Scoping Report
Drafted by Heather Britt
October 2009

Introduction

This report provides information on current evaluation methodologies and methods in gender training. The report was commissioned by the World Health Organization's (WHO) department of Gender, Women and Health (GWH). A monitoring and evaluation framework (M&E) is required for core gender mainstreaming capacity building activities within WHO[1], predominantly around two training tools (Gender Mainstreaming for Managers: A Practical Approach; and The WHO e-learning series on Gender and Health: Awareness, Analysis, Action). The purpose of the report is to inform decision making and design for the proposed M&E framework.

Specifically, the report aims to:

- Review and analyse different approaches for their implications for the GWH capacity building M&E activities.
- Summarize similarities and differences between each approach, highlighting areas of interest for GWH activities.

The report is modest in scope, and bounded closely by its purpose. The GWH focal person and the consultant worked together to identify likely sources of relevant information and to gather reports and comments from informants. A limited reading list on evaluating gender training beyond participant satisfaction was compiled through snowball sampling. Published work was collected through a search of the internet and various search engines through Reference Manager 10 [2]. Unpublished work, comments and suggestions were sought through a contact list of those working in gender and health, or from other UN agencies (e.g., gender units in other International Organizations). Specific attention was given to contacting those responsible for gender training for their insight and experiences with evaluating gender training. This search resulted in a very limited list of resources, and the list was expanded to include a few items related to M&E of capacity building and learning programs in general. The consultant reviewed the materials in order to draft this report.

Gender training aims to raise awareness, increase knowledge, build skills and change behaviours of participants. Gender training may also aim to empower facilitators and participants through the methods and process used to implement the training. Monitoring and evaluation activities, then, may need to capture both the intended outcomes of gender training as well as the process in which it is delivered.

[1] WHO's Gender Strategy (WHA 60.25) includes building capacity for gender analysis and planning as one of its four strategic directions. It is expected that this evaluation framework will contribute to evaluation of the capacity building strategy of the Gender Strategy.

[2] Parameters of the search were as follows: "evaluation + gender training," "impact + gender skills development," "monitoring and evaluation + gender training/capacity development."

Findings on Gender Training Evaluation

The review found:

1. Limited attention has been paid to monitoring and evaluation of gender training;
2. The bulk of M&E for gender training covers trainee satisfaction, and does not address outcomes such as increased knowledge and awareness, improved skills, changed behaviors or higher level organizational changes, and
3. No evidence of M&E systems which provide for the regular, on-going collection, analysis and use of M&E data to inform the design and implementation of gender training programs.

Selected quotes on M&E of gender training

Few training centers have conducted follow-up to determine how participants have used the training in their work.

The training…. lacked specific indicators that could be periodically measured to provide empirical evidence of progress.

Few training centers had conducted any type of assessment of pre- and post-learning (level 2) or longer term follow up to document level 3 and level four changes in performance or impact related to the training.

Because post-course follow up was almost universally weak, little written documentation exists to substantiate that the training had an impact on performance, programs, or policies.

In general, gender training may be characterized by an absence of M&E or routine tracking related to outcomes. A virtual discussion on gender training for security sector personnel concluded that "most gender training courses do not include an evaluation component, and there were very few examples of long-term impact evaluation of gender training."[3] Participants in the discussion represented many organizations from around the globe. Comments by informants of this review, as well as the remainder of readings, seem to indicate that this is a fair statement for gender training beyond the security sector. Examples of evaluations of gender programming were found, but very few examples of evaluation of gender training.

Findings Related to General Capacity Building & Learning Program Evaluation

Because of the modest scope of this review, it relied heavily on a recent study commissioned by the World Bank with similar objectives and the resources to conduct a thorough literature search and extensive interviews with training executives and evaluators.[4] The scope of the Evans review was considerably broader than capacity building for a single content area, such as gender, and included evaluation of learning programs related to professional development of any kind.

[3] Good and Bad Practices in Gender Training for Security Sector Personnel: Summary of a Virtual Discussion (June 2007) UN-INSTRAW

4 Are There New Approaches to the Evaluation of Learning Programs? David Evans. Evans' interview respondents were drawn from four organizational types: (a) international organizations with a similar mission to the World Bank, including the U.S. Agency for International Development (USAID) and the Inter-American Development Bank(IDB); (b) organizations devoted to training and research on training methods; (c) universities offering professional development programs or conducting research in evaluating training programs; and (d) private sector firms conducting staff training and development programs.

The limited use of M&E for gender training was found to be true of capacity development programming in general regardless of training content or purpose. "Most organizations rely on simple evaluations of participants' perception of learning." (Evans p. vi)

Not only is the implementation of M&E limited, but so too is the number of approaches used. Evans reports that the dominant framework used to evaluate staff development programs is that put forward by Donald Kirkpatrick 40 years ago[5]. Although Kirkpatrick's model has limitations, Evan's review found few new, cutting-edge models, frameworks, methodologies, or approaches in use. The popularity of the Kirkpatrick model is likely due to its simplicity and because it delineates five common sense areas that most evaluators would want to investigate.

Why is an Evaluation Framework Important?

Before reviewing the Kirkpatrick model and its alternatives, a discussion of the role of evaluation frameworks is useful. An evaluation framework influences the types of questions that are asked, the types of methods used, and therefore the information that is produced by evaluation exercises. Evaluation frameworks even influence how data is interpreted. How does it do that?

An evaluation framework is a conceptual model of the area under investigation. Where possible, a model should hold together coherently by representing the current knowledge of the field gained through research, theory and practice. In other words, it should represent accurately what is known about the area under investigation. The purpose of program evaluation is generally to provide answers to practical questions, not test or validate theory. However, a well-tested model can help formulate questions thereby shaping evaluation design to reliably provide answers to real world problems.

Where little is known, in an area where research and theory cannot shed light, an evaluation framework should lead evaluators to uncover the dynamics of change at work in their specific evaluation context.

In summary, the usefulness of an evaluation framework is its explanatory power. It should be able to guide an evaluation that can describe both what is happening, and *why* it is happening.

Four Kinds of Frameworks

How to assess the various available frameworks for the purpose at hand? The available models and approaches to evaluation of capacity building can be categorized according to their ability to shape evaluations that can answer both the "what" and the "why" satisfactorily:

1. Taxonomies
2. Logic models
3. Open learning approaches
4. Cynefin Framework

A **taxonomy** guides an evaluation by telling an evaluator where to look. It does that by defining a number of conceptual areas for investigation, areas where change might be expect to occur. However, it does not specify the relationships between these conceptual areas, or between them and the context

[5] Donald Kirkpatrick first proposed his framework in 1959 and 1960 in a series of articles in the Journal of the American Society for Training and Development.

in which they operate. A good taxonomy identifies useful areas to measure to find results, but even the best taxonomy cannot explain why the results are occurring.

In order to uncover the “why,” the evaluation must be guided by an explanation -- a **logic model**. A logic model describes the theory about how change occurs. It explains why a particular intervention can be expected to bring about or contribute to certain results. A good logic model must also include a description of the assumptions that support it; that is, what kind of context it can operate effectively in, and what factors, in addition to the intervention itself, must be operating to see the desired outcomes.

Frameworks based on a linear logic model outline pre-determined objectives for the capacity building initiative for both learners and the organization. These frameworks include the influence of factors other than the capacity building intervention in an attempt to the meet the attribution challenges for higher level results. A variety of M&E methods from the applied social sciences may be used for measurement.

Critics of logic models claim that most of the contexts in which we live and work are much too complex to be accurately described in a linear model. They state that in any context, a number of factors interact dynamically with a change intervention such as a training program. Proponents of this kind of thinking describe change in context as a complex, adaptive system and propose **open learning methods** as a means of discovering both what is happening and why it is happening. In contrast to logic models, open learning methods do not attempt to define in very specific ways the outcomes that are expected from an intervention, but instead set out to discover what is happening without preconceptions. These methods are well-suited for capturing unintended consequences and therefore, they are an improvement on both taxonomies and logic models in terms of describing results. Because of the complexity of causality, some applications of open learning methods may not always give as much attention to explaining why certain results are found. In that way, open learning frameworks may be difficult to distinguish from taxonomies. Evaluation methods included in this category include Most Significant Change (MSC) and various theories of change methods.

Following in the footsteps of the early systems thinkers and open learning advocates, Cynthia Kurtz and David Snowden[6] began examining contexts more closely, as well as the factors at play and the interactions between them. Their analysis refined the thinking about systems by pointing out that not all contexts are characterized by the same level of complexity. The **Cynefin Framework** allows for simple, complicated, complex and chaotic aspects of change, and suggests evaluation methods appropriate to the level of complexity of each aspect.

To summarize, evaluation frameworks for capacity building results can be categorized according to their approach to measuring or discovering results (the “what”) as well as their ability to address causality of varying degrees of complexity (the “why”). The sections below will examine how well each of the four kinds of frameworks -- taxonomies, logic models, open learning approaches and the Cynefin model – can provide guidance on design of an M&E system to provide information on the results of gender training.

[6] Kurtz, C. F. and D. J. Snowden. 2003. “The new dynamics of strategy: Sense-making in a complex and complicated world.” IBM Systems Journal, vol 42, number 3, page 462.

Kirkpatrick: The Most Common Evaluation Framework for Capacity Building and Learning[7]

Kirkpatrick is by far the most commonly used framework for evaluation of capacity building and learning programs. This framework, first put forward in 1959, is best characterized as a taxonomy. It outlines four levels for evaluation of training and capacity building:

Level 1	**Reaction**	**how learners perceive instruction or training**
Level 2	**Learning**	the extent to which learners change attitudes, gain knowledge, or increase skills as a result of training
Level 3	**Behavior**	how learners change their behavior as a result of training
Level 4	**Results**	the impact that has occurred at the organizational level as a result of training

Kirkpatrick's framework implies that each level builds on the previous one, but it does not define cause and effect relationships between them, and does not put forward a theory of change supported by research in the field. In fact, the implied linear causal relations among its levels do not stand up to logic, are not grounded in theory, and have not been demonstrated by research. The lack of research to further develop the framework represents a major shortcoming in the field of training evaluation which has been noted by various scholars.

Kirkpatrick Modifications

Over the years, a number of modifications to the original Kirkpatrick model have been introduced. Although the Kirkpatrick model has been criticized, is it possible that its various modifications have addressed these limitations satisfactorily? Evans describes several modifications that have been made to the Kirkpatrick model.

> **Level 1**: Alliger amended Level 1 to distinguish between participants' affective reactions and their judgments about utility; that is, the difference between whether participants like the training, and whether they consider it useful. Research has shown that participants' affective reactions are positively correlated with perceptions regarding utility. If trainees like the training, they are also likely to consider it relevant. More importantly, Alliger's research found positive correlations between utility reactions and immediate learning gains (Level 2), but almost no correlation between affective reactions and such learning gains. (Evans refers to Alliger et al's 1997 publication, see references).
>
> Also, Level 1 questionnaires may be designed to measure how participants plan to implement skills acquired in training or other planned actions resulting from the training. Strictly speaking, this second modification may be closer to a pre-test for Level 2, than a Level 1 revision.
>
> **Level 2**: Alliger's expanded version of Level 2 distinguishes between 3 types of learning:
>
> - Knowledge assessed at end of training
> - Knowledge assessed several months after training
> - Demonstration of new skills or behaviors immediately after the training
>
> **Level 5**: Return on Investment (ROI). This is an addition to Kirkpatrick's original four-tiered model that is intended to allow for calculating the value of training to an organization by quantifying the costs and benefits.

[7] This section draws heavily on the David Evans article (p. 9)

None of these modifications addresses the major limitations of the model, namely its lack of theoretically sound and research-verified causal links between the levels. Evans' review found that only Level 2 appears to have some validity. As mentioned, studies have found a relationship between the perceptions of relevance and utility, on the one hand, with learning gains, on the other hand.

Taxonomies Summary

Pros: The Kirkpatrick model (and its modifications) provides a common sense taxonomy which outlines a number of topics which are likely to be of interest. There are a large number of resources available to facilitate its implementation. Because it is so widespread, benchmarking may be possible.

Cons: Weak logic linking the various levels which are not addressed by the model's various modifications. The higher levels (3, 4 and 5) face measurement and attribution challenges; in other words, findings related to gender (learner attitudes, knowledge and behavior, as well as organizational findings) cannot be convincingly attributed to the training.

Frameworks based on Logic Models

Evaluation approaches based on logic models outline pre-determined objectives for the capacity building initiative for both learners and the organization. More sophisticated models include a thorough mapping of the additional factors influencing change at the various levels of the model and attempt to address challenges in attribution for higher level results. Attempts to approach capacity building evaluation through logic modeling may result in a model consistent with the levels outlined by Kirkpatrick, but with cause and effect spelt out. It is also possible that the exercise to determine objectives may identify areas outside the Kirkpatrick model. A variety of M&E methods may be used for measurement. Evaluation designs based on these models may capture unintended consequences, but the models themselves are constructed around pre-determined objectives.

In this review, Bresin may be considered a rudimentary version of an approach based on logic models. He encourages the very specific definition of higher level results (business outcomes) and the alignment of training to those results in order to both achieve and measure results. Evaluation frameworks based on logic models may be considered an improvement on the Kirkpatrick model, but some limitations persist. One of the most important limitations is the tendency to make exaggerated claims regarding outcomes and impacts of training. This is because logic models, like taxonomies, do not shed light on all the other factors that contribute to overall trainee and organization-wide effects. Furthermore, logic models may foster a blind spot for unintended consequences.

Logic Model Approaches Summary

Pros: In certain cases, the limitations to higher level measurement (Kirkpatrick levels 3, 4 & 5) may be addressed by improved logic modeling which accounts well for assumptions underlying predicted changes.

Cons: Exaggerated claims for capacity building outcomes and impact due to overemphasizing the role of training and omitting other contributing factors; overlooks unintended consequences.

Within this review, the proponents for open learning and theory of change approaches are represented by Ortiz and Taylor. The authors address the problems in attribution faced by logic models with higher level approaches from another perspective. They state that organizations are complex adaptive systems and linear models are insufficient to describe change. The link between strengthening technical capacities and achieving organizational level results is difficult to determine: "It is very difficult to find a direct correlation between the use of the new processes and higher level impacts." (Ortiz & Taylor, p 7)

"Whereas pre-programmed performance might be a good indicator of capacity development in some cases, *outputs and outcomes that are the result of emergent adaptive management/agile responses to complex environments are even more important proxies for capacity development."* (Ortiz & Taylor, p 10; emphasis theirs)

What are the consequences for capacity building M&E of focusing on the immediate, pre-determined results of specific capacities? Failure! "Most capacities have only weak links with immediate performance (which isn't inherently a good or bad thing), and forcing these linkages in M&E plans only sets us up for failure when it comes to the time for measurement." (Ortiz & Taylor, p 10) Ortiz and Taylor recommend using organizational learning and Theory of Change approaches for designing M&E systems for capacity building interventions. They mention specifically the Most Significant Change technique and an experimental TOC approach used by Keystone, but few details are provided.

Open systems approaches address a very real need in the evaluation field – to account for the conditions of complex change. Nevertheless, these approaches have several limitations. First among them is the overemphasis on complexity. The assertion that all elements of development programming and capacity building are emergent seems an exaggeration, and downplays the contributions of research on learning and education. Certain relationships within learning outcomes systems have been proven by research.

Another limitation of open systems methods is their tendency to relinquish addressing causality and attribution. Because of the complexity of causality, some applications of open learning methods do not give as much attention to explaining why certain results are found. In this way, they lose explanatory power and can only be distinguished from taxonomies by their improved capacity to capture unintended consequences[8].

<u>Open Learning Approaches Summary</u>

Pros: Addresses the conditions of complex change and captures unintended results.

Cons: Tends to overemphasize complex aspects of contexts and neglect research and theory in the field of learning and education; downplays the need to explain casual relationships. This approach and their associated methods are less well-known, and therefore, may be a harder "sell" in organizations which place a high premium on standard research methods.

[8] In this review the ProLEAD evaluation of a gender training program is such an example. This evaluation model outlines three domains of change (individual; networks & partners; public policy and health systems) and looks for changes over time (immediate, mid-term and long-term). The model does not specify the cause and effect relationships either across time or between the domains. In fact, the EvaluLEAD framework on which it is based specifically states that Attributing and documenting causal relationships between the program activities and the results by domains is not the aim (p 6).

The Cynefin Framework

Williams, Gujit and Rogers, in their recent presentation at the 3rd Impact Evaluation Conference[9], suggest that Snowden's Cynefin Framework has implications for guiding evaluators. Rather than characterizing change as either simple (appropriate for linear logic models) or complex (captured only by open learning approaches), they suggest that situations contain aspects that can be simple, complicated or complex, as suggested by Snowden's Cynefin Framework.

4 Categories of the Cynefin Framework[10]

Simple	Weak connections *between* elements, but a strong link to a central control element.
Complicated	Strong connections *between* elements, but each element is still has strong links to a central controlling element.
Complex	Strong connections *between* elements, but no central controlling element.
Chaotic	Weak connections *between* elements, and no central organizing core.

The framework is frequently represented by the following diagram[11].

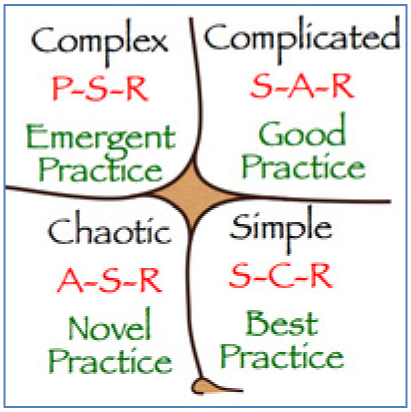

Figure 1: Cynefin Framework

The framework is not intended to label a situation or an entire system as a single category. Instead, elements of a situation demonstrate the behaviors and structures associated with one of the four categories. And over time, the nature of these elements may shift from one category of the framework to another.

The implication for evaluators is that different aspects of change require different evaluation approaches to discover and measure what is happening and why it is happening. The simple components of a change process are well-suited to linear logic modeling tools and M&E approaches. There may be some cases in which training may be directly related to specific pre-determined outcomes. For example, research has proven a connection between learner-perceived relevance and changed

[9] "Thinking Systematically about Impact Evaluation of Programs and Policies with Simple, Complicated and Complex Aspects," presentation by Bob Williams, Irene Gujit, and Patricia Rogers (1 April 2009), Cairo, Egypt.
[10] This version from Bob Williams manuscript.
[11] http://en.wikipedia.org/wiki/Cynefin

behaviors. Standard social research methods are suitable for complicated aspects of situations, while complex and chaotic components require M&E approaches that allow for learning in emergent situations, such as the approaches suggested Ortiz and Taylor.

Evaluators working with the Cynefin framework can identify the various aspects of the situation and select M&E approaches best suited to capture the various types of change. In fact, the framework identifies ways of knowing, or sense-making, that in unique combinations provide the most appropriate way to manage each aspect of a situation.

Cynefin Types of Sense-Making and Management

Sense	Collect sufficient data to identify the characteristics of this aspect of a situation
Categorize	Identify where these characteristics fit within known world
Analyze	Get the networks to find out the information and use expertise to choose the most appropriate means of response
Respond	Pick the proven appropriate response to that category
Probe	An experiment that makes patterns more visible and knowable by sensing.
Act	A strong intervention designed to shock a chaotic aspect of the situation back into some form of order

By assessing evaluation methods according to these approaches to knowing and managing it is possible to identify evaluation approaches and methods appropriate for each category of complexity. Snowden and those working with the Cynefin framework claim that most previous means of research and inquiry are suited for simple and complicated situations, but poorly adapted for the complex or chaotic. Cognitive Edge, a consulting company founded by Snowden, has developed a suite of methods to address this gap.

Evaluation Methods Appropriate for Levels of Complexity

Category	Sense-Making & Management	Possible Evaluation Approaches & Methods
Simple	Sense, Categorize, Respond	Logic models & standard social science methods
Complicated	Sense, Analyse, Respond	Social science research using experimental designs.
Complex	Probe, Sense, Respond	Facilitative & exploratory approaches means to collect data on results, factors & relationships. (MSC, SCM, CE narrative methods)
Chaotic	Act, Sense, Respond	CE narrative methods?

Cynefin Framework Summary

Pros: Addresses both intended and unintended consequences well. Addresses a various levels complexity related to causality. Easier to sell as a mixed-methods approach, by including well-recognized approaches and methods together with those better suited to complex adaptive systems. Able to address areas of interest comparable to those identified by Kirkpatrick, but with more explanatory power.

Cons: Less emphasis on intended results, and simple causality; not well-known.

Recommendations for Selecting an Evaluation Framework

Despite Kirkpatrick's dominance of the training measurement field, a number of alternative evaluation approaches to capacity building do exist[12]. This review has outlined four types of evaluation frameworks based on their explanatory power – taxonomies, logic models, open learning approaches, and the Cynefin framework. The table below summarizes their strengths in guiding evaluation design to capture findings (intended and unintended) and causality (of varying complexity).

Explanatory Power of Four Types of Evaluation Frameworks

	What?		**Why?**	
	Intended Results	**Unintended Results**	**Simple**	**More Complex**
Taxonomies	Good	Blind spot	none	none
Logic Models	Good	Low	Good	Low
Open Learning Approaches	Good	Good	Low	Low
Cynefin	Less emphasis	Good	Less emphasis	Good

While the Kirkpatrick model points to areas of measurement that have proved popular, it is not well-suited to capturing unintended results or providing explanation for findings. Taxonomies have no ability to answer questions about why results are occurring. Evaluations conducted according to taxonomies may provide answers about causality and attribution, but they are not guided by the taxonomy in doing so. Kirkpatrick, and other taxonomies, are the least powerful option for an evaluation framework.

The purpose of the proposed M&E system is to provide information useful for improving performance related to gender analysis and planning. With taxonomies, and to a lesser extent logic models as well, questions and answers would be focused on making changes only to the training itself. With more powerful evaluation frameworks, one can identify additional elements in the organization that may also be adjusted to improve overall performance in the area of gender analysis and planning.

The Cynefin framework provides the explanatory power of both logic models and open learning approaches, with the added benefit of knowing in what aspects of the evaluation context they are appropriate. In this instance, the whole is greater than the sum of its parts. Moreover, the evaluation can address areas of interest comparable to those identified by Kirkpatrick, but with more explanatory power.

Limitations of this Review

The learning evaluation field is a wide one, and an in-depth look is beyond the scope of this modest report.

[12] Evans notes several alternatives to the Kirkpatrick model, but doesn't come out in favor of any of them.

References

Alliger, G. M., S. Tannenbaum, W. Bennett, H. Traver, and A. Shotland. 1997. "A Meta-analysis of the Relations among Training Criteria." *Personnel Psychology*, 50: 341-58

Brinkerhoff, Robert O. 2005. "The Success Case Method: A Strategic Evaluation Approach to Increasing the Value and Effect of Training." Advances in Developing Human Resources Vol. 7, No. 1 February 2005 86-101. Online version at: http://adh.sagepub.com/cgi/content/abstract/7/1/86

Cheveldave, Michael. 2009. *"How do you make sense of thousands of ideas or stories?"* Cognitive Edge Pte. Ltd.

El-Bushra, Judy. 1996. Gender Training in ACORD: Progress Report and Critical Assessment. RAPP, London, October, 1996.

Evans, David. 2007. "*Are There New Approaches to the Evaluation of Learning Programs?"* Report No. EG07-125 World Bank Institute. Washington, DC.

Geneva Centre for the Democratic Control of Armed Forces, United Nations International Research and Training Institute for the Advancement of Women, Office for Democratic Institutions and Human Rights. 2007. "Good and Bad Practices in Gender Training for Security Sector Personnel: Summary of a Virtual Discussion." June 2007.

Grove, John T., Barry M. Kibel, and Taylor Haas. 2005. "EvaluLEAD: A Guide for Shaping and Evaluating Leadership Development Programs." The Sustainable Leadership Initiative, a project of the Public Health Institute, Oakland, California. January 2005

Annica Holmberg et al. 2007. "Making Sexuality and Gender Visible: An Evaluation of the Methodology Project on Gender and HIV/AIDS, The Africa Groups of Sweden 2005-2007." Project funded by Swedish International Development Cooperation Agency.

Jewkes, Rachel et al. 2007. Evaluation of Stepping Stones: A Gender Transformative HIV Prevention Intervention. Policy Brief Stepping Stones. March 2007

Kurtz, C. F. and D. J. Snowden. 2003. "The new dynamics of strategy: Sense-making in a complex and complicated world." IBM Systems Journal, vol 42, number 3, page 462.

Lin, Vivian and Sally Fawkes (2006). Evaluation Report for the WHO Kobe Centre of the PROLEAD WHO Health Promotion Leadership Development Program. July 2006.

LINGOs (Learning for International NGOs). End of Course Evaluation, Level 1. Designing & Developing Virtual Classroom Training.

Mark, Melvin M. 2002. "What works and how can we tell?" Presentation in March 2001. The State of Victoria, Department of Natural Resources and Environment, Agriculture Division.

Milward, Kirsty.2007. "Revisiting studies and training in gender and development – the making and re-making of gender knowledge" (Report from the International KIT conference, 14-16 May 2007), October 2007.

Office of the Auditor General of Canada. 2009. “Chapter 1: Gender-Based Analysis” in Report of the Auditor General of Canada to the House of Commons. Spring 2009.

Ortiz, Alfredo and Peter Taylor. 2008. “Learning Purposefully in Capacity Development: Why, what and when to Measure?” Institute of Development Studies, 25 July 2008.

Philips, Jack and Ron Stone. 2000. How to Measure Training Results: A Practical Guide to Tracking Six Key Indicators, McGraw-Hill. New York.

Ringheim, Karin and Manuela Colombini. 2008. “Evaluation of the WHO Training Initiative: Transforming Health Systems: Gender and Rights in Reproductive Health, 1997-2007.” World Health Organization (GRR/RHR), 5 May 2008.

Sadik, Nafis et al. 2006. Gender Equality: Evaluation of Gender Mainstreaming in UNDP, Evaluation Office, United Nations Development Program.

Siwal, B.R. 2005. “Basic Framework and Strategy for Gender Training” Eldis Document Store, 2005, available at http://www.eldis.org/fulltext/DOC20260.pdf

Stienstra, Jochum and Wim van deer Noort. 2008. “Loser, Hero or Human Being: Are You Ready for Emergent Truth?” in ESOMAR Congress 2008 Research Papers. Online version at: http://www.communicatieplein.nl/dsc?c=getobject&s=obj&objectid=138775

Stuart, Rieky and Aruna Rao, Jeremy Holland. 2008. “Evaluation of SDC's Performance in Mainstreaming Gender Equality.” Report for the Swiss Agency for Development and Cooperation (SDC). May 2008.

Tilley, Nick. 2000. “Realistic Evaluation: An Overview.” Presented at the Founding Conference of the Danish Evaluation Society, September 2000.

UNDP. 2001. Introductory Gender Analysis & Gender Planning Training Module for UNDP Staff

UNESCO. 2003. Tips and Good Practices for Conducting Gender Training for UNESCO Staff. September 2003.

United Nations System Staff College (no year). Workshop on Evaluation and Impact Assessment (EIA) of Learning and Development. Presentation on 17-19 December to UN learning community. .

Williams, Bob and Richard Hummelbrunner. 2010. "Cynefin" in *Systems Concepts in Action*. Palo Alto: Stanford University Press