to make a greater use of synthesis and there is the potential to increase their use especially in system-wide evaluations (see UNESCO and partners' example). However, this calls for more robust methodologies (such as meta-analysis) and the inclusion of broader sources of evidence. In the future, UN agencies could systematically compare the findings of its synthesis reports with wider research evidence on the same theme, or more critically appraise the methodological quality of individual evaluations being used in synthesis work.

3. Contribution Analysis

Initially developed by John Mayne, contribution analysis (CA) is an approach for assessing causal questions and inferring causality in real-life program evaluations.[8] It offers a step-by-step approach designed to help policymakers arrive at conclusions about the contribution their program has made (or is currently making) to development outcomes.

Six steps are taken to produce a credible contribution story:

1 Set out the attribution problem
Determine the **specific questions** being addressed. Has the Program influenced the observed result? Has the program made an important contribution to the observed result? What role did the intervention play?

2 Develop ToC and risks
Develop the **results chain** describing how the program is supposed to work. Based on that, develop the **theory of change** upon which the program is based. Include assumptions, risks, external influences

3 Gather evidence
First use **existing evidence** to test the ToC. What evidence is currently available about the occurrence of the various results? What about assumptions, risks, other factors?

4 Assemble and assess
Assemble the contribution story: is it reasonable to assume that the program has contributed to the observed outcomes? Then **assess it**. How credible is the story? Do results validate the results chain?

5 Seek out more evidence
Having identified where the contribution story is less credible, gather **additional primary or secondary data** to augment the evidence

6 Revise the story
Contribution analysis argues that **a reasonable causal claim** can be made if: (1) there is a reasoned ToC; (2) activities were implemented as set out in ToC; (3) the ToC is supported by evidence on observed results and underlying assumptions

Contribution Analysis

! ***Potential issues:*** *Reducing uncertainty; Inferring causality; Confirming ToCs*

Source: Authors' elaboration from Mayne (2008) and Betterevaluation.org

These steps are often part of an iterative approach to building the argument for claiming that the intervention made a contribution (or not), and exploring why or why not.

CA argues that if one can i) verify or confirm a theory of change with empirical evidence—that is, verify that the steps and assumptions in the intervention theory of change were realized in practice, and ii) account for other major influencing factors— then it is reasonable to conclude that the intervention in question has made a difference, i.e. was a contributory cause for the outcome.

The theory of change provides the framework for the argument that the intervention is making a difference, and the analysis identifies weaknesses in the argument and hence where evidence for strengthening such claims is most needed.

Advantages

The essential value of CA is that it offers an approach designed to reduce uncertainty about the contribution the intervention is making to the observed results through an increased understanding of why the observed results have occurred (or not), and the roles played by the intervention and other internal and external

---

[8] Adapted from https://www.betterevaluation.org/en/plan/approach/contribution_analysis

factors. It can also underpin a more rigorous approach to monitoring and reporting results during programme implementation, particularly in cases where there are potentially multiple drivers of change in complex social and economic systems, which can help programme specialists report the results of their work more accurately during implementation.

Disadvantages

Contribution analysis helps confirm or revise a theory of change; it is not intended to be used to uncover and display an implicit or inexplicit theory of change. The report from a CA is not definitive proof, but rather provides evidence and a line of reasoning from which we can draw a plausible conclusion that, within some level of confidence, the program has made an important contribution to the documented results.

Some challenges might arise when using CA with regards to:

- Reducing uncertainty about the contribution the intervention is making to the observed results.
- Inferring causality in real-life program evaluations.
- Confirming or revising a programme's theory of change – including its logic model.

In addition, CA is "agnostic" in terms of which evidence matters, since its key is to test alternative explanations. Alternatively, it can become prone to confirmation bias, thus falling into the typical theory-based evaluation problem of collecting evidence to corroborate one's prior ideas only. Evidence needs to be collected for both the main analysis and alternative explanations, using multiple lines.

Applications in the UN

MOBILE MONEY FOR THE POOR
MID-TERM EVALUATION
DECEMBER 2018

G:ENESIS
UNLOCKING VALUE

UN CDF
Unlocking Public and Private
Finance for the Poor

Contribution analysis was used for the mid-term evaluation of UNCDF's Mobile Money for the Poor (MM4P) programme. UNCDF is the UN's capital development agency focused on two main areas of work: local development finance and financial inclusion. UNCDF works in the LDCs aiming to support the unlocking of public and private capital for the poor.

The "Mobile Money for the Poor (MM4P)" programme supports different actors working to develop digital (mobile) financial services in multiple countries simultaneously. The programme is an example of a broader market systems approach to reducing poverty with the aim to strengthen the core supporting functions and rules within financial markets in a way that ultimately improves the terms of participation of the poor in the real economy in developing countries.

With over 1 billion adults still with no access to financial services worldwide, financial inclusion can be thought of as a system: (1) micro level – the different types of local level banks/financial services providers that work with the poor; (2) meso level – the different types of organizations that support the creation and continuation of markets in financial services; (3) macro-level or the regulators, supervisors of the system that provide licenses to market participants. The evaluation

challenge comes from the fact that the intervention is being made in a complex, dynamic and non-linear system.

MM4P systems Approach (source UNCDF)

To understand the relevance and performance of the programme, the evaluation team adopted a theory-based approach, rooted in the programme theory of change. Market development programmes like MM4P seek to stimulate change at multiple levels simultaneously and a theory-based approach provides a coherent framework within which different parts of the causal chain can be observed and empirically tested.

The approach also provides a means of understanding why and how things work - the complexity of market systems means that it is necessary to understand how a programme has interacted with other factors to achieve change, as attribution of market development to one programme or intervention (and the formulation of a counterfactual) is seldom possible.

Challenges

The evaluation team encountered two challenges:

- Limited engagement with market actors that did not receive programme support, thereby weakening the validity of the contribution argument (this speaks to the available budget for the evaluation); and
- Inability to conduct an independent survey with partner organisations.

The UNDP Independent Evaluation Office, which assesses the quality of all UNCDF evaluation reports, found that the CA approach provided a coherent framework where change could be evaluated at different levels, examining also how the programme has contributed to change within a complex market system. It also found that the evaluation could have provided more information around sustainability at the regulator level and that evaluators could have spent more time assessing the M&E aspects of the programme.

However, findings of the evaluation were found to be important, logically articulated, balanced and based on an adequate analysis of evidence. The use of the theory of change was judged as making evaluative arguments "particularly incisive".

Lessons Learned

The case of MM4P confirmed the importance of theory-based design when assessing the results of interventions supporting complex, dynamic, non-linear market and policy systems. In order to be able to assemble a credible contribution story, the quality of secondary data – and their range and ambition – become fundamental. Finally, being a real-world evaluation, it is important to be as robust and as rigorous as possible, but at the same time humble in what can be achieved.

4. Qualitative Comparative Analysis

Qualitative Comparative Analysis (QCA) is a method for systematic cross-case comparison, aiming to help users understand which combinations of factors are likely to influence an outcome. Compared to other case-based methods, QCA's added value is its ability to compare case-based information systematically, leading to a replicable (rigorous) generalisation of case-specific findings, which is normally considered an advantage of quantitative / variable-based / statistical methods.

Advantages

- QCA does not require a large number of cases in order to be applied (it works with small, medium, and large case numbers and is ideal for the 5-30 range). QCA retains some of the "thickness", richness or complexity of case-based in-depth information. It is, therefore, particularly appropriate when countries (or UN country office operations) are the unit of analysis.
- QCA is strong on construct validity and conceptual clarity: it forces evaluation teams to start from whatever theory is available and requires high conceptual precision (particularly when calibrating conditions). It exposes both theoretical and empirical limitations of knowledge, by maintaining a constant dialogue between data and theory.
- QCA is a rigorous qualitative method for assessing impact and synthesising information, answering questions about necessity and sufficiency of the intervention and other factors. Relations, explanations and generalisations emerge (which do not necessarily appear with other methods) that focus on intersections (a specific combination of factors) rather than correlations of factors. QCA procedures are potentially fully transparent and replicable, and also reliable / robust because they lead to the same findings when repeated under the same conditions.

Disadvantages

- QCA usually presents a steep learning curve – integrating the required expertise in evaluation teams is not always easy. Close collaboration between a subject matter expert and QCA expert is needed.
- There is usually some degree of unpredictability in the time and resources needed, depending on the number of iterations needed to refine the causal combinations that make sense according to theory and experience.
- Despite best efforts, sometimes findings are either too complex to be interpreted, or simple but counterintuitive and therefore equally difficult to interpret. Communicating findings to non-technical audiences can be a challenge.

Suitability

QCA is most suitable when:

- There are one or several theories of change and want to assess how different factors combine in different setting to produce the outcome. QCA requires conceptualising cases as combinations of characteristics (e.g. UN intervention plus political and cultural factors) that are suspected to causally influence an outcome (e.g. desired social change).
- Trying to find patterns among more than five but fewer than 30 case studies. QCA is ideal for this "no-man's land" where there are too few cases for statistical methods and too many for traditional case studies. It can also be used on more than 30 cases where the main purpose of the analysis is to test combinations of causal pathways rather than statistical correlations.

QCA cannot be used to analyse single cases or where there is missing data. The latter is an important limitation as data quality and availability typically differs substantially between cases. In QCA, if data is missing on a factor for one case, the entire factor needs to be deleted. So using QCA requires some creativity and willingness to rely on standard protocols to collate existing data and integrate missing data with primary data collection.

Applications in the UN

GEF IEO / UNDP IEO Joint Impact Evaluation of GEF Support to Protected Areas and Protected Area Systems

From 1991 to 2014, the Global Environment Facility (GEF) provided $3.4 billion in grants to 618 projects in 137 countries to help protect almost 2.8 million km^2 of the world's non-marine ecosystems. The evaluation used a theory-based, mixed-methods approach to assess the GEF's contributions to reducing biodiversity loss in these protected areas.

QCA was used to test causal combinations for intervention outcomes at two geographical scales: the sub-national protected area level and the national protected area system level. For the protected area level, 30 cases (27 successful / 3 unsuccessful) were used in the analysis; at the national level, 8 cases from 7 countries (5 successful / 3 unsuccessful) were used. Both analyses included cases with and without GEF support to assess if GEF funding was one of the factors that made a difference in outcomes. Results were triangulated with data from portfolio-wide document analysis, multiple regression analysis, quasi-experimental geospatial analysis and other methods to formulate the over-all evaluation findings.
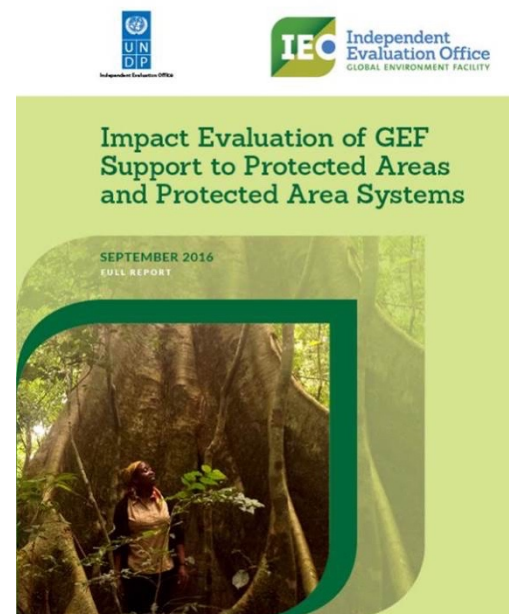
UNICEF Evaluation of Child Protection Systems Strengthening (CPSS)

CPSS has been one of UNICEF's two key corporate approaches to preventing and responding to violence, abuse, neglect and exploitation of children since 2008. UNICEF has made significant investments in developing, operationalising and implementing CPSS across more than 120 country offices.

In 2018, a global thematic evaluation assessed the relevance, effectiveness, efficiency and sustainability of this approach for learning, course correction and accountability. The evaluation included 24 case studies across the globe.

QCA was used to identify the most effective combination of UNICEF investments for child protection systems strengthening. These were: policy advocacy and advice; evidence and research; leveraging public

financial resources; and not paying for services and infrastructure. Investments in national capacity made the biggest difference by themselves, as a single investment.

Challenges

GEF IEO / UNDP IEO

- Need for very specific definitions/ criteria for presence ("1")/absence ("0") of outcomes and conditions understood by all involved so that scoring is comparable
- Results can be very sensitive to changes in scoring or causal recipes tested
- Need for complete data and a minimum level of detail for all cases across all factors
- Iterative nature can be time-consuming

UNICEF

- Getting to the zeros and ones: A lot of underlying qualitative data needed (the data collection protocol included 146 questions).
- Communicating the results: Multiple and sometimes seemingly contradictory analyses (which are all valid), including necessity analysis and subset sufficiency analysis, Boolean minimizations for both positive and negative outcomes.

Lessons Learned

- Conducting QCA cannot easily be added to the toolkit of every UN evaluator. Bringing in an evaluation methodologist is recommended.
- The learning curve is steep, but it pays off -- especially where high credibility is a requirement for an evaluation and multiple case studies are involved. QCA is usually considered systematic and rigorous by stakeholders.

5. Randomised Controlled Trials and Quasi-Experimental Designs

Randomized Controlled Trials (RCT)[9] and quasi-experimental designs[10] are widely used for impact evaluations that require evidence of causal attribution. An RCT is an experimental form of impact evaluation design in which the population receiving a programme or intervention (treatment group) is chosen at random alongside a control group from the same eligible population. If the eligible population is large enough, the results for the control group represent a "counterfactual" for the results observed in the treatment group (i.e. what would have happened to the treatment group in the absence of the intervention). RCTs test the extent to which planned impacts have been achieved through an intervention and are a powerful tool to answer questions of causality.

In a quasi-experimental design, participants are not randomly allocated to either a treatment or control group. The treatment group is either formed by self-selection or top-down selection (e.g. a policy maker), or a combination of the two. Quasi-experimental designs use statistical methods to construct a comparison group that serves as a counterfactual for the treatment group when randomisation is not practical. There are several techniques to identify a comparison group, such as statistical matching (e.g. propensity score matching), difference-in-difference, regression-discontinuity design or non-equivalent groups.

Advantages

- RCTs are a powerful way to provide evidence on the efficacy of an intervention by limiting biases that affect the internal validity of the findings.
- The random assignment of units allows programme implementers to determine whether observed effects are a result of the intervention and not anything else.
- Quasi-experimental designs are useful when pre-selection and randomization of groups are not possible, and can still provide credible evidence on intervention effectiveness.
- Quasi-experimental designs also include a comparison group which needs to provide a credible estimate of what would have happened to the treatment group in the absence of the intervention

Disadvantages

- Randomized allocation of people to the intervention vs. control group can be ethically questionable and logistically challenging in some cases.
- RCTs are expensive, can require lots of data collection and can be time-consuming (resulting in the loss of relevance as an intervention will have moved on by the time the trial is published).
- The reliability of an RCT is easily affected by the sample size (and the size of the standard errors). The smaller the sample size, the less statistical power the study has to detect the difference between an intervention and control group.
- Compliance and attrition are also important threats to the validity of RCTs, where systematic biases are introduced when participants in any group exit the sample. This is why validation of results through mixed-methods are critical, particularly in humanitarian and development settings.

---

[9] https://www.betterevaluation.org/en/plan/approach/rct
[10] https://www.betterevaluation.org/en/resources/guide/quasi-experimental_design_and_methods

- In quasi-experimental designs it can sometimes be challenging to devise credible comparison groups, while validity often requires comparison between groups in multiple sites which can be difficult to manage.
- RCTs and quasi-experimental evaluation designs are not appropriate for every type of evaluation question: they are best deployed to answer questions around 'what' has changed thanks to an intervention; they are less able to answer questions around 'how' or 'why' the changes have occurred.

Suitability

RCT and quasi-experimental designs are good when:

- The purpose of the evaluation is to determine whether an intervention caused the desired outcome(s) and evidence of causal attribution is needed.
- The programme is being implemented at a large scale, is innovative in testing the efficacy of a new approach, and when the programme you're evaluating is well-established with no major ongoing revisions.

RCT and quasi-experimental designs are not as useful when:

- The programme is small (<20 villages/schools) because the sample size is not large enough to detect a statistically significant effect size.
- The programme is constantly undergoing major changes (unless the evaluation is about those changes.

Applications in the UN

The example reviewed was of an impact evaluation of activities implemented under the Humanitarian Education Accelerator (HEA) conducted in partnership between UNICEF, UNHCR and DFID. A RCT was conducted by the American Research Institutes (AIR) from April 2017 – September 2018 in Kakuma refugee camp in Kenya to evaluate school attendance and learning and psychosocial outcomes of marginalized girls living in the camp and surrounding host communities at risk of dropping out of school. AIR also used mixed – method techniques (in-depth interviews, focus group discussions, Life History, participatory methods)) as an important means to triangulate findings and explain why some of the features from the impact evaluation were being observed.

Challenges

- Survey sampling and data collection delays: Locating study participants in the camps and communities was difficult. It was important to have back-up lists of households in addition to the control group.

- Attrition: It was challenging to track students because of high mobility.
- Study contamination: Ensuring that control group members did not receive the intervention during activity implementation.
- Distrust towards data collectors: Work with people already residing in humanitarian setting.
- Insecure environments: Remain flexible and have back-up plans if situation changes.

Lessons Learned

Where possible, and depending on the evaluation question to be answered, an impact evaluation should use mixed methods, using qualitative methods to complement and triangulate quantitative findings. Qualitative data helps to inform programme design and test assumptions of the theory of change. Including a process evaluation as part of an impact evaluation can yield valuable information on the quality of implementation. Evaluations should make use of local capacity and people on the ground in order to mitigate trust issues. Evaluations should also remain flexible and make use of back-up plans in insecure or volatile environments.

6. Outcome Harvesting/Outcome Evidencing

Outcome evidencing (Paz-Ybarnegaray, R., & Douthwaite, B., 2017) is a method that can be used in evaluations to unpack complex impact pathways in a participatory manner. It is similar to outcome harvesting, theory of change, and some participatory rural appraisal (PRA) techniques (e.g. fishbone analysis, problem tree, log frame, Venn diagram). Key interlocuters for outcome evidencing are the 'Change Agents' of the project/programme.

Key process steps to applying outcome evidencing (Source: FAO)
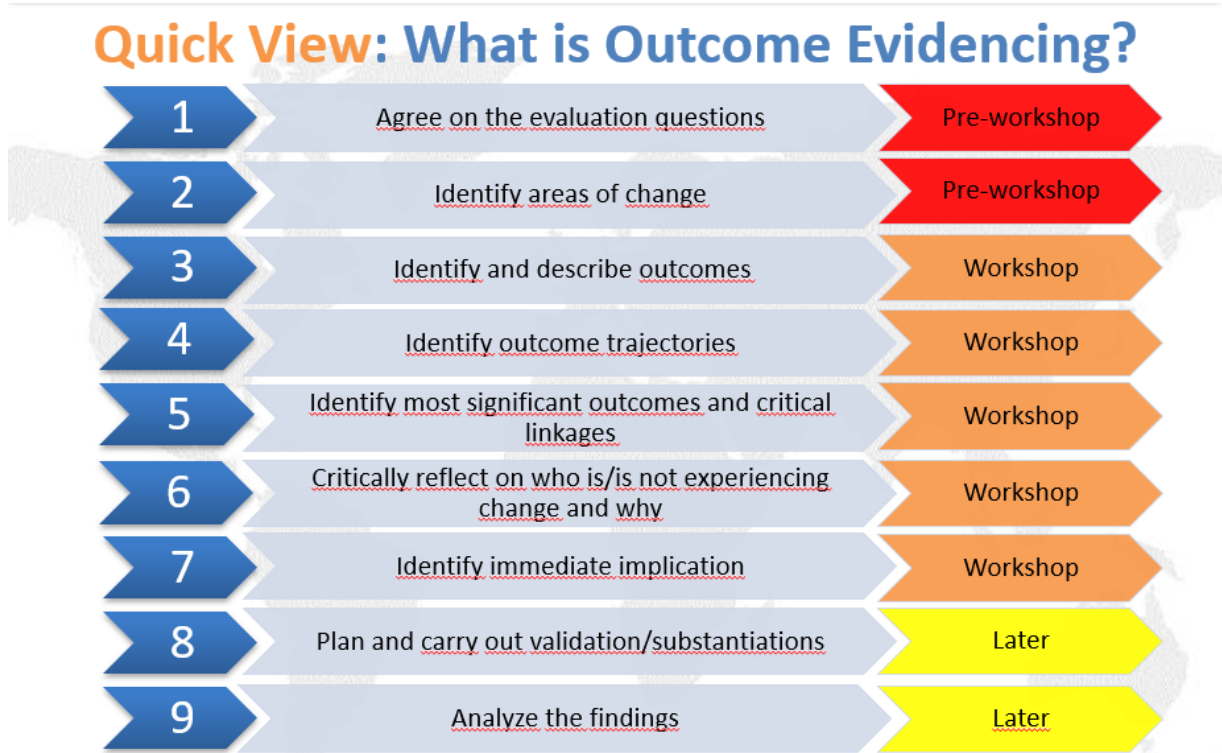
Modified key process steps to applying outcome evidencing based on Paz-Ybarnegaray, R., & Douthwaite, B., 2017 (Source: FAO)

Advantages

- The method delivers a more nuanced evaluation process, with complexity being unpacked by implementers (Change Agents) themselves.
- There is an emphasis on gathering contextualized information because the process is participatory, and context is needed to link outcome trajectories.
- It promotes ownership of the evaluation process and a strong utilization focus by encouraging the Change Agents to direct the conversation. It promotes self-realization, self-assessment, and encourages dialogue and critical thinking across independent Change Agents (i.e. where several NGOs implement distinct components of the project/programme)
- It provides the evaluator a set of "validation points" (evidencing part of the workshop) which can be used to streamline and prioritize field mission or substantiation of findings, making the evaluation cost-effective.

- It encourages the utilization of the evaluation process (not just the report) because it produces evaluative findings immediately (during the workshop), which can be validated, substantiated or translated to a recommendation that can be acted upon by the change agents.

Disadvantages

- It needs preparation. Because the tool includes various PRA tools, the implementation of the outcome evidencing workshop takes longer than most focus group discussions or theory of change workshops. This requires additional cost and time of both the evaluators and Change Agents.
- The outcome evidencing workshop is process intensive. It needs dedicated facilitators for breakout groups and on-the-spot analysis and gathering of evidence of findings.

Suitability

- When you want to increase the utilization of the evaluation.
- When you have limited time/budget for field validation and need prioritization/directed tools to gather evidence.
- When you are dealing with a complex project/programme with a diverse and often independent set of actors (often siloed to implement project/programme components).
- When you have a team of professional and experienced facilitators that are familiar with the subject matter and can handle real time analysis.
- When you want to a tool to collect possible case studies.
- When you can employ snow-ball process of further validation (i.e. add to fieldwork, quantitative, qualitative, triangulation).

Applications in the UN
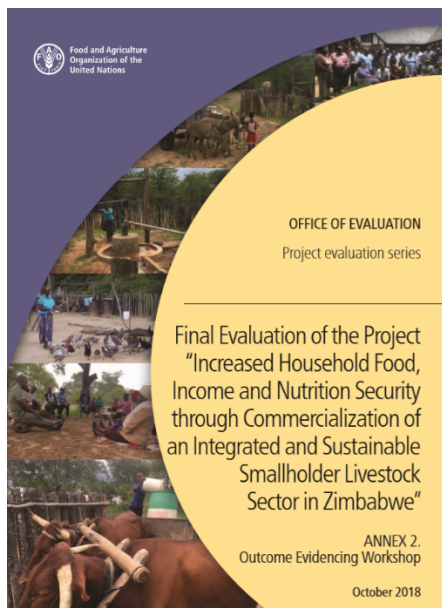
The Evaluation Office of the Food and Agriculture Organization of the United Nations (FAO) applied this method in evaluating a livestock project in Zimbabwe's Matambeland Province.

The project's objective was to encourage smallholder livestock farming as business through improving animal health, market linkages and improving policy for livestock. The interconnectedness of the project's intended results would contribute to improve resilience of communities, in the two dry land districts of Lupane and Nkayi, which is commonly prone to drought and surges of livestock diseases.

Outcome evidencing was used in the initial stage of the evaluation to unpack the complexity of the intervention and its implementation approaches and contextualize the results achieved. A workshop was held with twenty-five change agents, who were representatives from the government and two implementing partner organizations, from national, provincial and district levels. The participants identified six "Areas of Change" that were used to extract 150 results (output and outcome level) stemming from the many projects. Through several participatory analyses, the results were condensed into 78 outcome level results. These outcomes were used to build six multi-causal diagrams. A "critical chain of results" was then selected in each diagram. Each of

the 78 outcomes and linkages in the multi-causal diagram served as testable "evidence". Some of these were tested and substantiated through field visits, focus group discussions and interviews. As a modification of the methodology, a Strength-Weakness-Opportunities-Threats (SWOT) analysis was conducted for each "critical chain of results". This added an additional source of verifiable evidence for the outcomes.

Challenges

- Team members serving as facilitators were new to outcome evidencing. It was mitigated by conducting an online preparatory team meeting to familiarize them with the tool before conducting the outcome evidencing workshop in-country.
- Facilitators should ask participants "where can we see these changes" while s/he is facilitating the analysis. This is an important part of gathering evidence. If possible, assign a separate note-taker.
- Pen and paper were used to conduct the spray-diagram and link the outcome trajectories, which is easy to procure and use during the workshop.

Lessons Learned

Overall, outcome evidencing provided a systematic entry point to the project evaluation and increased the engagement of the stakeholders. The methodology helped aggregate, validate, and construct causal linkages around seemingly dispersed and independent project components, in tandem with other evaluation methods. The outcomes that were defined in the workshop were used for further triangulation and for snowball method.

The multi-causal diagrams and the addition of a SWOT analysis of "critical chain of results" were used to direct some of the field validation exercises through field visits, providing for more efficient use of evaluation resources. Since the project's components were being implemented by "change agents" at various government and implementing partner levels (i.e., national, provincial and districts), the workshop contributed to additional cross sharing and added to the utilization of the evaluation process.

7. Culturally Responsive Evaluation

More than a method, culturally responsive evaluation (CRE) can be considered "a system […] of practical strategies and frameworks that attend to culture during the various stages and phases of an evaluation" (Hopson, 2009, p. 433).

Some scholars and affected populations from the global South make clear that they have not found development evaluation, as they experienced it, to be responsive to the power differentials and political contexts that can sustain an oppressive status quo (Chilisa, 2012). They call for evaluators to be socially responsible and address injustice, oppressive privilege and inequity by employing culturally-responsive methodologies.

A five-branch tree of evaluation approaches
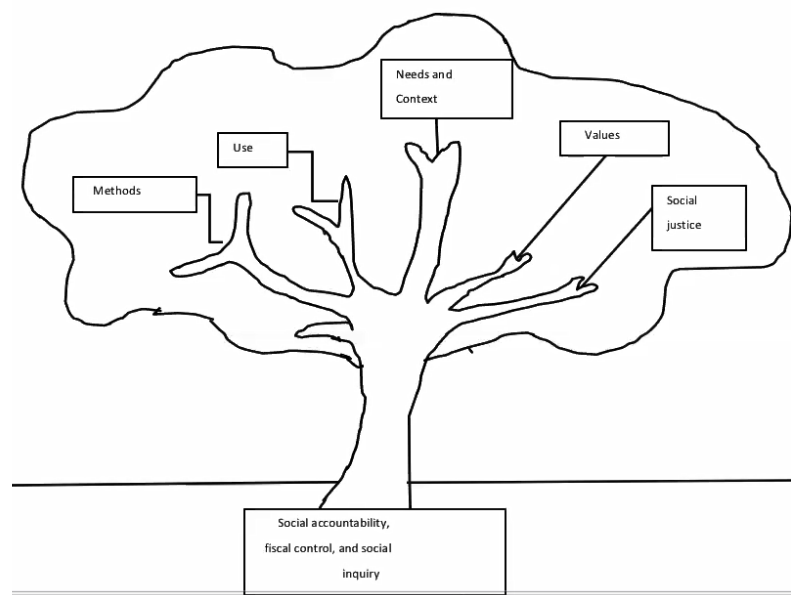
CRE is a theoretical, conceptual, and inherently political position that includes the centrality of culture in the theory and practice of evaluation. CRE recognizes that demographic, socio-political, and contextual dimensions, locations, perspectives, and characteristics of culture matter fundamentally in evaluation. Those who use CRE understand and value "lived experiences" that help to (re)define, (re)interpret, and make sense in everyday life. New explanations and understandings of evaluands and programmes emerge by privileging lived experiences, especially of marginalized communities, populations of colour or indigenous groups (see Hopson, 2009). To understand lived experiences and include them in data triangulation, the following open-ended or "in-depth" interview techniques may be of interest:

- **Storytelling**: The process of collecting stories, whether in groups (e.g. through group interviews or "story circles") or in individual interviews.
- **Relational interviewing** whereby evaluator and interviewee co-create meaning and co-produce narratives in a two-way dialogue rather than a one-way interrogation.
- Related methods are most significant change[11] and outcome harvesting/evidencing (page 23).

---

[11] https://www.betterevaluation.org/en/plan/approach/most_significant_change

There are different ways of recording the information, including standardized questionnaires and open-ended notes. Furthermore, different narrative techniques draw out, code and aggregate fragments of data, including options used in the Most Significant Change approach, such as voting, and the self-signification used in the SenseMaker™ software programme. New technology makes it possible to conduct quantitative analyses based on qualitative, emotional and open-ended narratives.

Overall, analysis focuses on the data's underlying logic (community histories, exploration of metaphors, myths and taboos, etc.) and 'interstices' (including what has not been said) rather than merely the collection or reconstruction of facts.

Advantages

- **Depth and context of data**: In-depth interviews elicit insights into people's thought processes and rationales and help articulate their 'story' in their own words. Stories can be systematically gathered and claims verified from independent sources or options (McClintock, 2004). This provides valuable information and feedback on the community's perceptions about an organisation's efforts. Stories may lead to different observations, emergent themes and patterns from those coming out of partner organisations' progress reports.
- **Focuses on issues such as social (in)justice, race/ethnicity and spirituality** that can be powerful explanatory factors but which tend to be overlooked in more traditional Western-centric evaluation.

Disadvantages

- Integration into more classical overall evaluation designs may be challenging due to a **clash of perspectives**: relational interviewing and the idea of a two-way dialogue from a positivist perspective would be seen to exacerbate the 'interviewer effect' – in which the interviewer's presence and behaviour bias the interviewee's response.
- **Resource-intensive**: In-depth interviews are time-consuming – in terms of identifying subjects, negotiating access, logistics, direct interviewing time, and lengthy transcriptions. A translator may be required to assist in both the interview and transcription.
- Strong **ethical protocols** are needed to avoid unintended transgressions of individual's comfort-zones and levels of privacy.

Suitability

- In the UN context, CRE works best with highly contextual evaluations (country or project evaluations). It is harder to employ with corporate/institutional or global thematic evaluations. It can support **developmental and formative evaluations** in particular, with a strong focus on reflexivity and learning. Storytelling, in particular, can support participatory change processes because it relies on people to make sense of their own experiences and environments. Narrative options can be integrated into on-going organizational processes to aid in programme planning, decision making, and strategic management.
- Conversely, in evaluations that have a strong performance measurement or accountability angle and come from a **positivist perspective** (seeking one "objective truth"), CRE may not be so helpful. It rests on the central premise that knowledge is socially constructed and that multiple truths co-exist.

Application in the UN[12]

UNDP

When assessing UNDP's contribution to Gender Equality and Women's Empowerment (2015), the evaluation teams looked at capturing institutional change through using a blending of "outcome harvesting" and "most significant change'" methods.[13] Thirteen country offices were visited, five of which had been actively involved in UNDP's innovative "Gender Seal" programme, which helps local staff identify connectors between institutional gender mainstreaming policies and activities, and UNDP's technical programme portfolio.

Staff were asked to identify the most significant change in their thematic area over the period 2008-2019. One example was the story of a project to promote the use of more efficient cooking stoves in Bhutan. While the women users of the new stoves were the target population of this project, during implementation it became clear that there were excellent female masons who could be part of the implementation of the programme. The result was that the project trained a cadre of women masons who then found employment in a male-dominated trade. Overall, staff felt that the Gender Seal and other Gender mainstreaming sensitization had helped shape this process of translating gender concerns into technical programme implementation. Focusing on a "significant change" and the outcomes associated with that change, helped locate the connectors between policy and practice.

UNESCO

Moving to an example of culturally – responsive programme implementation, an early example of community involvement was the "Mahaweli Community Radio (MCR)" project that was set up in 1981 in Sri Lanka. It was part of a huge irrigation scheme to resettle approximately one million people from all over the country. The UNESCO-executed project involved mobile teams travelling to villages located in newly-cleared jungle areas. The team spent four days in public areas, including village temples and community centres, to "explore the socio-economic and even psychological realities of village life". They selected an initial theme and returned to home-base to prepare. They then returned to the village three weeks later with mobile consoles and recording equipment. The villagers had, in the meanwhile, organized cultural shows for the recording. There was high participation – "sometimes more than 120 people presented music, theatre and poetry." Following the cultural show, the producers stayed in the village and invited the local people to help shape the final programme which was a mix of the cultural show and interviews. The final show was later broadcast by the national broadcasting corporation.

The community sometimes raised concerns and highlighted issues associated with the delivery of the support needed to start farming in these circumstances. The government was often not happy with this form of public airing of issues. This type of programme, and its push towards public accountability, has had its share of criticism. Some critics felt that it was a mouthpiece for national propaganda and others found fault with the costs and the dependency on donor funding. It may be interesting to revisit this approach for story-telling and community involvement in UN evaluations in this era of mobile phones and social media.[14]

---

[12] With thanks to Chandi Kadirgamar for raising these two interesting examples from her career as a UN evaluator during the discussion segment of the session.

[13] Ricardo Wilson Grau and Heather Britt 'Outcome Harvesting,' Ford Foundation, 2012, and Rick Davies and Jess Dart, "The 'Most Significant Change' Technique", 2005

[14] Source: "Mahaweli Community Radio by MJR David in "A Passion for Radio" Radio Waves and Community www.communica.org/passion

Concluding Remarks

The work of the Methods Group, through this series of useful discussions on approaches, techniques and methods that are relevant to evaluate the work of the UN system at different levels (community, national or global), clearly demonstrates the value added of UNEG as a learning community .

Its work has contributed to professionalization by attempting to formulate and starting to populate a universe of different evaluation designs of relevance to UN evaluators, drawing on good practice in evaluation and social science research more generally. As a result, UNEG has begun to document evidence of the application of these designs across a number of UN work areas including those outlined below.

Evaluations targeted at the community-level

To evaluate the impact/outcomes of the UN's support to community-level beneficiaries, two methods used by UNHCR/UNICEF, UNDP and FAO were discussed:

- RCT and quasi-experimental designs
- Outcome evidencing

Evaluations targeted at the national level

The group also discussed methods that cater for a more holistic evaluation of UN's work. For many agencies, its support consists of a combination of capacity development, policy advice and advocacy, which is often provided at national (or sub-national) levels. Methods used by UNCDF, UNDP/GEF and UNICEF to assess their contribution at the national level or to compare across country-specific contributions were:

- Contribution analysis
- Qualitative comparative analysis

Evaluations targeted at the global level

The UN also plays a major role articulating the efforts of many actors and partners working for a better world for all. In the case of specialized agencies, this role takes the form of multi-stakeholder support to the promotion and implementation of global goals (such as the SDGs) in their areas of expertise (environment, food, agriculture, education, industry, energy, health, gender equality, etc.). Given the complexities faced when assessing contributions of UN agencies at this level, but also the need to take advantage of evidence that already exist, the group discussed two methods proposed by UNFPA and FAO to conduct such evaluations, namely:

- Evaluability assessment and storyline techniques
- Synthesis and meta-analysis

Thinking of the group's contribution in this way can help determine its path going forward, in close cooperation with UNEG colleagues and in direct support of the UNEG strategy.

Building a community of practice of applied evaluation methods within and beyond the UN

Another salient feature of the work conducted by the Working Group in 2019, was the possibility to engage with evaluation experts in different fields and communities of practice. Some highlights included the presentation on impact evaluation by Thomas de Hoop from the American Institute for Research (AIR), Bond's "Choosing appropriate evaluation methods toolkit" by Barbara Befani, and the lively discussion on "Culturally responsive evaluation" led by Professor Bagele Chilisa. The contributions of guest speakers from partner agencies - such as Estelle Raimondo from the World Bank, Mark Engelbert from 3ie and Renata Mirulla from Evaluation Forward – added value to objectives set out in UNEG's strategy: namely, to enhance professionalization and capacity by developing innovative learning services and products for strengthening evaluation competencies and knowledge among UNEG members.

2020-2021 Objectives

The Group aims to continue with the peer-to-peer learning carried out to date. Given that there are many methods of potential utility that have not been discussed yet, there is a need to establish some criteria for determining future priorities and/or expanding into new areas of work.

The Group could contribute to broader discussions regarding the development or updating of methodological guidance for UN evaluations at the thematic, policy, strategy, and programme/project level.

By sharing experiences from UN agencies, external partners and evaluation communities of practice, the group at its core will continue to help evaluation practitioners navigate the universe of methods in a useful and structured manner – in order to tailor them to their needs.

Guest Speakers short bios

Barbara Befani: Dr. Barbara Befani is affiliated with the University of Surrey and the University of East Anglia but working mainly as an independent researcher / consultant. Her interests include 1) evaluation quality; 2) methodological appropriateness and comparative advantages and weaknesses of different evaluation methods; 3) causal inference frameworks for impact evaluation; and 4) specific hybrid, quali-quanti methodologies, like Qualitative Comparative Analysis (QCA) and Process Tracing (in particular its Bayesian formalisation, which she is extending to all forms of Theory-Based Evaluation). She's a former Secretary General of the European Evaluation Society.

Alexandra Chambel: Alexandra Chambel is Evaluation Adviser and lead evaluator of corporate evaluations at the Evaluation Office of UNFPA. She has over 25 years of experience in development of which 18 years working internationally (European Commission/EuropeAid, Evaluation Unit; EU Delegation Guatemala; UNDP IEO and UNFPA EO), holding advisory and management positions spanning from policy advise to evaluation. She counts with 20 years of **professional** practice on evaluation across the humanitarian-development nexus. Alexandra co-convenes the UNEG interest group on Joint Evaluations and actively contributes to the Decentralized Evaluation Interest group.

Bagele Chilisa: Professor Bagele Chilisa is based at the University of Botswana where she teaches Evaluation Theory and Practice; Feminist **Research** Methodologies; Policy Planning and Evaluation; and Quantitative and Qualitative Research Methods. She is a member of the UNDP Evaluation Advisory Panel and an instructor at the International Program for Development Evaluation Training (IPDET) among numerous other credentials.

Mark Engelbert: Dr. Mark Engelbert is an evaluation specialist at the International Initiative for Impact Evaluation (3ie). Mark manages 3ie-funded research and conducts evidence synthesis for 3ie's Agricultural Innovations and Immunization evidence programmes. He also manages 3ie's Development Evidence Portal, which is a largest-of-**its**-kind online database of rigorous impact evaluations and systematic reviews in international development.

Christine Fu: Dr. Christine Fu is Senior Evaluation Officer at UNHCR's Evaluation Service. She oversees the design and management of independent corporate evaluations ranging in focus from organizational effectiveness, **international** protection, public health, livelihoods and economic inclusion, cash-based interventions, workforce learning and development and private sector engagement. Prior to joining UNHCR, Christine has worked at USAID, INGOs, Red Cross and grassroots civil society organizations.

Andrew Fyfe: Andrew Fyfe is Head of Evaluation at the UN Capital Development Fund where he is responsible for designing and managing evaluations of UNCDF's work to unlock public and private finance for the poor in the Least Developed Countries. He is also an active member of the UN Evaluation Group, co-convening the Working Group on evaluation methods and contributing to UNEG normative work on peer reviews of small evaluation functions. He was elected as a Vice – Chair of UNEG in 2020.

Harvey Garcia: Harvey Garcia is an Evaluation Specialist in the Independent Evaluation Office of UNDP. He conducts independent country program and thematic evaluations. Before joining IEO, Harvey worked as an Evaluation Officer in FAO's Office of Evaluation in Rome. Prior to this he was the Head of the M&E Unit in FAO's Haiyan response in the Philippines. Harvey has more than 14 years

of experience in research and evaluation and has a degree in wildlife biology and a master's in natural resource management. He is the co-Chair of the UNEG Partnership Group.

Jeneen R. Garcia: Jeneen R. Garcia is an Evaluation Officer at the Global Environment Facility Independent Evaluation Office (GEF IEO). Since joining the GEF IEO, she has dedicated herself to testing innovative methodologies that bridge qualitative and quantitative approaches. She has co-developed conceptual frameworks for assessing impact in complex adaptive systems, applied to evaluations on various themes such as international waters, biodiversity, partnerships, environmental and socioeconomic synergies and trade-offs, and scaling-up of environmental outcomes. Jeneen is currently Program Co-Chair of the American Evaluation Association's Systems in Evaluation Topical Interest Group (SETIG)

Arwa Khalid: Arwa Khalid is an Evaluation Officer at the FAO's Office of Evaluation (OED). She oversees the design and management of independent corporate thematic, country programme and project evaluations. She also leads the Gender, Human Rights and Disability-working group in OED. She has co-developed OED guidelines for the assessment of gender mainstreaming in evaluations. Prior to joining FAO, Arwa worked at GIZ (when it was GTZ), CARE-Australia and the private sector.

Thomas de Hoop: Dr. Thomas de Hoop serves as a principal economist for AIR in Washington D.C. He has 13 years of experience designing, implementing, and leading mixed-methods impact evaluations and systematic reviews with a focus on South Asia, and sub-Saharan Africa. Currently, he oversees a research and evaluation portfolio with a focus on the impact, scalability, and cost-effectiveness of women's collectives and innovations in education.

Estelle Raimondo: Dr. Estelle Raimondo is an evaluation methods expert at the World Bank's Independent Evaluation Group, where she leads evaluations and advises and trains teams on a wide range of evaluation methodologies. She is also a faculty member of the International Program for Development Evaluation Training (IPDET) and adjunct faculty at the George Washington University. Her research has been published in several internationally peer-reviewed journals. She is also a Board member of the European Evaluation Society.

Carlos Tarazona: Carlos Tarazona is a development specialist with over twenty years of experience in the design, management and evaluation of agricultural and rural development programs. He was co-chair of the UN Evaluation Group strategy task force (in charge of developing the UNEG Strategy 2020-24), and is currently a Senior Evaluation Officer in FAO and co-coordinator of the UNEG Working Group on Methods.

Tina Tordjman-Nebe: Dr. Tina Tordjman-Nebe is Senior Evaluation Specialist at UNDP's Independent Evaluation Office (IEO) in New York, which she joined in July 2020 from UNICEF's Evaluation Office. She conducts independent country programme and corporate evaluations, and coordinates the IEO's "Reflections" series, offering lessons based on rapid evidence assessment of past UNDP evaluations. In 2019/20, she co-convened the UNEG working group on ethics and the UNEG Working Group on Methods.

Pietro Tornese: Pietro Tornese is Evaluation Analyst at UNCDF in New York. Prior to joining UNCDF, Pietro was UN/DESA Fellow in Havana. He has also worked for Ernst & Young, Open Evidence, and the Italian Agency for Development Cooperation in Latin America and the Caribbean. As an independent consultant, Pietro has co-authored a number of evaluations and policy studies for the European Commission. He holds a Master of Economics from Bocconi University.

Henri van den Idsert: Henri van den Idsert is a Senior Evaluation Officer at UNHCR who oversees a broad range of evaluations under his portfolio, including country strategy evaluations, livelihoods, inter-agency humanitarian evaluations, and emergency preparedness and response. Henri has over twelve years of experience in monitoring and evaluation, his experience extends across the humanitarian-development nexus, having worked for UNDP in South Sudan, the Dutch Ministry of Foreign Affairs in the Hague, NGOs in the humanitarian and sexual reproductive health and rights sector, as well as public policy and development consultancies in Somalia and Kenya.

References

Evaluability assessments and storyline approaches

1. BetterEvaluation.org (accessed 29 April, 2020): https://www.betterevaluation.org/en/themes/evaluability_assessment
2. Davies R (2013). Planning Evaluability Assessments: A Synthesis of the Literature with Recommendations. DFID Working Paper 40. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/248656/wp40-planning-eval-assessments.pdf
3. OECD-DAC (2010). Glossary of key terms in evaluation and results based management. Paris: OECD-DAC. Available at: http://www.oecd.org/development/peer-reviews/2754804.pdf
4. Joint Evaluation of the UNFPA-UNICEF Joint Programme on the Abandonment of Female Genital Mutilation: Accelerating Change Phase I and II (2008–2017)
5. Thematic Evaluation of UNFPA support to the prevention, response to and elimination of gender-based violence and harmful practices (2012-2017)
6. Evaluation of FAO's Work on Gender: http://www.fao.org/3/ca3755en/ca3755en.pdf

Synthesis and meta-analysis

7. Synthesize data from one or more evaluations, Better Evaluation brief https://www.betterevaluation.org/sites/default/files/Synthesize%20-%20Compact.pdf
8. FAO evaluation reports and management responses: www.fao.org/evaluation
9. IFAD Annual Report on Results and Impact of IFAD operations (ARRI): https://www.ifad.org/it/web/ioe/arri
10. 3ie's evidence maps: https://www.3ieimpact.org/evidence-hub/evidence-gap-maps
11. 3ie systematic review repository: https://www.3ieimpact.org/evidence-hub/systematic-review-repository
12. Campbell Collaboration Library: https://onlinelibrary.wiley.com/journal/18911803
13. Synthesis of WFP's country portfolio evaluations in Africa (2016-2018) https://www.wfp.org/publications/synthesis-wfps-country-portfolio-evaluations-africa-2016-2018
14. Synthesis of Four Evaluations of the Impact of WFP Programmes on Nutrition in Humanitarian Contexts in the Sahel (2018) https://www.wfp.org/publications/four-evaluations-impact-wfp-programmes-nutrition-humanitarian-contexts-sahel-synthesis
15. Meta-syntheses of UNODC's evaluation reports: https://www.unodc.org/unodc/en/evaluation/meta-analyses.html

Contribution analysis

16. Befani, B. & O'Donnell, M. (2016) Choosing Appropriate Evaluation Methods: A Tool for Assessment and Selection. Bond.
17. Befani, B., & Mayne, J. (2014). Process Tracing and Contribution Analysis: A Combined Approach to Generative
18. Causal Inference for Impact Evaluation. (B. Befani, C. Barnett, & E. Stern, Eds.) IDS Bulletin, 45(6), 17-36.

19. Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. The Canadian Journal of Program Evaluation, 16 (1), 1-24.
20. Mayne, J. (2011) Contribution Analysis: Addressing Cause and Effect in Evaluating the Complex, K. Forss, M. Marra and R. Schwartz (Eds.), Transaction Publishers; Piscataway, New Jersey.
21. MM4P evaluation
22. Evaluation at UNCDF
23. Market systems and the poor – DCED
24. Measuring Results & the DCED Standard
25. CGAP Guidance – Measuring Market Development

Qualitative Comparative Analysis

26. Befani, B., 2016. Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA)
27. https://www.betterevaluation.org/en/evaluation-options/qualitative_comparative_analysis
28. GEF IEO (Global Environment Facility Independent Evaluation Office). (2016). Impact Evaluation of GEF Support to Protected Areas and Protected Area Systems, Evaluation Report No. 104, Washington, DC: GEF IEO. Accessed May 2020: https://www.gefieo.org/evaluations/impact-evaluation-gef-support-protected-areas-and-protected-area-systems-pas-2016
29. UNICEF Evaluation Office (2018). Strengthening child protection systems: Evaluation of UNICEF strategies and programme performance. UNICEF, New York. https://www.unicef.org/evaldatabase/index_103557.html

RCT and quasi-experimental designs

30. http://hea.globalinnovationexchange.org/programs/humanitarian-education-accelerator-hea
31. http://hea.globalinnovationexchange.org/blogs
32. https://inee.org/system/files/resources/Policy-brief-Rigorous%20Research%20in%20Humanitarian%20Contexts-Final.pdf
33. https://blogs.worldbank.org/impactevaluations
34. https://www.betterevaluation.org/en
35. https://www.unicef-irc.org/publications/752-randomized-controlled-trials-rcts-methodological-briefs-impact-evaluation-no-7.html
36. https://www.unicef-irc.org/publications/753-quasi-experimental-design-and-methods-methodological-briefs-impact-evaluation-no.html
37. White H. (2011), *An introduction to the use of randomized control trials to evaluate development intervention* International Initiative for Impact Evaluation 3ie, New Delhi, Global Development Network. Retrieved from http://www.3ieimpact.org/media/filer/2012/05/07/Working_Paper_9.pdf
38. Bulpitt C.J. (1996) *The Advantages and Disadvantages of Randomised Controlled Trials. In: Randomised Controlled Clinical Trials*. Springer, Boston, MA

Methods to support evaluation design

39. BetterEvaluation.org (accessed 29 April, 2020): https://www.betterevaluation.org/en/themes/evaluability_assessment

40. Davies R (2013). Planning Evaluability Assessments: A Synthesis of the Literature with Recommendations. DFID Working Paper 40. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/248656/wp40-planning-eval-assessments.pdf

41. Critical Moments Reflection Methodology. A Reflective Method and How to Use it with Groups of People. Available at: http://web.mit.edu/cil/web_scripts/www/Critical%20Moments%20Methodology%20Brief%20CoLab.pdf

42. Mayne, John. "Useful theory of change models." Canadian Journal of Program Evaluation (30.2), 2015.

43. Mayne, John. "The COM-B Theory of Change Model". Working Paper, 2018.

Outcome evidencing

44. FAO. (2018). Final Evaluation of the Project "Increased Household Food, Income and Nutrition Security through Commercialization of an Integrated and Sustainable Smallholder Livestock Sector in Zimbabwe" FAO Office of evaluation, Project evaluation series http://www.fao.org/evaluation/evaluation-digest/evaluations-detail/en/c/1171794/

45. Paz-Ybarnegaray, R., & Douthwaite, B. (2017). Outcome evidencing: a method for enabling and evaluating program intervention in complex systems. American Journal of Evaluation, 38(2), 275-293.

Culturally responsive evaluation

46. https://www.betterevaluation.org/en/evaluation-options/stories

47. https://www.betterevaluation.org/en/evaluation-options/interviews

48. Chilisa, Bagele. Indigenous Research Methodologies. United States: SAGE Publications, 2019.

49. Fujii, Lee Ann. Interviewing in Social Science Research: A Relational Approach. Taylor & Francis, 2017.

50. Hopson, Rodney K. "Reclaiming knowledge at the margins - Culturally responsive evaluation in the current evaluation moment". In K. Ryan and B. Cousins (Eds.), International Handbook of Educational Evaluation. Thousand Oaks, CA: Sage, 2009

51. McClintock, Charles. Using Narrative Methods to Link Program Evaluation and Organization Development. The Evaluation Exchange, Vol. 9, No. 4. 2004: 14-15.