

Personal Benchmark Evaluation:
Assessing the Contribution of Training to Development Capacity

by Aliza Belman Inbal

B.A. 1990, University of Toronto
M.Phil. 1993, University of Cambridge
M.S.M. 2001, Boston University Brussels

A Dissertation submitted to

The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

January 31, 2011

Dissertation directed by

Kathryn E. Newcomer
Professor of Public Policy

The Columbian College of Arts and Sciences of The George Washington University certifies that Aliza Belman Inbal has passed the Final Examination for the degree of Doctor of Philosophy as of September 28, 2010. This is the final and approved form of the dissertation.

Personal Benchmark Evaluation: Assessing the Contribution of Training to Development Capacity

Aliza Belman Inbal

Dissertation Research Committee:

Kathryn E. Newcomer, Professor of Public Policy and Public Administration,
Dissertation Director

Jennifer Brinkerhoff, Professor of Public Administration and International
Affairs, Committee Member

Ryan Watkins, Assistant Professor of Educational Technology Leadership,
Committee Member

@ Copyright 2011 by Aliza Belman Inbal
All rights reserved

Abstract of Dissertation

Personal Benchmark Evaluation: Assessing the Contribution of Training to Development Capacity

In recent years, the international community has increasingly prioritized capacity building as essential to development. Yet, lack of capacity is still commonly cited as one of the primary reasons for poor development outcomes in many countries. Given the limited success of many capacity development initiatives to date, there is increasing recognition of the lack of adequate evaluative tools for both measuring the success of capacity development initiatives and learning how to improve these initiatives. This dissertation focuses on one of the most prevalent types of capacity development interventions, that of training. It includes a survey of 20 of the top development training institutes, mapping present practice and unmet evaluation needs. It then presents a new, participatory model for evaluation of international training courses targeting developing country professionals. The model is designed to address some of the limitations to feasibility, utility and credibility of current evaluation models through the use of participatory evaluation methods. The new "personal benchmark" evaluation model includes three participant questionnaires: pre-course, end-of-course, and several months post-course, with participants at each stage defining for themselves their goals, and then in the next phase assessing the extent to which they were reached. The personal benchmark system is designed to both support evaluation of training and to facilitate better course design and continued improvement of training. The dissertation presents results from initial field testing of the model and makes recommendations for needed research in order to validate the model and test its applicability of the model in other contexts.

Table of Contents

Abstract of Dissertation	iv
List of Figures	vi
List of Tables	vii
Glossary of Terms	viii
Chapter 1: Introduction	1
Chapter 2: A Review of the Literature on Training Evaluation	26
Chapter 3: Methods	85
Chapter 4: Development Training Institute Interviews and Building of the Conceptual Model	112
Chapter 5: Field Test of Dissertation Evaluation Model	142
Chapter 6: Discussion and Conclusions	170
Bibliography	190
Appendix A: Summary of Capacity Building Studies	218
Appendix B: List of Interviewed Development Training Institutes	223
Appendix C: MCTC Gender and Local Governance Course Description	224
Appendix D: MCTC Support Systems for Microenterprises Course Description	227
Appendix E: Results of Development Training Institute Interviews	233
Appendix F: Gender Course Pre- and End-of Course Questionnaires	239
Appendix G: Microfinance Pre-Course and End of Course Questionnaires	244
Appendix H: Structured Course Organizer Assessment Questionnaire	248

List of Figures

Figure 2.1: The IEG Three-Tier Training Evaluation Model	77
Figure 4.1: Prevalence of Levels of Evaluation	115
Figure 4.2: How Do DTIs Learn?	116

List of Tables

Table 1.1: Definition of Capacity Development by International Development Agencies and Research Institutes	7
Table 2.1: Summary of Evaluation Needs from Literature Review	83
Table 3.1: Interview Protocol for DTIs	94
Table 4.1: Summary of DTI Evaluation Needs	128
Table 5.1: Learning Goal Table	144
Table 5.2: Training Use Goal Table	147
Table 5.3: Post-Course Evaluation Questionnaire	149
Table 5.4: Sample of Open-Ended Responses in Post-Course Questionnaires	158
Table 6.1: Does Personal Benchmark System Address Identified Needs?	181

Glossary of Terms

ADB	Asian Development Bank
CD	capacity development
CIDA	Canadian International Development Agency
DAC	Development Assistance Committee of the OECD
DTI	development training institute
ECDPM	European Center for Development Policy Management
IDRC	International Development Research Centre
IEG	Independent Evaluation Group of the World Bank
IMF	International Monetary Fund
INGO	International Non-Governmental Organization
ITCILO	International Training Center of the International Labor Organization
JAI	Joint Africa Institute
JICA	Japan International Cooperation Agency
MCTC	Mount Carmel Training Center
OECD	Organization for Economic Cooperation and Development
SIDA	Swedish International Development Agency
TA	technical assistance
UNDP	United Nations Development Program
WBI	World Bank Institute

Evaluating the Contribution of Training to Development Capacity:

A Personal Benchmark System for Training Evaluation

Chapter One: Introduction

This dissertation explores methods for evaluating the contribution of training to organizational and institutional capacity building in the developing world. It assesses current methods, analyzing their weaknesses, and proposes a new model for training evaluation.

In recent years, the international community has increasingly prioritized capacity building as essential to development. While a range of interventions are used to build capacity, such as policy advice, on-the-job technical assistance and mentoring schemes, training remains one of the primary tools used to achieve capacity enhancement goals. The international donor community has invested an estimated 25% of all development funds in capacity building in recent years (Whyte, 2004:10), or about 24.6 billion dollars annually – almost double the amount spent on capacity building in 1969 (Fukuda-Parr, Lopes & Malek, 2002:4). Yet, lack of capacity is still commonly cited as one of the primary reasons for poor development outcomes in many countries (OECD, 2006:11, Watson, 2006:4). Today, lack of government institutional capacity is recognized as one of the main impediments to achieving the Millennium Development Goals -- the eight development targets which UN member states have committed to achieving by 2015 (World Bank, 2004: 51).

Given the limited success of many capacity development initiatives to date, there is increasing recognition of the lack of adequate evaluative tools for both measuring the success of capacity building initiatives and learning how to improve these initiatives. As was noted by Whyte in a landscape review of donor trends in capacity building, donors have paid considerably less attention to monitoring and evaluating their capacity development support than to evaluating other types of programming. This omission has been attributed in part to the lack of evaluative tools that can accurately assess the contribution of any one intervention to capacity building goals (Earl, 2002:5; Whyte, 2004:19).

At first glance, training differs from other forms of capacity building interventions, in that as there is a very well-developed and almost universally-used model for evaluating training: the “Kirkpatrick model”. Donald Kirkpatrick’s model was originally developed in the United States for corporate training in the US and other developed countries (Kirkpatrick, 1998). It proposes evaluation on four levels, starting with participant post-course satisfaction and then examining results at the learning outputs, workplace behavior change outcomes and organizational impact levels.

It is questionable, however, whether the Kirkpatrick model provides adequate tools to measure the results of training in developing countries, particularly when training is targeted towards public sector or local non-governmental organization (NGO) personnel. Such contexts differ in several important ways from the developed-country, corporate contexts that Kirkpatrick had in mind when developing his model. Measuring training impact is likely to be more difficult in public sector and non-profit institutions, which tend to have much more complex, less clearly defined and less easily measured

goals than profit-oriented businesses. In addition, whereas in developed country businesses, the organization benefiting from training is generally the one that finances the training, in development contexts, training funders, beneficiaries and providers are more often than not separate entities located in different countries. As is discussed in this dissertation, the separation between training funders, providers and beneficiaries can make gathering accurate, useful evaluative information on training both more challenging and more costly. This dissertation analyzes the utility of Kirkpatrick and other existing models for evaluation of the contribution of training to capacity building to public sector organizations in developing countries, and offers a new model that might address shortcomings in present practice.

This chapter provides the policy context for the dissertation research. It begins by establishing working definitions for certain terms used within the dissertation. The chapter then explores current thinking on capacity building for international development, paying particular attention to both the definition and role of capacity building in strategies for social, economic and political development, and to the evidence of the success of the international community's attempts at supporting capacity building of developing world agents. This sets the stage for the dissertation's discussion of monitoring and evaluating training's contribution to the capacity of developing world organizations and institutions. It does so both by demonstrating the importance of capacity building for development and identifying common problem areas in current practice that are relevant for the ensuing discussion on training evaluation. Finally, the chapter presents the main research questions to be addressed in this dissertation, briefly

outlines the methodology to be used, and then discusses the intended contribution of the research to practice and theory.

Definitions

Before beginning the discussion on evaluation of training for capacity development, it is useful to clarify what is meant by certain terms used frequently in the course of this dissertation:

Capacity building/capacity development: As is discussed later in this chapter, there are many different definitions currently used for these terms. Perhaps the most basic of these is the UNDP definition: "The ability of individuals, institutions and societies to perform functions, solve problems, and set and achieve objectives in a sustainable manner. Capacity development is the process through which the abilities to do so are obtained, strengthened, adapted and maintained over time" (2007:2). There has been some debate in the international development community on the terminology of "capacity building". UN agencies, for example, now prefer the term "capacity development", to "capacity building", on the premise that the latter term better implies building on and enhancing existing country capacities. Other organizations, including the World Bank, use the terms interchangeably. This dissertation takes the World Bank approach.

Development training institutes (DTIs): donor-funded institutes primarily training participants from the developing world with the aim of building capacity to achieve social, political or economic development goals.

Donor-funded: For the purposes of this dissertation, donor-funded training and/or capacity building interventions should be taken to include activities financed by any foreign development cooperation agents, including bilateral donors, multilateral organizations, international NGOs and international financial institutions such as the World Bank and the International Monetary Fund (IMF).

North/south: In this dissertation “the North” is used (in accordance with practice in much recent international development literature) to refer to high-income countries, and “the South” to refer to low and middle-income countries, regardless of their actual geographic location. Thus, for example, Australia would, in this definition, fall into the category of Northern countries, and Mongolia into the category of Southern countries.

Training – This dissertation used the World Bank Independent Evaluation Group's definition of training (2008: 5), as learning programs which (a) have specific learning objectives (b) are scheduled activities occurring in a setting other than where the knowledge will be applied, and (c) are conducted by lecturers, trainers or facilitators.

The Importance of Human Capacity Building for Development

International development strategies have evolved over the past few decades to place increasing emphasis on national capacity building as a key driver of development. Post-WWII development theorists tended to see development as being solely a matter of investment. It was commonly believed that, for developing countries to progress, all that was needed was an influx of financial capital which could be used to build infrastructure and create jobs. By the late 1970s, however, the developing world was on the verge of crisis after decades of borrowing foreign capital to finance large-scale infrastructure

projects had not only failed to jump-start economies, but had resulted in crippling debt for some countries. Faith in the power of investment was replaced in academic and practitioner literature with an insistence on the necessity of economic policy reform for growth (World Bank, 1979).

Throughout the 1980s and through much of the 1990s, donors used structural adjustment programs to compel developing countries to liberalize their economies, balance their budgets and shrink their governments. However, when these reforms also failed to result in significant economic progress in the least developed countries, it became apparent that, just as capital investments not guided by sensible public policies would not result in transformation of the developing world, so economic reform would fail if policies were imposed from the outside without the indigenous state capacity to adapt policies to local needs and to implement them in an efficient, effective and transparent manner (Fukuyama, 2004; Levy, 2004:1; Madavo, 2005: 23; World Bank, 1997.)

In response to the failure of past approaches, the fundamental importance of building national capacity for development has been emphasized by development theorists and practitioners alike since the early 1990s. Thus, for example, capacity development of aid beneficiaries was particularly emphasized in the Paris Declaration - the OECD's major 1992 policy declaration on new orientations for development assistance of donor countries (OECD, 1992:9). And, in 1999, the U.N. General Assembly called for more systematic attention to capacity building by all U.N. agencies.

While there is no one internationally sanctioned definition for capacity, or capacity development, the international community seems to be moving towards a shared

understanding of what is meant by the term. Table 1.1 provides some examples of how capacity and capacity development have been defined by various international development agencies and research institutes:

Table 1.1: Definition of Capacity Development by International Development Agencies and Research Institutes

UNDP	Capacity is the ability of individuals, institutions and societies to perform functions, solve problems, and set and achieve objectives in a sustainable manner. Capacity development is the process through which the abilities to do so are obtained, strengthened, adapted and maintained over time.” (2007:2)
OECD/DAC	Capacity is the ability of people, organizations and society as a whole to manage their affairs successfully (2006: 12). ‘Capacity development’ is understood as the process whereby people, organizations and society as a whole unleash, strengthen, create, adapt and maintain capacity over time. (2006:9).
CIDA	Capacity building includes activities, approaches, strategies and methodologies that help organizations, groups and individuals to improve their performance, generate development benefits and achieve their objectives over time. It often involves broad participation, building on local interests and expertise, offering opportunities for learning and linking at micro, meso and macro levels to build ownership and sustainability (Cited in Whyte, 2004: 24).
ECDPM	Capacity is that emergent combination of attributes that enables a human system to create development value (Morgan, 2006a:8).
ADB	Securing a country’s ability to manage its own affairs....capacity development symbolizes a shift toward measures that enable national empowerment and a more balanced relationship among funding agencies, national governments, and other organizations.” (2008: 2).
JICA	Capacity development refers to the ongoing process of enhancing the problem-solving abilities of developing countries by taking into account all the factors at the individual, organizational, and societal levels. (2006:2).

Notable in these definitions of capacity development is they all make the link between capacity and functional purpose. While capacity is inherently a potential state, “latent as opposed to kinetic energy” in the words of one researcher (Morgan, 2006a:7), it

is not without purpose. Each of the definitions listed in Table 1.1 refers to capacity as being in service of a particular active purpose; “to create development value”, “to make development possible”, “to improve their performance, generate development benefits and achieve their objectives over time”, “to set and achieve objectives” and so on. In other words, “capacity” is always “capacity to...” Implicit in this conception of capacity is the expectation that capacity is to be used. In this way, capacity differs from knowledge, and capacity building goes beyond learning.

Also notable in many of the definitions offered in Table 1.1 is the multi-level nature of capacity building. Capacity building may be pitched at the level of individuals, communities and organizations, institutions and inter-organizational networks, country, society, region, and/or the overall enabling environment. Donors may differ in terms of how they define these levels of capacity building, as well as which are the principal levels that they target in their capacity building interventions (Whyte, 2004:26). However, there is broad recognition that multiple levels of intervention are likely to be necessary to bring about change. For example, the World Bank’s 1997 World Development Report “The State in a Changing World” defined state capability as including both “the administrative or technical capacity of state officials” and the “deeper, institutional mechanisms that give politicians and civil servants the flexibility, rules, and restraints to enable them to act in the collective interest.” (World Bank, 1997:77). A World Bank Independent Evaluation Group (IEG) evaluation of capacity building in Africa defined capacity building as necessarily involving three levels: human (individual skills), organizational (the internal structures, processes, systems, staffing, and other resources to achieve goals) and institutional (the rules of the game) (IEG, 2005:7).

The UNDP, for its part, has defined capacity building as necessarily encompassing individual skills, the institutional environment in which individuals operate, and the opportunities available at the societal level that enable people to use and expand their capacity (Fukuda-Parr, Lopes & Malek, 2002). Like the World Bank, the UNDP has acknowledged the necessity of addressing capacity challenges at more than one level. As argued in a 2006 UNDP Capacity Practice Note: “Attempts to address capacity issues at any one level, without taking into account the others, are likely to result in developments that are skewed, inefficient and, in the end, unsustainable.”

A final notable aspect of these definitions of capacity building is that many of them refer to it as a dynamic process over time rather than as the result of a few discrete interventions. This reflects an emerging consensus in support of a dynamic “systems perspective” to the issue of capacity building (Morgan, 2006b). Systems perspectives conceptualize “capacity building” very differently from more traditional “activity-based” conceptions. The activity-based conception of the nature of capacity building defines capacity building according to the type of intervention – in this conception, training, technical assistance and mentoring might all be seen as forms of capacity building, whereas, for example, budgetary support to organizations would not be seen as capacity building, even if this support enhanced an organization’s capacity to achieve its goals (OECD, 2006).

Systems-based perspectives differ in that they define capacity development not according to the type of intervention but according to its goals. According to this school of thought, capacity refers to the ability of individuals, organizations and/or institutions to achieve their goals (DeLange & Feddes, 2008). Proponents of systems perspectives on

capacity development argue that enhancing an organization's capacity to achieve its goals is first and foremost an endogenous process within institutions requiring not only, or even always, knowledge transfer, but also changes in organizational structures, power relationships, incentives, and/or resources to succeed. In this conception, for example, not only training and technical assistance, but also budgetary support, purchase of equipment, and policy advice may be seen as capacity development interventions, because all may enhance an organization's capacity to achieve a given development goal (De Lange & Feddes, 2008; Morgan, 2006b; Taylor et. al., 2008; Whyte, 2004).

Donor, professional and academic literature on capacity building today tends to look at capacity building from this “dynamic systems” perspective rather than the “activity-based” perspective, recognizing that capacity building generally requires a multifaceted approach, addressing the capacities, resources and incentives relevant to the achievement of any given goal at multiple levels, including the individual, organization, institution and system. (De Lange & Feddes, 2008; IEG, 2008, JICA, 2006:2; Madavo, 2005; OECD, 2006, Taylor et. al,2008:6). However, international development structures still lend themselves more to the “activity-based” input-driven perspective. There is a strong tendency for international development actors to equate capacity building with training and technical assistance. Thus, for example, the training institute of the World Bank - the WBI has long claimed to be the “capacity building arm” of the World Bank, despite the fact that its activities to day have been almost exclusively restricted to training.

Indeed, there is considerable evidence to suggest that, despite rhetoric about the necessity for more dynamic, systemic approaches to capacity, the reality as manifested in

development capacity-building initiatives still remains firmly ensconced in the ‘activity-based’ paradigm (Morgan, 2006a:4). The development cooperation landscape is populated with far more training institutes and technical assistants offering discrete, single-level interventions than capacity building practitioners who are able to advise on and support dynamic, integrated, multi-level processes of endogenous change.

This tendency towards an “activity-based” approach to capacity building influences the way that capacity building is financed and evaluated. Donor, UN and international financial institution (IFI) budgetary structures generally account for allocations towards training or technical assistance, but not for capacity building as a whole. For this reason, while, as mentioned earlier, the figure of 25% of donor funds is often cited as the amount spent on capacity building, this figure is based on a calculation of training and technical assistance rather than on a broader definition of the range of activities and support that might be necessary to help individuals, organizations or institutions achieve their goals (Taylor et. al., 2008). In addition, evaluations tend to look at the direct results of specific interventions without taking a broader look at the contextual factors that may have affected the extent to which those interventions could be effective.

This gap between the theory of capacity development as a multi-faceted, systemic process within organizations, and the practice of viewing capacity development as something that can be achieved through one-off discrete, external interventions is only one way in which donor practice does not conform with capacity development theory. Indeed, much recent practitioner literature points to similar gaps between theory and

practice in several other aspects of capacity development. As Whyte's review of donor, foundation and international institution-financed capacity building found:

There is an emerging consensus among donors about the dos and don'ts of capacity building. In fact, the review of the literature rather depressingly suggests that the success factors for capacity building have been known for at least a decade and are reiterated or recast by later reviewers rather than being newly discovered insights (Whyte, 2004: 73).

What is "depressing", to Whyte and others, is that this shared understanding of good practice has not led to substantially improved results. As observed in a comprehensive 2001 UNDP study on capacity building (Fukuda-Parr, Lopes & Malek, 2002):

Donors can ship out four-wheel-drive vehicles, or textbooks, or computers; they can dispatch expatriate experts...they can run multiple seminars and courses that improve the individual skills of thousands of people. However, the capacity of local institutions and of countries as a whole has still not appeared adequate to meet the challenges of development. There have been positive micro-improvements, but not the kind of macro-improvements that build and sustain national capacity for development. (p. 3)

The reasons for the development community's failure to substantially improve capacity in much of the developing world was discussed in a recent meeting of several research institutes from both Northern and Southern countries, donors, and foundations on improving capacity development, hosted by the Institute of Development Studies in Sussex (Taylor & Clarke, 2008). Meeting participants suggested that one of the primary

reasons for the poor overall results of the international development community's capacity building results lies in the gap between theory and practice of capacity development. "There are cases where values, principles and concepts of Capacity Development are expressed in project or program frameworks, but these rarely are seen to translate into practical action...Our (literature) survey and discussions have shown quite strongly...that what is known does not necessarily translate into what is done." (Taylor et. al., 2008:11).

In order to determine whether such a gap between theory and practice indeed exists and, if so, what are the main ways in which it is manifest, the dissertation author completed a review of recent practitioner studies and evaluations on capacity development practice (for the full review, see Appendix A). The author looked at all available multi-program studies by major development organizations, such as the UNDP, OECD-DAC and the World Bank, in the 2002-2008 period, that have assessed past capacity development practice and made recommendations with regard to future practice. The review found almost complete agreement between the eleven studies reviewed on the major shortcomings of capacity development interventions. The list of the identified common issues strikingly illustrates the present gap between the theory and practice of capacity building for development, and the effect of this gap on the results of capacity building endeavors.

Analysis of the studies reveals a reality which is far from the ideal of capacity building as a dynamic, endogenous, goal-oriented long-term process whose goals are accomplished through integrated support addressing individual, organizational and institutional capacity gaps. The following is a synthesis of the primary themes common

to the studies reviewed here regarding how practice falls short of capacity development theory.

There is a need for capacity building interventions that better address specific country or organizational needs. All of the studies reviewed called for better assessments of needs on all programming levels. Needs assessments are essential to ensure that capacity development programs build on present organizational strengths and address existing capacity gaps impacting organizational performance. According to all of the reports reviewed, more thorough needs assessment can and should be conducted on multiple levels. The OECD's Development Assistance Committee (DAC) (2006) points out the need for better understanding of country and societal needs, goals and characteristics as a fundamental building block of better capacity development programs. In addition, the DAC stresses the importance of using needs assessment to better target interventions to specific organizational needs. As stated in a DAC publication on good practice in capacity building: "Capacity needs assessments should begin with the question, 'Capacity for what?' and avoid the trap of providing generic training on broad topics, disconnected from the capacity and performance of specific organizations." (2006:8) The UNDP (2002) asserts that current capacity development interventions based on transfer of "textbook" knowledge rather than an understanding of local context and knowledge are "catastrophically wrong." The World Bank's Independent Evaluation Group (IEG) for its part (2005; 2008), finds that capacity development interventions are rarely designed on the basis of organizational assessments that first determine what organizational needs ought to be targeted, and only after specify the best combination of interventions to address these needs, and, finally, determine in detail the knowledge,

skills and attitudes that must be developed in target organizations in order to achieve capacity development goals. The Asian Development Bank (ADB) (2006) and European Center for Development Policy Management (ECDPM) (Land, Hauck, & Baser, 2009) both affirm that better emphasis must be placed on organizational assessments which also map present capacities in order to make it possible to build on what exists. JICA's 2006 study of capacity development programs underlines the importance of sufficient analysis of environmental challenges and constraints to application of learned capacities in workplace contexts as an important part of needs assessment. In sum, the studies reviewed all suggest that donor-funded interventions are not based on adequate understanding of what capacities exist, what "capacities to..." need to be enhanced for the achievement of specific development goals and what means are best to achieve capacity development goals.

There is a need for more integrated capacity building support that addresses individual, organizational and institutional capacity gaps in a well-sequenced, holistic manner, as well as greater attention to systemic attributes such as governance, the policy environment, incentives and political and social aspects of capacity building. In other words, not only should better needs assessments be conducted at all levels, but also good capacity development should endeavor to address these multiple levels of needs through well sequenced programs involving a range of needed interventions. All of the studies reviewed emphasize that interventions that build human capacity without addressing factors at organizational, institutional and country levels such as policy, power, and resources are likely to fail. This has been a major theme of the DAC's work

on capacity development over the past few years. Thus, for example, in its 2006 guide to good capacity development practice, the DAC states (OECD, 2006):

The traditional “capacity building” tools of technical cooperation and training have often proved ineffective in helping to improve performance because they have not been linked to the necessary organizational and institutional developments...This implies approaching capacity development in an integrated way, so that individual skills and the organizational settings in which they can be put effectively to work are created simultaneously. (p. 30)

The importance of integrated capacity building programs is echoed by the UNDP (2002; 2003) and the World Bank (IEG, 2005; IEG, 2008). The ECDPM, for its part, has devoted considerable effort over the past few years to mapping elements of organizational, institutional and country capacity which it contends must be better addressed in order to improve the success rates of capacity development interventions (2009) and JICA (2006) stresses the need to better address external environmental factors likely to impede the application of skills and knowledge gained through capacity development programs.

There is a need for greater country and organizational ownership of capacity building and a more endogenously-led process. Indeed, lack of sufficient ownership was cited as a primary concern in all eleven of the studies reviewed. For example, the UNDP's 2002 study of capacity development finds that technical cooperation is generally driven more by donor supply than beneficiary demand, leading to a lack of commitment on the part of the beneficiary. Lack of sufficient beneficiary ownership is attributed in

the Asian Development Bank (ADB) study (2008) to lack of good mechanisms for building ownership and demand. Without strong beneficiary input, it extremely difficult to do accurate assessment of needs (IEG, 2007). In addition, to the extent that capacity development is an endogenous process occurring within organizations, without strong levels of commitment to capacity development goals within the organization, capacity development interventions are likely to fail (Land, Hauck, & Baser, 2009; OECD-DAC, 2006; UNDP, 2003).

There is a need for longer-term capacity development programs rather than short-term, one-off events. This theme is related to the above two issues. In-depth needs assessments are often feasible only for longer-term programs where the cost of the assessment would not be disproportionately large in relation to the size of the program itself. Second, longer-term relationships are often needed in order develop trust and effective collaboration between program partners and to reinforce beneficiary ownership (JICA 2006). Moreover, as capacity is generally a complex endogenous process occurring on multiple capacity levels (human, organizational, institutional and societal), one-off, short term interventions are unlikely to be sufficient to address the multiple types of capacity gaps that may need to be addressed to achieve capacity goals. Among the organizations which have called for a shift to longer term programs have been the UNDP (2006) and JICA (2006).

Monitoring and evaluation is a particularly poor aspect of capacity development program management. Among the problems with present monitoring and evaluation regimes, cited by the reviewed capacity development studies, are lack of systematic data collection (ADB 2008, JAI 2008), insufficient understanding of useful

indicators of capacity development (JICA 2006; Land, Hauck, & Baser, 2009), insufficient focus on results measurement (World Bank, 2005) and lack of quality control of programs (IEG, 2005). Lack of evaluation is a matter of concern for two reasons. First, good monitoring and evaluation schemes are important for program learning: they can help improve the success of capacity building programs by establishing feedback loops which can enhance present and future capacity building interventions. Second, monitoring and evaluation schemes can help program stakeholders (donors, beneficiaries and others) assess the extent to which donors' capacity building policies are indeed translated into capacity building practice, and to hold to account practitioners that do not meet these policy standards.

All but one of the studies reviewed here look at capacity development interventions of all types, including training, technical assistance and other forms of capacity-development related programming. The only study completed in this period which looked specifically at training interventions delivered by a range of training institutes (as opposed to a review of the courses of one training institute) is the comprehensive World Bank IEG review of training led by the dissertation author, completed in 2007. The IEG flagged the same problem areas as the other capacity development studies cited here. Specifically, the IEG review found that among the main failures in training management practice were: insufficiently targeted training courses due to lack of adequate needs assessment, lack of sufficient "client" ownership of training and input into training content, lack of integration of training with other forms of capacity development interventions in order to address capacity gaps likely to impede the transfer of learning to workplace environments, and lack of efficacy of one-off, short-term

courses as a means of capacity development. In addition, the IEG study found that, of eight training management processes reviewed by the study (including, among others, needs assessment, participant selection, training design, provision of needed follow-up support), the one training management process that was almost invariably poorly performed was that of monitoring and evaluation. The IEG study identified the lack of adequate monitoring and evaluation as a significant shortcoming of 29 of the 37 World Bank financed training programs reviewed, but did not discuss reasons for the lack of good monitoring and evaluation of training for capacity development. Nor did the study explore how the international community might try to improve training evaluation practice in international development contexts. As is discussed in greater length below, this dissertation takes as its point of departure these two questions.

In sum, the dissertation author's review of 11 of the most comprehensive reviews of capacity development programs done by major donors, think-tanks and international organizations in the 2002-2008 period identified several common areas where international community capacity development practice falls far short of theories of successful capacity development, including the area of monitoring and evaluation. This dissertation attempts to directly address the lack of adequate monitoring and evaluation of capacity development. It identifies present weaknesses in present-day development training institute (DTI) monitoring and evaluation practice, assesses the reasons for those weaknesses and proposes an improved model for monitoring and evaluation of training. In addition, while it is outside of the scope of this dissertation to present a full analysis of how to address identified shortcomings to capacity development practice, beyond that of training evaluation, the dissertation touches how a monitoring and evaluation regime can

be leveraged in order to help address other shortcomings, such as cost infeasibility of in-depth needs assessment, low levels of ownership, and the need to address associated organizational and environmental capacity constraints that affect implementation of learning in the workplace. All of these constraints are discussed in greater detail in the context of this dissertation.

Research Outline and Intended Contributions

In order to build a new model for evaluation of training in international development contexts, the dissertation explores two research questions:

Q1. What are the existing weaknesses in present models for evaluating training in development contexts?

Q2. What are the key elements of an evaluative model that could feasibly provide more useful and credible information for learning and accountability purposes?

On the basis of the data gathered to answer the first question, a new "personal benchmark" training evaluation system was developed. Research on present practice and on potential improvements specifically targeted DTIs, as defined earlier in this chapter. The research focused on evaluation of courses that met the following three criteria, due to both the prevalence of such courses in international development and the particular difficulties associated with evaluating them: (a) Courses of short duration (three days to three months); (b) courses with participants coming from multiple organizations or countries; (c) courses aimed at supporting participants' performance in their present workplaces through building of awareness, knowledge or skills.

The decision to focus research on courses meeting the above three criteria was informed by the brief review of capacity development literature summarized above. Despite the fact that the international development policy community has for the past decade strongly advocated transitioning away from short-term one off training interventions towards longer-term, multi-level capacity development programs, short-term, one-off courses are still the most dominant form of training done by DTIs. Thus, this dissertation took as its point of departure present training institute practice rather than the theoretical ideal discussed in the literature of long-term, multi-level, strategic capacity development programs offering a range of different sorts of support.

By answering the above research questions, the dissertation endeavored to contribute to the academic and practitioner literature on training evaluation and evaluation of capacity development in international development contexts. First, research aimed to provide a map of present DTI evaluation practice, constraints and needs. To date, no survey of DTIs has been undertaken to map present practice. In this, the international development field differs greatly from Northern, corporate contexts which have benefited from several academic and practitioner surveys mapping the prevalence of various types of training evaluation (Brandenburg, 1982; Parker, 1986; Saari et al., 1988; Training, 1996). A more detailed understanding of present DTI practice and unmet evaluation needs can facilitate academic and practitioner dialogue on how to improve DTI monitoring and evaluation practice. Second, the dissertation presents a possible system for addressing identified evaluation needs and constraints. In doing so, it attempts to go beyond Kirkpatrick's four-level model to provide evaluative tools tailor-made to DTI needs and realities.

As is detailed in Chapter Three of this dissertation, the personal benchmark system was developed using the following methodology: First, a comprehensive literature review on training evaluation practice examined relevant academic and practitioner perspectives on training evaluation methodology - in both Northern corporate and Southern development-related contexts. Second, 20 of the largest and most reputable DTIs in a variety of fields and geographical regions were interviewed with regard to their evaluation practices. These interviews were used to provide data on evaluative needs, constraints and unique good practices of DTIs. Third, a new theory of training evaluation, leading to the design of new evaluation system, was developed on the basis of the literature review and DTI interview findings. Fourth, this model was applied in two international training courses in the Mount Carmel Training Center in Haifa, Israel, in order to gain further insights into its feasibility and utility.

Summary Description of the Personal Benchmark System

The system, as developed in this dissertation, consists of three questionnaires. The first questionnaire is administered to training participants and their supervisors prior to the course to gather information on their workplace contexts and learning needs, asking them to define their learning goals and to explain how they hope this learning will enhance participants' ability to do their work. The second questionnaire, administered at the end of course to participants, asks participants to review the individual learning goals which they set for themselves at the beginning of the course and rate the extent to which they believe that they have achieved these goals. The questionnaire then asks participants to name the three most useful things that they gained in the course and to explain how they expect to use these things in their workplace. Finally, the third

questionnaire, administered several months to one year after course completion, redistributes to participants their responses from the end-of-course questionnaire asking them to comment on the extent to which they were able to use what they learned at work as expected. Where they have not been able to apply what they learned as expected, the questionnaire asks what were the main obstacles which they faced to workplace application of learning.

There are a number of unique features to the model proposed here. First of all, it is highly participatory, asking participants to define for themselves learning and training use goals rather than defining for participants the parameters by which the course will be evaluated. Participatory methods are employed in order to raise participant ownership of the evaluation process with the aim of improving response rates and generating richer, more nuanced information on the effects of training on participants from very different country and workplace contexts. Second, it supplements participant's qualitative responses with quantitative measures of those responses. Rather than having standardized close-ended questions asking participants to rate training results, it ask participants to provide ratings to the individual learning and workplace use goals, which they had previously chosen for themselves. This is intended to give evaluators a more nuanced understanding of what sort of training results the quantitative numbers represent. Third, it includes a compulsory participant needs assessment phase for all courses. This is intended in part to support continuous improvement of training courses over time. Finally, it places particular emphasis on determining not only workplace application of training but also the environmental context and constraints in which participants try to apply training, in order to either better integrate training in longer-term organizational

capacity development programs or, at least, to highlight the need for such ongoing interventions.

These aspects of the personal benchmark system were also intended to address the six common shortcomings of capacity development interventions identified in this chapter's review of capacity development studies. The system directly addresses the issues of lack of needs assessment, participant ownership and monitoring and evaluation. In addition, by gathering data on how workplace environment affects training use, the model can be used to identify human, organizational and institutional capacity gaps that may hinder the application of training, thereby providing a basic roadmap for the sort of longer-term complementary capacity development interventions that may be necessary to enable training success. In other words, an attempt was made in design of the system to take into account these other needs through a holistic approach fusing together program planning and instructional design needs as well as developmental, formative and summative evaluation. However, this study focuses on analyzing the contribution of the personal benchmark system to better summative and formative evaluation alone. While the personal benchmark system was designed to facilitate program planning and instructional design as well as evaluation of training, it was considered beyond the scope of this dissertation to explore evidence on how the system may be used for these purposes.

Chapter Outline

This dissertation has six chapters. After this introductory section, Chapter Two presents a literature review in three parts. The first part of the literature review discusses quality criteria for evaluation, asking what should be the goals of evaluation and by what

standards evaluation models should be assessed. The second part assesses current models of training evaluation, focusing on the Kirkpatrick model, as the dominant one in use today. It analyzes possible shortcomings of the Kirkpatrick model in general and, in particular, for developing country, public sector contexts. The third part of the literature review presents alternatives to the Kirkpatrick model from the literature, as well as some newer models for evaluating capacity development that may provide useful tools and ideas for a new training evaluation model.

Chapter Three defines the research questions addressed in this dissertation and presents the methodology that was used to explore existing weakness in training evaluation and develop a new evaluation system for use in DTIs. It also presents potential limitations to the research methodology. Chapter Four then present results of the DTI interviews and the logical process by which a new theoretical model of training evaluation was constructed and Chapter Five details the conduct and results of a field test of that evaluation model. Chapter Six discusses the results of research, draws conclusions with regard to how training evaluation may be made more feasible, useful and accurate, discusses research limitations and identifies directions for future research.

Chapter Two: A Review of the Literature on Training Evaluation

The following literature review is divided into three distinct parts. The first part outlines the desired purposes of evaluation and quality criteria for valuation of evaluation methods. The purpose of this is to define the standards by which the personal benchmark system was evaluated. The second part presents the literature on the most broadly used training evaluation model – that of Kirkpatrick - including in-depth analysis of the possible weaknesses of this model. Finally, part three paves the way to elaboration of a new evaluation model by looking at both academic and practitioner literature on alternative evaluation tools and methods which go beyond the Kirkpatrick model and could enhance the usefulness of training evaluations.

Evaluation Purposes and Quality Criteria

This section defines the purposes of evaluation and the quality criteria by which evaluative models may be assessed. These criteria guided the dissertation’s assessment of present and possible evaluation models.

What is the purpose of evaluation?

Broadly speaking, there are two widely-cited categories of evaluation goals: summative and formative (Edelenbos & van Buren, 2005; Iverson, 2003; Patton, 1986; Russ-Eft & Preskill, 2001; Scriven, 1991). Summative evaluation, often done after program completion, aims at assessing the impact or effectiveness of the program. Summative evaluations are primarily concerned with accountability: the extent to which programs can correctly claim to have achieved their goals. Summative evaluations are useful for decision making on continuation of the program or future use of similar ones.

Formative evaluation, for its part, is aimed at gathering information that can help improve and enhance programs. Formative evaluations are also referred to in the literature as “learning” evaluations. Learning evaluations are aimed more at understanding what must be done to achieve objectives than at, in the words of Edelenbos and van Buren (2005:603), “determining the score”.

In addition to formative and summative evaluation goals, Michael Quinn Patton has suggested a third category of evaluation, that of developmental evaluation (Patton, 1996; Russ-Eft & Preskill, 2001:20). If summative evaluations ask: “What has the program achieved?”, and formative evaluations ask: “What can be done to improve the program?” developmental evaluations ask the question: What is to be done and how can it be achieved?. Developmental evaluation largely occurs in the planning stages of a program or in its early phases. Questions addressed in this form of evaluation include: What are the right sets of activities, processes and strategies for this program? What is the best program design? What goals and objectives should the program have? Developmental and formative evaluations are similar in that they look at the process of implementation rather than assessing results. In fact, many theorists do not recognize developmental evaluation as a distinct category, separate from formative evaluation (Edelenbos & van Buren, 2005; Iverson, 2003; Patton, 1986). However, Patton (1996) distinguishes between the two by claiming that they serve different purposes:

Formative evaluation remains connected to summative evaluation and defines a preliminary stage of assessment, while the program and its evaluation are getting ready for the real thing: summative judgment...
Developmental evaluation serves development-oriented programs that

have as their purpose the vague, general notion of development..... The process often involves engaging participants in setting and achieving their own goals. Program designers observe where they end up and make judgments about the implications of what has happened for future programming and re-engineering. They never expect to arrive at a steady state of programming because they're constantly tinkering as participants, conditions, learning, and context change. (p. 136)

Developmental evaluations may provide useful information for decision makers. However, in practice, evaluators are rarely present in the planning phases of programs, and the evaluative function has typically been perceived as formative and/or summative rather than developmental. Indeed, many would argue that what Patton calls developmental evaluation is simply a re-naming of program planning and not a form of evaluation at all (Edelenbos & van Buren, 2005; Iverson, 2003). Moreover, in the day-to-day work of evaluators, the lines between summative and formative evaluations are often blurred. Evaluators are frequently asked to provide information which can be used for both learning and accountability purposes. Thus, for example the OECD-DAC (1992) defines aid evaluation as:

...an assessment, as systematic and objective as possible, of an on-going or completed project, program, or policy, its design, implementation and results. The aim is to determine the relevance and fulfillment of objectives, developmental efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible

and useful, enabling the incorporation of lessons learned into the decision-making processes of both recipients and donors. (p. 132)

The literature on training evaluation, for its part, tends to include both learning and accountability goals. One of the most widely-cited guides to training evaluation (Philips and Stone, 2002: 56) identifies ten objectives of training evaluation: (a) Determining success in accomplishing training objectives; (b) Identifying strengths and weaknesses in the training process; (c) Comparing the costs and benefits of a training program; (d) Deciding who should participate in future programs; (e) Testing the clarity and validity of tests, cases and exercises; (f) Identifying which participants have had the most success with training; (g) Reinforcing major points made to the participant; (h) Gathering data to assist in marketing future programs; (i) Determining if training was the appropriate solution for the specific need; (j) Establishing a database that can assist management in making decisions.

While evaluators are frequently charged with gathering information for both learning and accountability purposes, there is widespread recognition that summative and formative evaluations generally are based on very different types of data, which often need to be collected in very different ways (Patton, 1996:66). Moreover, there is often an inherent trade-off between learning and accountability aims. Devoting resources and attention to proving program impact will limit the amount of resources that can be devoted to gathering information that can help improve program implementation. Moreover, program implementers and beneficiaries are likely to react differently to evaluators, depending on their primary evaluative goals. It may be more difficult for evaluators charged with the task of “allocating scores” to hold an open dialogue with

program implementers about potential program improvements than it would be for evaluators whose sole function is to facilitate program learning.

The trade-off between summative and formative evaluative data collection is widely recognized in the literature (Brinkerhoff, 1987; Fitz-Enz, 1994; Smutylo, 2001). Some experts prioritize the collection of summative evaluative information that can accurately determine program impact. Others argue that proving impact diverts resources from what they see as the more useful formative function of improving programs through understanding and assessing which elements of a program appear to be working and which do not. Of course, even formative evaluations must be based on some sense of the program's effects, but the levels of proof and specificity of program outcome data needed for formative evaluation are generally lower. Thus, while some evaluation experts call for use of more rigorous quasi-experimental and experimental studies to accurately determine program impact (Bamberger, 2000; Savedoff, et al, 2006), others claim that such evaluations divert resources from more useful questions. Specifically, the argument has been made that (a) proving program impact is often prohibitively costly or impossible and that (b) formative evaluative information aimed at learning how to improve programs is generally far more useful to decision makers.

Thus, for example, John Mayne (1999:5) argues that "Measurement in the public sector is less about precision and more about increasing understanding and knowledge. It is about increasing what we know about what works in an area and thereby reducing uncertainty... We need to include softer and qualitative measurement tools in our concept of measurement in the public sector." Similarly, Patton, in his book on utilization-focused evaluation (1986:128), asserts that, where evaluation resources are limited,

information on program implementation is generally more useful to decision makers than information on program results.

The same perspective has been voiced by evaluators in both the international development and training fields. Terry Smutylo (2001) of Canada's International Development Research Centre writes about the dangers of attempting impact evaluations for most international development programs. In fact, he argues that donor demands for more impact evaluations pose "threats to learning from evaluations of development efforts." Instead, he asserts that limited aid evaluation dollars should be focused on learning how to increase effectiveness of programs in relation to ultimate goals, by studying program processes and evaluating the direct effects of a program on behaviors, actions and relationships rather than by evaluating the ultimate contribution of programs to the achievement of development goals.

This sentiment is echoed by several training evaluation experts. For example, Robert Brinkerhoff claims in his book "Achieving Results from Training" (1987) that, "...the primary payoff from evaluating HRD (human resource development) is the improvement of programs. Although accountability data are needed, "proving" the value of HRD is not the main reason to evaluate it. When evaluation data are systematically collected and used to make programs work better, the "proof" argument tends to take care of itself." And Fitz-Enz (1994) suggests that,

One of the self-imposed barriers to seeking the concrete value of training programs is the mistaken notion that the trainer must present proof, with a capital "P", that training had a special effect on the organization's bottom line. So many factors affect the human and financial performance of a

company, it is said, that isolating the pure contribution of training is impossible....The objective of a valid and reliable evaluation effort that assigns a specific value to the outcome of a training program is simply this: to demonstrate that there is a probable correlation between the training event and a subsequent change in quality, productivity, sales or service. (p.56)

By what criteria should evaluation be assessed?

One of the most commonly referenced criteria for judging the value and quality of evaluations are those of the Joint Committee on Standards for Educational Program Evaluation (1994). These standards include the following four criteria:

- Utility: providing useful information to decision-makers.
- Feasibility: realistic, prudent, diplomatic and frugal.
- Propriety: in the sense that they are legal, ethical and show due regard to those involved in the evaluation and those affected by results, and;
- Accuracy: conveying “technically adequate information about the features that determine worth or merit of the program being evaluated.”

One of the notable features of these standards is that they take a broad view of the elements of evaluation quality. In the words of Daniel Stufflebeam, one of the principal authors of the standards, the standards are designed to go beyond the “narrow view that any good evaluation should meet only the experimental design requirements of internal and external validity.” (Stufflebeam, 2004: 100). The Joint Committee supplements the accuracy category with three other categories of standards, recognizing that there may be possible tradeoffs between these four groups of criteria, and leaving it to the evaluator to

judge how to reconcile potential conflicts rather than suggesting that one category takes primacy over the others.

The Joint Committee standards have been chosen as the most relevant set of evaluation criteria for the purposes of this dissertation. While there are other commonly-used evaluation standards, these generally refer to the ethical responsibilities of the evaluator, rather than the quality of the evaluation, and thus are less suitable for this analysis. Thus, for example, the American Evaluation Association's guiding principles include that evaluators display honesty, integrity and competence and respect the security, dignity and self-worth of evaluation stakeholders (2004). While it is of course important that evaluation methods not contradict these ethical guiding principles, there is no indication that there are significant ethical shortcomings in any of the presently-used or proposed models of training evaluation. Consequently, the Joint Committee standards are far more relevant and instructive with regard to judging the quality of models of training evaluation.

This dissertation's analysis of the Kirkpatrick model of training evaluation, as well as of the new personal benchmark system developed in the context of this research, was guided by the Joint Committee's evaluation standards. Analysis focused on the standards of usefulness, feasibility and accuracy as there is no suggestion in the literature that basic standards of propriety are commonly not met in training evaluation. The following is a more detailed description of each of the relevant criteria as they relate to this dissertation's analysis of training evaluation:

Feasibility. The Joint Committee's feasibility standards call for evaluation systems that are as easy to implement as possible, efficient in their use of time and

resources, adequately funded, and viable from a number of other standpoints. There are three standards under this attribute including practical procedures, political viability, and fiscal viability. Feasibility is judged from the perspective of those administering the evaluation. In the case of international training courses – evaluation is almost always conducted by training institute representatives – either course organizers themselves, the evaluation department of the training institute, or external consultants hired by the training institutes to conduct the evaluation.

Utility

Utility, as a subjective concept, depends on which decision makers are using the information generated through evaluation. In the case of training evaluation, this can include training institute personnel, the workplace organizations of participants that chose to send their employees to training courses, or the donors who financed training. This dissertation primarily looks at utility of evaluative information for training course organizers and other training institute officials. This is due to the fact that these agents are likely to be the primary users of evaluative information.

Accuracy

The Joint Committee definition of accuracy does not conform with the Oxford English Dictionary definition of “correct in all its forms” - a definition which suggests conformity with an objective factual reality. Rather, the Joint Committee standards inject a certain degree of relativity to the construct by referring to “technically adequate information” and a degree of utility by referring to the purpose of that information, namely, to “determine the worth or merit of the program.” Thus, this definition views accuracy through the utilitarian prism of whether the information obtained is accurate

enough to determine the worth or merit of the program without requiring that data be held to the most rigorous standards of empirically-proven truth.

The dissertation embraced this more utilitarian concept of accuracy. In particular, in its analysis of the present use of the Kirkpatrick framework, it focused on the issue of construct validity – the extent to which constructs actually measure the things that they are intended to measure. For example, in determining whether training participants have learned from the course, survey questions soliciting participant self-reports on the extent that they have learned will only be seen as accurate and credible to the extent that there is evidence that self-reports of learning reflect actual learning.

Training Evaluation - The Kirkpatrick Model

The most frequently employed model for training evaluation was developed almost 50 years ago by Donald Kirkpatrick. The Kirkpatrick model (1968) traces a results chain that begins with participant satisfaction and learning outputs, leading to workplace behavior change outcomes and then to impact on target organizations. To assess progress along this chain, the model defines four levels of training evaluation:

1. Reaction: Satisfaction of participants as measured by structured questionnaires at the end of the course or participant interviews.
2. Learning: Learning gains as measured by tests at end of course (or pre- and post- tests) or assessment of in-class projects.
3. Behavior: Changes in workplace behavior, as measured by interviews with /surveys of participants, supervisors, colleagues, subordinates and/or participant observation.

4. Results: Impact of training on organizational performance and achievement of organizational objectives.

While variations on this model have been proposed by other writers over the past half century, as is detailed in the next sections, the Kirkpatrick model has remained the point of reference and dominant paradigm for virtually all training evaluation (Brown & Reed, 2002:2; Holton, 1996:5; Newstrom, 1978; Wang & Spitzer, 2005). In fact, a recent review of new approaches to evaluation of learning programs commissioned by the World Bank Institute found that of the over 20 experts interviewed for the research, none used or even knew of any approach other than Kirkpatrick's (Evans, 2007:4). However, while use of Kirkpatrick's "four level" terminology is pervasive amongst training evaluators, there are those who argue that Kirkpatrick's framework is not an evaluation model at all, but rather a taxonomy of outcomes (Holton, 1996:5). This claim is based on the fact that, Kirkpatrick defines what he sees as the four levels of training results without being prescriptive about what methods of evaluation should be used at each level, or asserting that there is any causal link between the levels. Indeed, Kirkpatrick's theory of training evaluation assumes that it is possible to look at various levels of training results in isolation from other levels.

Kirkpatrick plus models.

While the Kirkpatrick model remains the standard for virtually all training evaluation conducted today, a number of "Kirkpatrick-plus" models have suggested building on the Kirkpatrick model. One category of such models is the "fifth levelists", those who propose adding a fifth level to Kirkpatrick's four-level framework. For example, Phillips & Stone (2002) add a fifth level of return on investment. Hamblin

(1974), for his part, suggests a different fifth level of evaluation, that of ultimate value. Hamblin argues that it is often useful to distinguish between measuring change at the organizational level and measuring how that change impacts the ultimate criteria by which an organization judges its success, whether these criteria involve monetary return on investment or other cost-benefit considerations. Kaufman & Keller (1994) also suggest a fifth level of evaluation. In their version, the fifth level should be societal value of training.

Other evaluation specialists suggest building on Kirkpatrick's levels not by adding to them but by better elaborating what is to be measured at each level, and by providing techniques for how this could be done. For example, Warr & Bunce (1995) suggest that level one questionnaires should be broken down into three distinct categories: reported enjoyment of training, its perceived usefulness, and its perceived difficulty. Lee & Pershing (2002) elaborate at length possible standards for level one questionnaire design. Mathieu et al. (1992) propose an interim category of attitude change between the categories of reactions and learning. Kraiger, Ford & Salas (1993), for their part, use cognitive theory to suggest a classification scheme dividing Kirkpatrick's level two (learning) into cognitive (verbal knowledge, knowledge organization, cognitive strategies), skill-based, and affective (attitudinal and motivational) learning outputs. Hamblin (1974) and Faerman & Ban (1993) provide a variety of different measurement techniques in order to aid evaluators in measuring Kirkpatrick's four levels.

All of these above authors share a basic acceptance of the Kirkpatrick framework. While some add a fifth level and others better define the composite elements of the levels

or provide new tools for measuring them, all of them begin with the premise that training evaluation should be primarily a summative exercise, measuring training achievements along a results chain which starts with trainee reactions, and then moves to learning outputs, behavior outcomes, organizational impacts and, in some cases, beyond. In other words, they all share the basic assumption of the Kirkpatrick framework that the primary purpose of training evaluation is the assessment of results, rather than generating information on program design and implementation which can be used to improve program performance. The Kirkpatrick approach, however, has been widely criticized by other theorists, precisely for this assumption. Kirkpatrick's focus on results measurement is criticized on the basis of its utility, feasibility, and accuracy, to use the terminology of the Joint Committee standards outlined at the beginning of this chapter. The following section documents the literature criticizing the Kirkpatrick framework.

The Kirkpatrick framework's feasibility.

A feasible evaluation, according to the Joint Committee on Standards for Educational Evaluation (1994), is one that has practical procedures, political viability, efficient and effective resource use and effective project management strategies. Where evaluation strategies are lacking in feasibility, evaluators will have difficulty correctly and effectively implementing them. The evidence on use of the different levels of the Kirkpatrick framework suggests that the framework may not be feasible in many training contexts. As is reviewed in the following paragraphs, there is extensive literature documenting the lack of training evaluation beyond level one, in both corporate and international development contexts. While level one evaluations are almost always administered at the end of training courses, attempts to evaluate how training has affected

the work behavior (level three: training outputs) of training participants or the achievement of broader organizational goals of their workplaces (level four: training impact) are seldom made. This section presents the evidence on the lack of full use of the Kirkpatrick framework, and explores some possible reasons why pursuing higher levels of training evaluation might be largely infeasible in many training situations.

Kirkpatrick, and later authors who have expanded upon his framework, all insist that good training evaluation must involve at least some evaluation of the effect of training on workplace behavior and organizational performance. However, in reality, research indicates that only a small proportion of training organizations systematically go beyond level one in training evaluation. Surveys of American corporate training programs have found only limited use of any evaluative tool beyond participant reaction questionnaires (Brandenburg 1982; Parker 1986; Saari et al. 1988; Training 1996). The lack of higher levels of training evaluation has been recognized as a problem for decades. Nearly 40 years ago, a 1970 review of management training evaluation studies published over the previous two decades found that less than a third evaluated the effects of training on individual job performance or on organizational results (Clement, 1981). The low incidence of level three and four evaluation has remained consistent or, possibly, deteriorated, since this study was completed. Robinson & Robinson (1989), in a survey of 150 training directors, found that while 97% administered level one evaluations at least occasionally (77% did such evaluations 81-100% of the time) and 90% conducted level two learning evaluations at least occasionally, 31% never did level three evaluations and 60% never conducted level four. A 1999 American Society for Training and Development report found that of the courses reviewed by the ASTD's benchmarking

forum, 89% were evaluated at level one, 29% at level two, 11% at level three and only 2% at level four (cited in Pershing & Pershing, 2001). In sum, to cite one review of training evaluation models: “Training evaluation is a bit like eating five portions of fruit and vegetables a day; everyone knows that they are supposed to do it, everyone says they are planning to do better in the future and few people admit to have gotten it right” (Tamkin et al, 2002:ix).

While there is no survey information or other forms of data available on the reasons why higher levels of evaluation are rarely undertaken in corporate training programs, a number of possibilities have been suggested by training and training evaluation experts. Morgan and Casper (2000) argue that training organizations restrict themselves to level one evaluations because higher levels are too costly to perform. Others suggest that an “accountability vacuum” exists that limits the impetus for, and thus the political viability of, training evaluation. For example, Robinson & Robinson (1989) argue that human resource departments in companies are generally only held accountable for the design and delivery of training programs, not for their effects on workplace organizations. Thus, human resource departments are only required to report on participant training days, participant satisfaction and/or learning. Robinson & Robinson argue that disconnect between line managers and human resource development managers results in neither being held accountable for demonstrating the effect of training on the workplace. Brinkerhoff similarly asserts (2003) that:

Anyone involved professionally in training in today’s organization cannot help but observe that the typical training practitioner is so busy organizing and delivering training that there is very little effort paid to....follow-up

assessment of training results. The concern of the training staff, and often the formal budget reports of the training unit, is with measure of activity: training hours provided, training seats filled, numbers of programs delivered and so forth. Almost never are long-range results of training assessed, nor are training staff terribly concerned with them (p.42).

In other words, training departments are generally given neither the political-managerial support, nor the time or budgets to evaluate the effect of training on participant workplace behavior or on organizations. This suggests that, in the absence of strong managerial commitment to more extensive evaluation, post-course evaluation may neither be politically viable nor procedurally practical.

This “accountability vacuum” may be exacerbated where training is being done externally, through dispatch of trainees to outside courses, rather than through in-house training. Saari et al (1988) demonstrated that companies are far more likely to assess the effects of training on workplace performance and organizational outcomes when courses are held internally within the organization. They found in a review of 1,000 randomly selected mid-size and large US companies with formal management training programs, that, while a full 98% of companies with company-specific management programs conducted at least some evaluation of their in-house courses, 32% of companies who sent their employees to short external courses did nothing to evaluate their effects and 42% of companies who sent their employees to executive MBA programs did not evaluate their results, despite the fact that an average of \$14,000 per person was spent on such training.

This absence of post-course training evaluation appears to be present in international development contexts as well. A recent review of training evaluation

theory and practice commissioned by the World Bank Institute (WBI) found that training organizations in the private sector conduct post-course training evaluation more frequently than the WBI (Evans, 2007). A benchmarking exercise conducted by the author of this dissertation for the World Bank Independent Evaluation Group found that of five reviewed training organizations working in developing countries [JICA (Japan), MASHAV (Israel), WBI (World Bank Institute), ITCILO the International Training Center of the International Labor Organization, and INWENT (Germany)] only one, INWENT, systematically conducted evaluations of its programs beyond level one (IEG, 2008).

There are a number of reasons why evaluating training's effect on workplace behavior and organizational performance is likely to be even less feasible in international development contexts than in developed country, corporate contexts. First, much of the training done in the former contexts is done through external, rather than in-house courses, which, as Saari et al.'s research suggests, may exacerbate the "accountability vacuum". Second, evaluation of training workplace behavior outcomes and organizational impact in international development contexts can be more expensive to execute than evaluations done for in-house corporate training programs. Unlike corporate training programs, training for international development generally involves separate entities, in different countries, financing, providing and benefiting from training. The training financier-provider-beneficiary divide is likely to necessitate greater effort and expense to obtain information on how training has affected participants' on-the-job performance and their organizations' efficacy, because it is often the donor who is commissioning and/or conducting the evaluation from the "outside" rather than the

beneficiary observing from the “inside” how training has affected organizational performance (Iverson 2003, 15). This can complicate access to information, sometimes necessitating dispatch of foreign consultants to one or more countries to gather information, thus considerably raising the cost of evaluations. Thus, for example, a recent evaluation of the post-course workplace performance outcomes and organizational impact of one international training program conducted by a UN training organization necessitated the dispatch of an external consultant to several beneficiary countries, costing approximately \$100,000 to conduct (Taken from an interview with Franz Lenglet, Deputy Director, ITCILO, September 2006).

While these feasibility factors suggest the likely low incidence of level three and four evaluations for training done in international development contexts, no studies prior to this dissertation have actually been undertaken thus far documenting the collection, or lack thereof, of higher level training evaluation data by development training institutes (DTIs). Nor have there been any attempts to systematically document the reasons why information on training effects on workplace behavior or organizational performance is so rarely collected.

The Kirkpatrick framework’s utility.

The utility of the Kirkpatrick framework has been criticized by practitioners and theorists for its narrow focus on evaluation of results. Some argue that, for training evaluation to serve both accountability and learning purposes, it is not enough to evaluate training results, as the Kirkpatrick model prescribes, but, rather one must also determine the reasons for training success or failure. In addition, some theorists and practitioners see a strong need for training evaluation that is not only formative and summative in

nature, but also developmental, assisting training practitioners in design of their courses rather than just assessing implementation or results.

Much of the criticism of the utility of Kirkpatrick's framework stems from the fact that the effects of training on workplace performance are largely dependent not on the quality of the training program itself, but on intervening variables external to the training course such as organizational support and resources for implementing learning (Brinkerhoff & Apking, 2006; Holton, 1996; IEG, 2008). For this reason, Brown & Gerhardt (2002), Brown & Reed (2002), Holton (1996) and Brinkerhoff (1988, 2006) have all suggested that Kirkpatrick's framework does not provide sufficiently useful information for decision makers interested in improving the results of training. In Holton's (1996) words,

The only conclusion possible using data within the four-level model is that something is wrong with the training program. However, if the many intervening variables that remain unmeasured are considered, it is quite possible that the training program is well designed and that the problem lies outside the classroom with some element of the organization, job or individual. Thus, to the extent that evaluations should provide information to make correct decisions about HRD [Human Resource Development] interventions, the four-level taxonomy fails to do so when the outcomes are not those desired. (p. 8)

Holton (1996) emphasizes the importance of building into evaluative schemes information on how trainee characteristics influence success. Holton argues that the causal linkages between the four levels in Kirkpatrick's model are unclear, and that

further research is needed to develop a fully-specified model of successful training, detailing how intervening variables such as motivation to learn, trainability, job attitudes, personal characteristics and workplace conditions affect training success. Holton calls for the development of a new model for training evaluation that would take into account these and other factors likely to affect training success.

Hamblin (1974) and Brinkerhoff (2006), for their part, have placed particular emphasis on the importance of building into training evaluation an understanding of the organizational factors affecting training success. They have both argued that useful training evaluation should reinforce the partnership between management, trainers and trainees and provide information not only on training but on the 'performance system' factors that facilitate or impede training impact. Brinkerhoff criticizes the Kirkpatrick framework for looking at training in isolation from the organizational context in which it is done. In doing so, Brinkerhoff (2006) argues that Kirkpatrick designates trainers as the primary client of the evaluative exercise, whereas, in his view, it is the line manager who should be seen as the primary consumer of training evaluation. In other words, while trainers may only be concerned with what goes on in their classrooms, line managers need to understand how training inputs interact with individual, organizational and institutional inputs and conditions in order to help achieve organizational goals. To support these managers, he argues that training evaluation should be recast as evaluation of the entire managerial and performance system needed to achieve training results, including not only training, but also organizational factors such as managerial support and availability of resources for implementation of learning.

If we want to evaluate "transfer" or behavioral change from training, we must evaluate the managerial and performance system, not training. The primary feedback channel should be to the owners of the performance system: line management and senior leadership. If we are really serious about improving the performance that training can contribute to, then we need to be equally serious about getting all the players in the performance partnership involved in understanding their roles, their contributions, their successes, and their failures. (p. 43)

Others criticize the utility of Kirkpatrick by arguing that training evaluation should most usefully begin in the developmental phase, in order to ensure better use of training resources. Given the strong influence of factors external to training on the likelihood of training success, some authors have argued that there is a need for evaluative frameworks which are not only formative and summative in nature, but also developmental, helping to better align training content with organizational realities (Brinkerhoff, 2006; Brown & Gerhardt, 2002). According to this perspective, evaluative techniques should be used in the planning phases of a training program in order to determine the likelihood of training success in a given organizational context and to ensure that participants are properly selected and training is well-targeted to organizational needs. Thus, for example, a developmental evaluation of training might aid in ensuring that training programs are matched to organizational needs, that training content is appropriate for the participants' level of expertise and that the necessary resources and managerial support exist in the workplace for implementing new skills acquired through training.

The Kirkpatrick framework's accuracy.

The final grounds upon which the Kirkpatrick framework has been criticized can be placed under the category of accuracy. Specifically, what is often criticized is the construct validity of information obtained at the different levels, that is, whether evaluations actually measure the things they are used to indicate. These critiques are not so much criticisms of the Kirkpatrick framework in theory as they are criticisms of how the framework is commonly used and what information each level of evaluation is thought to produce. Due to the expense, effort and time necessary to conduct higher level evaluations, training evaluators tend to use level one and two evaluations as proxy indicators for information on higher levels of training outcomes (level three) and impact (level four). They generally do so by asking participants to assess, as part of level one questionnaires, how useful and relevant training is to them. However, while level one and two evaluations may provide accurate information on participant satisfaction and learning levels, research indicates that they are not accurate predictors of whether participants can and will use what they've learned on the job.

Accuracy of level 1 evaluations.

Level one evaluations provide information on participants' degree of satisfaction with training. They may ask participants for their perceptions of how much they learned or the relevance, interest and usefulness of training content, the effectiveness of the instructor or the quality of training materials (Keller, 1996:7). The results of end-of course participant satisfaction questionnaires are often cited by training institutes as evidence of course outcomes. However, while such questionnaires may provide accurate information on participant satisfaction, extensive research has documented that

participant satisfaction, or even participant perceptions of learning and utility, does not correlate with either learning or, more importantly, with the likelihood that trainees will be able to successfully apply what they have learned on the job (Lee & Pershing, 2002; Bramley & Newby, 1984; Warr et al., 1999). Thus, while level one evaluations may be useful to monitor and improve variables such as training content and instructor performance, they cannot be seen as indicators of the likelihood of successful learning outcomes, behavior outcomes or organizational impact.

Rodin & Rodin (1972) found a negative correlation between participants' self-assessments as recorded in level one evaluations and actual learning. Dixon (1990), as well as Warr & Bunce (1995), found no correlation between participant perceptions of whether they had learned or assessments of course enjoyment, usefulness, and difficulty and participant learning. Faerman & Ban (1993) found only a moderate relationship between participant ratings in level one evaluations of course utility and subsequent job behavior, while Noe & Schmitt (1986) found no link between the two. Alliger et al.'s (1997) meta-analysis of training evaluation studies conducted between 1959 and 1997 also concluded that there was no correlation between participant satisfaction and learning.

Notwithstanding the above findings, some research has suggested that differentiated level one evaluations may provide more accurate information than general questions on course satisfaction about the likelihood of learning being used in the workplace. Thus, Alliger et al. (1997) found that, when participant satisfaction questionnaires asked both whether participants had enjoyed training (affective reactions) and whether training was useful for their jobs (utility reactions), enjoyment was uncorrelated with on-the-job behavior change, but there was a modest correlation

between utility reactions and behavior change. Alternately, end-of-course questionnaires can be used to go beyond participant's attitudes to gather information on participants' workplace organizational contexts. For example, Motorola University reports that the following workplace-related questions on their level one questionnaires were stronger predictors of the on-the-job impact of training than evaluation questions on course content, quality and self-assessment of learning (IEG, 2008:36):

1. My managers know what I was taught in training.
2. The training was built to match the way I need to do my job.
3. Job aids are available to support what I learned in this training.
4. The procedures taught in this training are ones I can use on the job.
5. I feel my co-workers will help me if I have problems using this training on the job.
6. In general, applying training on the job helps employees in their careers with Motorola.

In sum, while level one questionnaires may give accurate information on participant enjoyment and overall satisfaction, on quality of training content, lecturers and materials and on workplace contexts, they are less accurate as indicators of learning or likely utility of learning in the workplace. There is a problem of construct validity with level one questionnaires which ask for participants to report on whether they learned from the course and whether they feel the course will be useful to them in the workplace. Participant self-reports are not likely to give accurate pictures of the true extent of learning or the likelihood that training content will actually be applied at work.

Accuracy of level two evaluations.

Possibilities for accurately measuring learning vary somewhat, depending on factors such as the type of learning being measured. It is generally easier to measure acquisition of hard skills or knowledge gains than soft skills or attitudes. However, even where it is possible to provide good measurements of learning, this cannot be seen as a proxy indicator for increased “capacity to” rather than just “knowledge of”. While one might expect that evaluations of participant learning (level two) would be good predictors of whether learning is used on the job, research indicates that this is not the case. Even where learning has occurred, trainees may have difficulty understanding how to apply what they’ve learned in a workplace context (Erikson 1990; Ghodisian et al 1997). Clement (1982), in a study of the training of 50 new supervisors, found no relationship between the magnitude of learning gains and that of work behavior change. More important predictors of workplace behavior change, he found, were similarity between workplace environments and training content, opportunity to practice learned skills, and the expectations of peers and immediate supervisors. Alliger et. al (1997) found that learning evaluations which measured behavior or skills application change in-course only weakly predicted use of those behavior and skills once participants had returned to the workplace. They concluded that such behavioral learning evaluations were even weaker predictors of workplace performance change than participant perceptions of learning usefulness as recorded in level one questionnaires.

Level three accuracy.

There are many possible ways of accurately measuring the effect of training on trainees’ workplace performance. These may include observation of participants in the workplace, measuring performance improvements against pre-course baseline data, 360

degree evaluation surveys of trainee performance, in-depth interviews with supervisors, work colleagues, clients and/or policy makers, among many others. However, as illustrated in the above section on feasibility of evaluations, these methods are generally cost-prohibitive in the sort of short-term, multi-country or multi-organization DTI courses which this dissertation addresses.

The least costly, and most common, method by which training institutes measure post-course workplace behavior change or performance is through participant questionnaires administered three months to two years after course completion (Tamkin, 2002, Tascherau, 1998). However, there are serious concerns with the accuracy of participant self-reports of training results. Participants may not be the best qualified to assess their own performance. One study of 360 degree performance evaluations which compared self, superior, subordinate and peer assessments of performance to assessment center performance test results found that self-reports were the least accurate of all assessments as compared to performance test results. In fact, the study described self-reports as “highly inaccurate”, with the least competent employees likely to give themselves the highest ratings (Atkins & Wood 2002). Similarly, Carless & Robert-Thompson (2001), in a study of 545 participants in a Royal Australian Airforce leadership training course, found that there was a low level of agreement between self- and trainer assessments of leadership qualities, with the worst performers most greatly overestimating their skills.

The inaccuracy of participant self-reports of use of training may be exacerbated in international development contexts where participants receive training as a free good, sometimes with perks such as per diems or trips abroad. Beneficiaries may see

themselves as having an interest in stating that training was successful in order to ensure that they will continue to receive donor funds. There may also be a lack of confluence between beneficiary goals and funder goals, making even the definition of what would constitute a positive training outcome a matter of debate. Moreover, cultural differences may also complicate communication and data analysis (Iverson, 2003: 15).

Level four accuracy.

Obtaining accurate information on organizational impact (level four), has long been recognized as a difficult, and in some cases, impossible, endeavor (Dionne, 1996:280; Fitz-enz,1994:56; Pulley, 1994:20; Warr et. al., 1999:355). Three reasons are primarily cited for the difficulty. First, not all training goals may have easily monitorable indicators. This is particularly true in international development contexts, where many goals, such as good governance, human rights or community empowerment, are far more difficult to measure than private-sector variables such as profits or client satisfaction (Iverson, 2003:14; Smytulo, 2001; Whyte, 2004:19). Second, as non-training related factors, both endogenous and exogenous to the organization, can have a strong effect on organizational performance, it is often difficult to isolate the effects of training from that of other changes or interventions (Kaufman & Keller, 1994: 372). Attributing impact may be particularly difficult for many international development programs, where success can often only result from complex programs involving multiple, interrelated interventions. Thus, identifying the contribution of training to changes in organizational performance indicators may be particularly difficult (Smutylo, 2001). Finally, even where it is possible to disaggregate the impact of training on organizational performance from other interventions, these effects may only be apparent in the long-term – far too

late to be picked up in post-program evaluations (Brinkerhoff, 2003). For these reasons, several authors have suggested abandoning altogether attempts at level four assessments of program impact for capacity building interventions in international development contexts and have recommended instead focusing on better level three evaluations (IEG, 2008; Smutylo, 2001).

Re-examining the concept of accuracy.

The above analysis suggests that it is extremely difficult, given the resources commonly available to training institutes, to feasibly gather accurate information on training results at the level of participant behavior change outcomes and organizational impact. This is partly because level one and two evaluations can in no way be used as proxy indicators for higher levels of evaluation and level three evaluations may be difficult to administer. Moreover, participants may not themselves have an accurate understanding of how training has affected them and their organizations. This presents a certain "catch 22" of training results evaluation. On the one hand, participant self reports at any level cannot be viewed as an accurate gauge of participant workplace behavior outcomes or organizational impact. On the other hand, the expense of forms of post-course data collection other than self reports in written questionnaires may be prohibitively costly for most training institutes. However, this does not necessarily imply that the solution is to develop better, less costly tools to improve accuracy of information gathered on the level of participant behavior outcomes and organizational impact. Instead, the work of some theorists would suggest that we should abandon the criteria of accuracy for participant self-reports of training results altogether.

Fowler would argue that accuracy is not the correct term at all for the sort of data gathered through participant self-reports (1995: 151). According to Fowler, positivist concepts of accuracy are limited in their applicability to measurement of subjective attitudes. In other words, when asking a respondent for his or her opinion on a matter, it is impossible to speak of one opinion as being more “accurate” than another. Instead, the relevant question is merely whether the survey has captured the nature of the respondent’s opinion in a sufficiently nuanced manner. Fowler suggests that we should ask not whether respondents’ opinions have been captured accurately, but whether these opinions as recorded reflect sufficient variation to be useful to the surveyor. For example, if 99% of survey respondents choose the same answer option on a close-ended question on their opinions, the data generated by the question is not, according to Fowler, a particularly useful one to the surveyor. Rather, a good survey question is one which better captures the variations of opinion, however subtle, that are likely to exist between respondents.

Fowler’s approach to survey questions soliciting respondents’ attitudes is reminiscent of Lincoln & Guba’s approach to the concept of validity (1999: 402). Lincoln and Guba have developed a full range of criteria for assessing research quality from a non-positivist perspective. Lincoln & Guba do not see the role of a social science researcher empirically proving objective fact (whose existence they dispute), but rather, understanding and representing the multiple representations of reality that subjects of research may have. For this reason, they reject positivist conceptions of construct validity or accuracy, and instead propose the quality criterion of credibility. Credibility, as defined by Lincoln & Guba, is gauged by how closely information gathered reflects

the actual perceptions of those being studied. For participant self-reports to be established as credible, they must merely reflect participants' perceptions of reality rather than objective truth.

Thus, both Fowler and Lincoln and Guba reject the need for approaches which solve the accuracy problem of training evaluation by finding feasible methods for gathering more accurate information at Kirkpatrick's level three (outcomes) and level four (impact). Instead, their approaches are more compatible with training evaluation theorists who argue for the greater utility of data aimed at improving training results rather than proving them (Brinkerhoff, 1987; Fitz-Enz, 1994; Smutylo, 2001). All these approaches suggest that the relevant goal for improving training evaluation is not enhancing accuracy, but rather enhancing the nuance and detail of our understanding of participant's own perceptions of the contribution of training to them and to their organizations.

Towards a New Model for Evaluating Training

The shortcomings of the Kirkpatrick framework and of its present use, suggest a need for better tools for assessing the contribution of training to capacity building in international development contexts. This is not to say that the Kirkpatrick model can never be used to effectively evaluate the learning outputs, workplace performance outcomes and even, in some cases, the organizational/institutional impact of training. However, the challenges discussed in the above section point to the difficulties of using the Kirkpatrick framework as presently used to generate information which may be used to enhance training programs or understand their effects.

The following section presents the literature on alternative approaches to training evaluation and evaluation of other capacity building interventions in international development contexts. Some of the approaches presented below are predicated on the argument advanced by several of the academics and practitioners cited earlier in this chapter that, for evaluation to generate adequate information for learning and accountability purposes, it must go beyond summative determination of training results. Others explore methods for results evaluation which, while not incompatible with Kirkpatrick's framework, go beyond the sort of measurement tools normally associated with training evaluation. All share a conception of evaluation as something which goes beyond the Kirkpatrick's summative, ex-post-facto approach. They all suggest that evaluation should be part of a broader performance system aimed not only at assessing program effects but also at enhancing program planning, improving program performance and/or empowering program participants. In brief, these approaches fall into four categories:

1. Examining training process and training participants' workplace context in order to better understand and improve training results.
2. Adding a developmental component to evaluation – i.e., using evaluative techniques to help better plan training and control quality up-front.
3. Enhancing the participatory nature of evaluations both to improve the credibility of findings and to enhance the efficacy of training.
4. Adopting a “Goal-Free” approach to results evaluation.

This section explores these four categories of evaluative tools, discussing their use in training evaluation as well as in other contexts. In addition, the chapter presents two

fairly recent examples of evaluations/evaluative techniques in international development contexts – that of the IDRC’s Outcome Mapping, the World Bank’s Independent Evaluation Group’s three-tier training evaluation. These models are presented as potentially instructive ones for the development of a new training evaluation model in the context of this dissertation.

Process and context evaluation of training.

As documented earlier in this chapter, one of the primary critiques of Kirkpatrick is that evaluation of training only through measurement of results does not provide sufficiently useful information for program improvement. For this reason, several researchers and training evaluation experts have recommended evaluation of training participants’ workplace contexts and of the training management process in order to be able to better identify reasons for training success and failure.

Process evaluation is defined by Michael Quinn Patton (1986) in the following manner:

[Process evaluation]...focuses on the internal dynamics and actual operations of a program in an attempt to understand its strength and weaknesses. Process evaluations focus on why certain things are happening, how the parts of the program fit together, and how people perceive the program. This approach takes its name from an emphasis on looking at how a product or outcome is produced rather than at the product itself. (p.139)

Context evaluation of training, for its part, looks at the organizational context in which trainees will be applying learning in order to determine factors which may inhibit or promote training transfer (Stufflebeam, 2007).

Process and context evaluation of training are particularly useful when the goal of training goes beyond pure learning to that of “transfer” of learning - defined by Baldwin & Ford (1988:63) as, “...the degree to which trainees effectively apply the knowledge, skills and attitudes gained in a training context to the job.” In other words, where training endeavors to result in workplace performance change rather than just passive learning, it is useful to gather information regarding variables beyond learning achieved. As transfer may be influenced by many factors outside of the classroom itself, context and process evaluations enable evaluators to look at a range of variables that may have influenced training results. A copious amount of research exists analyzing the factors which contribute to transfer of training. Among the frequently cited factors:

1. Trainee characteristics such as age, experience, intelligence, motivation (Brown and Reed 2002; Fecteau et al, 1995; Ford et al, 1992, Foxon, 1993, Mathieu et al, 1992; Naquin & Holton, 2002; Noe, 1986; Richey, 1992; Warr & Bunce, 1995).
2. Training design and relevance of training to organizational goals (Baldwin and Ford,1988; Brinkerhoff & Montesino, 1995; Carter, 2002; Foxon, 1993, Salas & Cannon-Bowers, 2001).
3. Workplace incentives for implementation of training (Baldwin & Ford,1988; Brinkerhoff ,1989; Fecteau et al 1995).

4. Organizational structure and culture (Brinkerhoff, 1989; Huczynski & Lewis, 1980; Kupritz, 2002, Richey, 1992; Rouiller and Goldstein, 1993; Tannenbaum,1997; Tracey et al 1995, Warr et al, 1970).
5. Support of colleagues and managers for implementation of learning (Brinkerhoff & Montesino, 1995; Facticeau et al, 1995; Foxon, 1993; Georgenson, 1982; Huczynski & Lewis, 1980; Richey, 1992; Richman-Hirsch 2001; Tracey et al., 1995).
6. Opportunity to practice skills on the job and post-course follow-up support of trainees (Arthur & Bennet, 1998; Brown & Reed, 2002; Ford et al, 1992; Richman-Hirsch 2001).
7. Pre-training interventions to motivate and focus trainee (Brinkerhoff and Montesino, 1995; Tannenbaum et. al., 1993; Stufflebeam and Wingate 2005).

Of these categories, several authors emphasize organizational environment and support of managers and colleagues as being particularly important. According to one study of 30 articles in the training literature citing factors inhibiting transfer, the negative effect of an unsupportive organizational climate on the transfer process accounts for 42% of inhibiting factors to training transfer, with failure of supervisors to encourage and reinforce application of training the single most commonly cited factor (Foxon, 1993). More broadly, as noted in the previous section, Brinkerhoff argues that the greatest determinant of training success is not the quality of the training program itself, but the “performance system”, including the resources, organizational support and incentives on

the job for application of training and the extent to which systems are in place to ensure that learning is properly targeted to workplace needs. This emphasis on factors outside of the classroom is also evident in the World Bank-IEG's (2008) evaluation of World Bank-financed training, which found that the primary determinants of training success were targeting of training content to organizational needs and the availability of resources and incentives for transfer.

Given the importance of individual trainee characteristics, organizational environment and managerial support, several authors have proposed training evaluation systems which examine both the organizational context in which training is done and the processes used in order to ensure that training content and methods are linked to individual and organizational needs. Berthnal (1995) suggests moving beyond the Kirkpatrick framework of training results evaluation to assess both training process and the variables in the work environment that facilitate or impede the effect of training. Holton et al. (1997) proposed that organizational "transfer climate" be measured through examining the variables of supervisor support, the extent to which trainees have the opportunity to use learning on the job, peer support, supervisor sanctions if training is not used, personal outcomes of transfer (payoffs ranging from positive payoffs such as raises to negative ones such as reprimands), and the extent to which group norms discourage use of training. In addition, Holton et al. proposed looking at two training design factors: content validity (the extent to which content of training reflects job requirements) and transfer design (the extent to which training gives trainees the ability to use learning on the job, including through practice and real world applications). Bramely (1996) suggests investigating a wide range of training management processes, including: the quality and

specificity of training objectives, the balance between theory and practice in course content, the basis upon which didactic methods have been chosen and the use of feedback and performance assessment in-course. Bushnell's (1990) Input, Process, Output, Outcomes model proposes supplementing Kirkpatrick's evaluation of training results (here described as outputs and outcomes) with evaluation of training inputs such as trainee qualifications, instructor abilities, instructional materials, facilities and money, and assessment of the extent to which training is planned, designed and delivered according to good practice principles.

Developmental and formative evaluation of training.

Some authors go beyond proposing ex-post facto summative evaluation of training context and process to suggest that evaluation techniques also be used for developmental and formative evaluation of training, helping training providers better design training as part of a design-to-evaluation continuum (Brown & Gerhardt 2002). These evaluation models are all predicated on the belief that up-front evaluation of needs and of issues in the organizational context can make a significant difference in the success of training programs (Brinkerhoff & Apking 2006; Salas & Cannon-Bowers 2001). Salas & Cannon-Bower's comprehensive review of research on training over the preceding decade found that, "Many training programs fail to reach their goals because of organizational constraints and conflicts, which could have been identified and ameliorated before training was implemented. Hence, conducting an organizational analysis is an important first step in training design."

Stufflebeam's Context, Input, Process, Product (2007) model proposes the following evaluative steps, beginning, where possible, in the planning phase of training (for context and input):

1. Context: Assessment of the extent to which the program targeted to needs, problems, assets and opportunities;
2. Input: Assessment of the extent to which alternative approaches, competing action plans, staffing plans and budgets have been evaluated for feasibility and cost-effectiveness;
3. Process: Ongoing assessment of implementation of training plan to both improve and judge training's value;
4. Product: Identification and assessment of outcomes, both intended and unintended, short- and long-term.

Warr, Bird & Rackham's (1970) CIRO (Context, Input, Reaction, Outcome) model proposes developmental context and input evaluations as well as post-course "Kirkpatrickian" participant reaction and learning transfer outcome evaluation. In their model, a context evaluation is used pre-course to determine training needs. The input phase of their evaluation model looks at available resources and asks how they can be deployed to maximum effect. For example, the CIRO model includes investigation of whether training should be done internally in the organization or externally, how much time should be spent on training, and so forth.

Both Brinkerhoff & Apking (2006) and Brown & Gerhardt (2002) propose that pre-course developmental evaluation assess not only organizational needs but also the transfer climate. Brown & Gerhardt (2002) argue that training evaluation should assess

the extent to which the work environment is likely to support transfer of learning as well the alignment of training objectives, content and instructional methods with organizational objectives and strategies. To do so, they emphasize the importance of involving management in pre-course assessments, portraying the role of the evaluator as bringing together the training and management functions in order to ensure that training beneficiaries are more involved in the design and development of training.

Brinkerhoff & Apking (2006) also propose using developmental evaluation techniques not only to assess the transfer climate but to change it. Like Brown & Gerhardt, they say that training should be used to determine whether workplaces have positive transfer climates and what can be done to improve transfer climate. They argue that improving the impact of learning can best be done through ensuring that (a) training content is focused to organizational needs; (b) trainees understand the way that training is supposed to affect their workplace performance; and that (c) supervisors support training goals and communicate this support to trainees. They propose a number of tools in order to help enhance the trainer-trainee-management “performance system” partnership, making both training participants and their direct and upper level management more involved in training design by having managers take an active role in defining training goals, communicating these goals to trainees and providing follow-up support for training.

In sum, the above authors all advocate that training evaluation be developmental and formative in nature as well as summative. They all argue that a planning-to-evaluation continuum starting with needs assessment and goal definition and ending with evaluation of training results can significantly enhance the effects of training. Some

research substantiates this conclusion. For example, Richey (1992) argues that training designers can improve the likelihood of training transfer by taking into account transfer climate in training design. Richey's research on instructional design for adult learning indicates that factors outside of the classroom are generally more important than quality of training design for training transfer. Based on surveys of almost 200,000 participants in three major corporate-wide training programs in work safety, Richey concludes that, "trainee and organizational climate entry characteristics can account for up to 80% of the knowledge retained after training, leaving instruction and instructional materials responsible for only a maximum of 20% of the variance." However, she adds that training design can be used to compensate for shortcomings in the organizational climate so that, "With excellent design, instruction may count for up to 60% of the variance." (1992:37). Richey details what is meant by excellent design, emphasizing the importance of in-depth front-end analysis of performance discrepancies on the job as well as the organizational context for transfer (management behavior, co-worker behavior, physical work conditions and employee empowerment), trainee characteristics, motivation and incentives. According to her approach, analysis of organizational needs and environment can be used to ensure that training programs target needs and deal with constraints to transfer.

Participatory evaluation of training.

Other benefits to training quality may accrue not as a direct product of such front-end analysis but through use of participatory methods for assessment of training needs and organizational context factors. Participatory evaluation methods, in which program beneficiaries play an active role not only in determining progress toward program goals,

but also in defining what those goals should be, have long been recognized as means to increase ownership of and commitment to programs, enlist stakeholder expertise to reinforce validity of findings, and enhance use of evaluation findings (Bogareert, Bhagat & Bam, 1981; Brandon, 1998; Cousins & Earl, 1995a, Patton, 1982).

Those advocating use of highly participatory evaluative methods for training generally do so in the belief that participatory methods can enhance both managers' support for training programs and the motivation of trainees to learn and apply learning. As mentioned previously, several empirical studies have concluded that the support of managers is one of the most important factors influencing transfer of training (Facteau et al, 1995; Foxon, 1993; Georgenson, 1982; Richey, 1992; Richman-Hirsch 2001; Tannenbaum, 1997; Tracey et al., 1995) For example, Huczynski & Lewis (1980), in a study of two management training courses found that management style and attitudes of the trainee's supervisor were the most important factors in management training transfer. Brinkerhoff & Montesino (1995) found in a study of almost 100 trainees in several courses that when supervisors had pre- and post- course talks with trainees, discussing course content and encouraging its use on the job, trainees reported significantly higher usage of skills gained than in a comparison group who had not discussed the course with their supervisors. On the basis of these and similar findings, several evaluation experts have proposed involving managers in the planning and evaluation of training to enhance the commitment and involvement of management to training and thus augment the likelihood of transfer success (Brinkerhoff & Apking 2006; Dubin et al, 1974; Georgenson, 1982; Gill, 1989; Tannenbaum 1997).

Participation of both managers and trainees in planning and evaluation of training may also increase trainee motivation, in two ways. Trainee motivation may be raised by giving trainees a sense that their input is valued and that course content will take into account their specific needs. In addition, the Brinkerhoff & Montesino study cited above suggests that greater managerial involvement in the training process may improve training results through raising trainee “intentionality” - i.e. trainee understanding of the reasons for training and its intended applications. On the basis of their research, Brinkerhoff and Montesino argue that conducting brief pre-course manager-participant conversations on the purpose of training and on-the job application will significantly increase participant learning gains over time. This finding is reinforced by a study carried out by Warr, Bird & Rackham (1970:54), that found that the use of a pre-course questionnaire increased learning of training content.

The importance of intentionality has also been highlighted in research concluded by Stufflebeam and Wingate (2005) that found that pre-course self-assessment of trainees can be used to heighten intentionality, increase motivation and provide useful information for training needs assessment. Stufflebeam and Wingate distributed a highly detailed pre-course subject questionnaire as part of the application process for a training course in evaluation methods. These forms were used to determine prior knowledge of the training subject, explore how training content related to trainees’ present job responsibilities and determine expectations with regard to future course content. Trainees were then asked post-course to assess learning gains in the same subject areas. While, as the authors of this study acknowledge, self-reported learning gains were not validated through use of learning or performance tests, and thus may have been subject to bias, the authors report

that the questionnaire was valuable both as a needs assessment tool and for evaluative purposes.

Notable in both these above studies is the link that is made between training and pre-and post-course interventions (in the case of Brinkerhoff & Montesino) and between planning and evaluation of training (in the case of Stufflebeam & Wingate). Both of these methods have the effect of more greatly involving training participants and their supervisors in the planning, follow-up and evaluation of training. This link between planning and evaluation may be particularly important in the case of training courses such as the ones studied in this dissertation, which are managed by training institutes external to participant workplaces. Where trainers have little familiarity with the workplace contexts and needs of individual participants, involving participants and/or their managers in pre-course planning can help improve training content. In order to ensure that “external” training is relevant for trainee needs and workplace contexts, Hamblin (1974) argues that, “External training courses cannot be properly evaluated except by means of a concentrated and cooperative effort by all three parties: the training organization, the client organization, and the trainee.”

Goal-free evaluation.

One final approach which merits attention is that of Scriven’s Goal-Free Evaluation (1991). Scriven argued that evaluating results of educational programs against pre-formulated goals is problematic, in that the unintended consequences of programs may at times be as important as the goals that were prescribed in advance. In his words:

Evaluation with respect to goals does not even include all the anticipated effects and gives much too limited a profile of the project...Since all projects either fall short of their goals or over-achieve them, why waste time rating the goals, which usually aren't what is achieved? (p. 58).

Instead, Scriven proposed that the evaluator see his role as mapping all relevant program consequences. Doing so, he argued, is useful not only as a summative evaluative exercise, but also for formative purposes. Better understanding of actual program benefits, including those that were not foreseen in advance, makes it possible for program managers to take these unforeseen, but possibly important, benefits into account in later phases or iterations of the program.

Recently, the goal-free approach has been adopted experimentally by the Dutch Ministry of Foreign Affairs in its efforts to develop better tools for evaluation of capacity development. The Ministry, in cooperation with several Dutch development partners, has launched an experimental evaluation of capacity development programs financed by its aid program (De Lange & Feddes, 2008). Using a lengthy list of possible categories of organizational capacities (for example, organizational leadership capacity and negotiation capacity) the evaluation method seeks to map all changes in capacity occurring over a period of time in a target organization, without reference to the goals or activities of Dutch-funded capacity development interventions. Only once evaluators will have determined what changes in capacity occurred over the program period are efforts made to determine which of these changes may be attributable to Dutch-funded interventions and, subsequently, the positive and negative factors that influenced the extent to which Dutch-funded interventions were able to contribute to organizational

capacity development. The logic behind this approach is twofold: first of all, by adopting a goal-free approach, the Dutch evaluation is recognizing the complex effects that capacity development interventions often have and the necessity of a more in-depth understanding of these effects for those wishing to improve programs. Second, this approach relates to capacity development as an endogenous process within organizations affected by a wide range of internal and external factors. The assumption is made that in order for interventions to better contribute to capacity development, it is necessary to better understand the range of factors affecting capacity, including those external to the aid program. As stated in the terms of reference of the Dutch experimental evaluation:

...capacity development is an endogenous non-linear process that is strongly influenced by a range of internal and external factors. This observation has major implications for donors and recipients and their efforts to find the most effective strategy for identifying and addressing capacity challenges. (p.6)

This approach may be particularly relevant for the sort of multi-organization training courses which are the subject of this dissertation's research. First, the necessity of understanding the complex factors influencing capacity development is at least as important for training interventions as for other forms of capacity development interventions, as has been illustrated earlier in this literature review. Moreover, as such multi-organization courses generally bring together training participants who differ widely in their educational backgrounds, experience, country contexts and workplace functions, participants are likely to benefit from the same training course in very different ways that can not necessarily be predicted in advance and thus cannot easily be measured

through close-ended survey questions. While some courses teaching hard skills, such as accounting courses, may have very specific learning and capacity development goals, more often than not, multi-organization DTI courses may have very different benefits for different participants. Goal-free evaluation may provide a means for developing a richer, more nuanced understanding of the effects of such training courses on participants, which can then be used to both assess programs and further refine them.

Evaluation of Capacity Building in International Development Contexts.

The authors presented above suggest that training evaluation may be made more useful, feasible and with greater construct validity by moving away from frameworks that aim primarily at proving training impact towards ones that aim at improving it through use of process and context evaluation and developmental evaluation. In addition, the use of participatory and goal-free evaluation methods in order to get a richer picture of training benefits was explored.

The case for improving over proving, for process-oriented evaluation, for integrating summative, ex-post-facto evaluation with formative and developmental activities, and for more participatory methods of determining program goals and benefits has also been made in the international development context, both for training courses and more general capacity building interventions. The following section highlights two relatively recent models of evaluation of capacity building and/or training programs in international development contexts – the IDRC’s Outcome Mapping and the World Bank-Independent Evaluation Group’s Three-Tier Training Evaluation model. Both of these models are notable in that they share some of the philosophical underpinnings of the training evaluation models presented above.

IDRC's Outcome Mapping.

The IDRC's Outcome Mapping approach is designed to evaluate capacity building programs targeting single issues or organizations. It is intended for evaluation of a wide range of capacity-building programs. The approach explicitly does not attempt to determine program impact at the organizational level, but, rather, seeks to capture program outcomes, defined as "changes in the behavior, relationships, activities, or actions of the people, groups and organizations with whom a program works directly" (Earl, 2001:1). The choice to focus on outcomes, rather than on measurement of program impacts, is based in the argument that program impacts are difficult both to measure and to attribute to the effects of specific programs. The Outcome Mapping approach argues that achievement of development objectives often requires several different program interventions and is dependent in large part on variables external to program control. As a result, the assumption is made that identification of impact is not a particularly useful evaluation objective, for program accountability, learning and improvement. Instead, Outcome Mapping focuses on outcomes which can be logically linked to a program's activities and that enhance the possibility of development impacts. In this manner, Outcome Mapping attempts to learn how to increase a program's effectiveness in relation to its impact objectives without directly measuring impact.

Outcome mapping goes beyond ex-post facto evaluation, incorporating elements of program design and monitoring. The first stages of outcome mapping bring together program stakeholders in an attempt to establish consensus on the changes that the development program will help bring about, and on the strategies it will use to achieve these goals. At this stage, project beneficiaries, donors, and project implementers decide

on goals, define strategies to be used to achieve goals and set monitorable benchmarks of progress. Goals and progress markers are defined in terms of behavioral change desired as the result of the intervention. Once the program is being implemented, program stakeholders are expected to monitor it by keeping journals in which they assess progress in the implementation of defined program strategies (i.e. monitoring of the program process) as well as progress towards achievement of defined outcomes (i.e. monitoring program outcomes).

Among the notable features of outcome mapping is its heavy reliance on participatory methods, involving program beneficiaries in both the planning and, like Stufflebeam and Wingate, self-evaluation of programs, with the express purpose of encouraging ownership and use of findings. Thus, the evaluative act becomes, to quote the IDRC's manual on the approach, "a consciousness-raising, consensus building and empowerment tool for those working directly in the development program" (Earl, 2001:4). IDRC has found this method to be effective, claiming that program staff are "often more critical of themselves than an external evaluator would be" (Earl, 2001:81).

A second important feature of Outcome Mapping is its combination of both process and results (outcome) evaluation. Information is collected on the program's success in both implementing its strategies and achieving desired outcomes. A well-implemented capacity building program may not achieve its objectives due to external factors. Monitoring both process and results enables evaluators to better define and monitor aspects within a program's sphere of influence, while attempting to trace how process leads to results. In doing so, outcome mapping aspires to generate information

that can be used to improve program performance while collecting credible information on results for purposes of program accountability.

A third interesting aspect to the IDRC approach is its implicit attempt to shift accountability requirements from demonstrable impact to progress towards impact. It does so by focusing monitoring and evaluation only on those program implementation processes and outcome results within a program's direct sphere of influence. Finally, the IDRC approach is notable in its inclusion of a developmental phase, linking program planning to evaluation in a participatory manner that enables program beneficiaries to play an active role in defining objectives, shaping program content and evaluating the success of the program.

World Bank Independent Evaluation Group's three-tier training evaluation model.

The World Bank's Independent Evaluation Group (IEG) recently developed and used a new evaluation methodology in order to evaluate the contribution to the capacity of target organizations and institutions of the estimated \$720 million that the World Bank invests in training interventions in low and middle income countries annually (IEG, 2008). The author of this dissertation was also the task manager and principal author of that evaluation, leading a research team in a review of hundreds of World Bank-financed training courses around the world in most major fields in which the World Bank works, targeting groups ranging from farmers to senior policy makers. The training evaluation methodology developed by the dissertation author and her team for this World Bank evaluation was similar to that of IDRC's Outcome Mapping and to other training evaluation models discussed above, in that it combined a process evaluation with a results

evaluation that focused on the outcomes level. In all cases, evaluated training had already taken place at the start of the evaluation, and thus it was not possible to collect baseline data which could be used to assess training results.

The evaluation used a three-tier model depicted in Figure 1. These tiers were: (a) A results evaluation, placing particular emphasis on workplace performance outcomes, due to the difficulties inherent in attributing organizational impact; (b) An assessment of the validity of the assumptions that would need to hold to progress from one level of the training results chain to the next; and (c) A process evaluation of the presence and quality of good training management practices.

The model was applied in thirty-seven field studies of training programs and in an international survey of over 500 former training participants, six months to two years after they had received training. Data on training results were gathered in both the international survey and the field studies. In the international survey, participant perceptions of training outcomes were assessed, using a combination of close-ended and open-ended questions. Respondents were asked to report the extent to which the course had impact on their work. They were given four response options: (a) “substantial positive changes to the way I perform key or primary functions of my work”, (b) “minor changes to key functions”, (c) “changes to secondary work functions” or (d) “little or no change”. Interviewers conducting the survey were asked to probe for several examples of what respondents thought had changed as a result of taking the course.

While it was hoped that the open-ended section would help gain a better understanding of the results of training, in practice, it highlighted the questionable validity of participant self-reports. For example, of the approximately 40% of

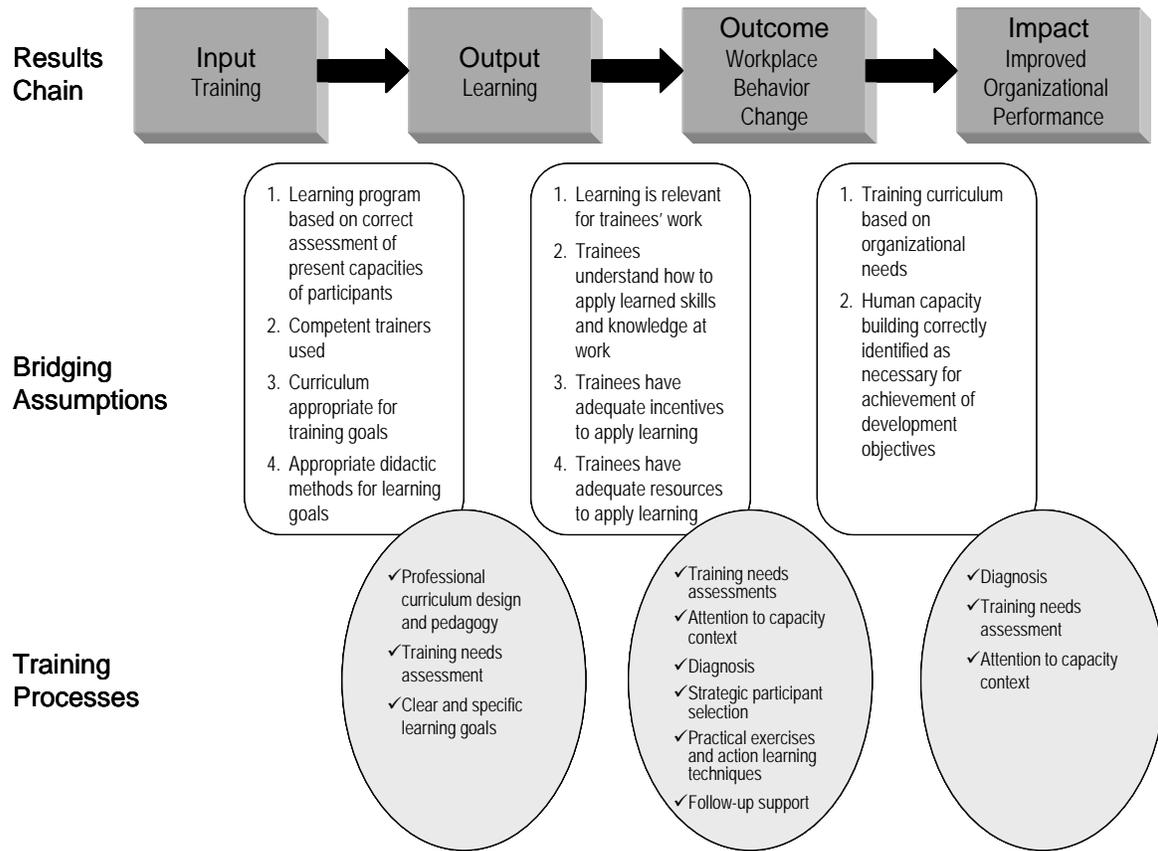
respondents who answered that training had resulted in small changes to key work functions or changes to secondary work functions, fewer than one sixth could articulate any concrete changes at all. It was impossible to judge whether the respondents' inability to cite actual changes points to a positive response bias leading participants to report changes to work as a result of training, even when there were none, or simply difficulties that participants may have had articulating changes that had occurred.

The training field studies, which included review of available documentation as well as in-depth interviews with former training participants, their supervisors, government-policy makers, training providers and the World Bank task team leaders responsible for supervising training financing provided far more credible information with regard to training outcomes. However, even in the field studies it was not always possible to find evidence of training results. In addition, while information could be gathered on training outcomes, it was impossible to empirically demonstrate training impact in most of the cases reviewed.

Consequently, a second tier of evaluation was added. In order to better understand the likelihood of positive training results, the validity of bridging assumptions necessary to get from one level of training results to the next was verified. For example, in order for training to result in organizational impact, the bridging assumption that trainees had the necessary resources and support to implement training and that that training content was based on organizational needs and would have to hold. Otherwise, learning would not lead to changed workplace behavior or changed workplace behavior would not have an impact on capacity to achieve organizational goals. Verifying the validity of bridging assumptions was attempted through survey questions which asked

training participants to comment on questions such as the relevance of training content to their organization's needs, the availability of resources to implement learning and the support of their managers. However, while these surveys provided some picture of the validity of bridging assumptions, the credibility of responses is unclear. For example, while training would have to be relevant to broader organizational goals in order to impact achievement of those goals, training participants may not always be in the best position to determine training relevance to such goals.

Figure 2.1: The IEG Three-Tier Training Evaluation Model



Source: IEG 2008

To supplement information gathered about training results and the validity of bridging assumptions, an international training participant survey was also used in order to determine the presence and quality of important training processes. A literature review conducted in the context of the IEG evaluation identified the training management practices recognized in the literature as important for the achievement of training results (IEG, 2008). The prevalence and quality of these processes in the reviewed training programs were examined both through the field studies and through the international survey. For example, on the basis of ample evidence on the importance of on-the job

follow-up support to learning retention and implementation (OECD, 2006; Stammers, 1975; Taylor et al., 2001), trainees were asked if they received follow-up support, whether this helped them implement what they had learned on the job, and, if not, whether that was because no follow-up support was necessary or because the follow-up support was inadequate for their needs. The assumption was made that participant reporting on the “objective” presence of processes such as follow-up support and use of practice exercises in class was likely to be less influenced by potentially biasing considerations like gratitude for receiving free training abroad. Thus, the participant survey was used to extract valuable information on training quality through questions on training process.

The third tier of Figure 2.1 indicates the training processes identified by the IEG as important for the achievement of results at different links of the chain. The evaluation looked at processes for the planning of training, including diagnosis of organizational needs, training needs assessment and strategies for selection of participants, the quality and variety of didactic methods and the presence of adequate on-the job follow-up support for the implementation of learning.

In addition, three factors that do not neatly fall into the definition of “training process” or results, but were determined to have an important impact on training results were included in the process evaluation: the presence of complementary organizational or institutional capacity support to ensure that trainees had the resources and incentives to implement learning (for example, financing purchase of necessary computer software to enable trainees to use computer applications learned in course), client (i.e. beneficiary

organization) commitment to training objectives and client involvement in diagnosis and design of training.

It should be noted in this regard, that the purpose of the process evaluation was to provide evidence on the existence of good training processes likely to lead to training results. The assumption was made that, as process evaluations focus on factors within the sphere of influence of training organizers, they inherently focus on the elements most important to assess to improve training management performance, and, thereby, training results. Process evaluations may be particularly useful in the case of training, where there is a high degree of consensus as to what constitutes good process. While there is no guarantee that training designed on the basis of adequate diagnosis and needs assessment will succeed in achieving its objectives, it is highly unlikely that training which does not include these process elements will meet organizational needs. In consequence, the three-tier evaluation method used survey questions on training process as a relatively low-cost method for obtaining accurate information on the likelihood of training contributing to organizational capacity building. Data on training processes was thereby used, to some extent, as a proxy for evidence on training results.

Summary

This chapter reviewed relevant literature on training and evaluation. In the first part, quality criteria for the assessment of evaluation models were determined. The review outlined three possible purposes of training evaluation: learning, accountability and developmental functions of evaluation. In addition, three relevant criteria for assessment of training evaluation, as taken from the Joint Committee on Standards for

Educational Program Evaluation, were identified: utility, feasibility and accuracy. The third concept, that of accuracy, was later challenged for training evaluation through participant self-reports and an alternate criterion of credibility, that is adequately capturing respondents' perceptions of reality rather than reality itself, was proposed.

The second part of the literature review examined the literature on the most commonly used training evaluation model: Kirkpatrick's four levels of evaluation of training results. This review outlined two types of critiques of Kirkpatrick's four level framework: The first set of critiques focuses on the claim that presently available tools do not enable feasible collection of accurate data on training results. Specifically, while collecting data on participant satisfaction (Kirkpatrick's Level One) and learning (Level Two) are fairly straightforward affairs, feasible collection of higher level data on participant behavior outcomes and organizational impact is far more problematic. Participant self-reports often do not accurately indicate actual behavior change or organizational impact, and a range of empirical studies have demonstrated that lower-level evaluations, such as end-of-course evaluations are not accurate predictors of training transfer to the workplace.

While some theorists have tried to address the problem of gathering data on training transfer by more precisely focusing level one questions on predicted training use, this solution has not significantly addressed construct validity problems with using level one questions as proxy indicators for training use. Alternately, the experience of the IDRC with outcome mapping suggests that perhaps more participatory methods of evaluation could increase the level of training participant commitment to and ownership of the evaluation process, and thus potentially raise the thoughtfulness and honesty of

participant responses to questions. However, the literature review did not uncover any experimental data to confirm or disconfirm this possibility. In addition, the possibility was broached of receiving richer data on training results through goal-free evaluation methods that enable participants to define for themselves the contribution of training to their work, rather than merely getting them to rate pre-assigned categories. In this case as well, although the Dutch are now experimenting with this method for evaluation of large capacity building interventions, the literature review revealed no documentation of goal free methods being used for training evaluation purposes.

The second category of critiques of Kirkpatrick's framework focus on the usefulness of its exclusive reliance on ex-post-facto results evaluation. Those who critique the Kirkpatrick framework on this basis argue that evaluation should fulfill more than an ex-post-facto, summative function. Instead, the argument has been made that training evaluation should serve developmental and formative functions as well as summative ones. According to this logic, evaluation should be seen in the context of the entire training management performance system. It has been suggested that the evaluation activities can and should be used not only to assess training impact, but also to inform future training, aid in program planning, identify obstacles to training evaluation in the workplace context, improve motivation and intentionality of participants and increase ownership of training processes by participants, as well as their managers and workplace organizations. Approaches such as Stufflebeam's Context, Input, Process, Product (2007) model and Bushnell's (1990) Input, Process, Output, Outcomes model take a far broader view of the type of information that should be gathered in order to enhance the developmental and formative functions of evaluation.

While the literature review unearthed a wide range of critiques of present training evaluation practice and suggestions for improvement, the Kirkpatrick framework of ex-post-facto results evaluation, broken down into four levels that can be assessed independently of each other, remains the only commonly used training evaluation system, in both corporate and international development contexts (Evans, 2007). Indeed, it would seem that, just as there exists a gap between theory and practice on capacity development in international development contexts, there is a similar gap in the field of training evaluation. As of yet, however, no systematic research has been undertaken to document present development training institute evaluation practice, demonstrate where it falls short of both theory and of training institutes' own perceived needs and investigate how those needs could be better addressed. This dissertation endeavors to address this lack through its survey of development training institutes.

Table 2.1 sets the stage for the following chapters by providing a summary of potential improvements to evaluation discussed in the literature, linking them to unmet evaluation needs identified in Chapters One and Two. The table also categorizes each of the identified needs according to the quality criteria categories identified earlier in this chapter, namely, credibility, feasibility and utility. This table is revisited in following chapters, as the basis for the design and assessment of evaluation model to be proposed in this dissertation.

Table 2.1: Summary of Evaluation Needs from Literature Review

Identified Need	Theorized solution
Utility	
<p>Improved and more participatory participant/target organization/country needs assessment as the basis for program planning. (ADB, 2008; Brown and Gerhardt, 2002; Brinkerhoff, 2006; ECDPM, 2009; IEG, 2005, 2007; OECD-DAC, 2006:8; UNDP, 2002; IEG, 2007; UNDP, 2002, 2003.)</p>	<p>1. Inclusion of a developmental evaluation phase. (Patton, 2006).</p> <p>2. Increased beneficiary ownership of needs assessment. Use of participatory needs assessment and evaluation methods where clients, supervisors set personal training goals and report on their achievement. (Brinkerhoff & Apking, 2006; Brown & Gerhart, 2002; Earl 2001; Salas & Cannon-Bowers, 2001). This can also help improve training results through increasing participant “intentionality” – i.e. conscious focusing on their individual goals (Stufflebeam & Wingate, 2005).</p>
<p>Improved level 3 evaluations, including greater measurement of workplace environment factors contributing to training success, because workplace environmental factors account for the majority of the variance on training success. (Brown and Gerhardt 2002; Brown and Reed 2002; Holton 1996; and Brinkerhoff 1988, 2006.)</p>	<p>Use of post-course evaluation forms to identify constraints to application of learning in the workplace. This information can then be used to find ways better support trainees in implementation of learning through training on related issues like change management and other accompanying interventions in the workplace. (Berthnal 1995; Holton et al. 1997; Richey 1992; Stufflebeam 2007; Warr, Bird & Rackham 1970).</p>
<p>Evaluation questions which provide increased variation in responses on training quality in order to provide richer information on training results (Fowler 1995).</p>	<p>Use of goal-free evaluation methods to identify a range of learning and training use goals and achievements, providing ratings for each. (Scriven, 1991; DeLange & Feddes, 2008).</p>
<p>Reinforcing partnership between management, trainers and trainees in order to ensure that actual organizational needs are addressed and to improve training results by ensuring that participants receive support from their managers for implementation of learning. Evaluation should be used to</p>	<p>Use of participatory needs assessment and evaluation methods where training participants, together with their supervisors, set personal training goals and report on their achievement (Brinkerhoff & Apking, 2006; Brinkerhoff & Montesino, 1996; Dubin et. al., 1974;</p>

enhance partnership between training providers, training participants and their workplace managers in order to increase likelihood of positive training results (Brinkerhoff 2006).	Georgenson, 1982; Gill 1989; Hamblin 1974; Tannenbaum 1997).
Improved information on quality of training processes and their contribution to training results. (Brinkerhoff 1987; Mayne 1999; Patton 1986; Holton 1996; Brinkerhoff and Apking 2006).	Continued/increased focus in end-of-course and post-course evaluations on training management process related questions such as quality of lecturers and need for follow-up support. (Berthnal 1995; Bramley, 1996, Bushnell, 1990; Earl 2001, IEG, 2008; Stufflebeam 2007; Warr, Bird & Rackham, 1970).
Feasibility	
To be feasible, evaluations should be inexpensive, relatively non- time-consuming evaluation methodologies that take into account training providers' lack of time and resources to evaluate training's workplace outcomes and impact. (Iverson 2003, 15; Morgan and Casper 2000; Tamkin et al, 2002:ix).	
Credibility	
Better predictors of participant use of learning in the workplace. Level one and two evaluations are inaccurate predictors. (Bramley & Newby 1984; Lee & Pershing, 2002; Warr et. al. 1999). Level three (behavior change) questionnaires measure attitudes with regard to utility rather than actual utility, participant self-reports often inaccurate (Atkins & Wood, 2002; Carless & Robert-Thompson, 2001; Iverson, 2003: 15).	More focus on level three evaluations. Use of participatory methods of training evaluation to increase thoughtfulness of participant responses and "ownership" of evaluation process (Earl, 2001).

Chapter Three: Methods

This chapter begins by restating and elaborating upon the dissertation research questions first presented in Chapter One. It then outlines the methodology used to develop the personal benchmark system, and, finally, defines limitations and assumptions of the study.

Dissertation Methodology

Two research questions guided this study

Q1. What are the existing weaknesses in present models for evaluating training in development contexts?

Q2. What are the key elements of an evaluative model that could feasibly provide more useful and credible information for learning and accountability purposes?

Data collected to answer these two questions was used to build a new evaluation model designed to overcome weaknesses in present models. The study consisted of five main phases:

1. A comprehensive review of the literature on capacity building for development, evaluation of capacity development and training evaluation was completed.

2. DTI representatives were interviewed to map present training evaluation practice and unmet needs of DTIs and to gather examples of good practice.
3. A conceptual model was developed for evaluation of multi-organization or multi-country DTI training courses that may better address training evaluation needs identified in the literature review and DTI interviews.
4. The model was field tested in two international courses at the Golda Meir Mount Carmel Training Centre in Haifa, Israel in order to explore feasibility of the personal benchmark system and collect preliminary data on the utility of the model.
5. Potential benefits and shortcomings of model were assessed and further research needed to validate the model was identified.

Each phase built upon the contributions of its predecessors, with the purpose of gaining a better understanding of present DTI evaluation practice and needs, and developing a new model of training evaluation. Taken in total, the methodology pursued can be seen as an exercise in applied theory-building, with the aim of developing a more feasible, useful, and credible theory of training evaluation that could be applied by DTIs for short-term, multi-country training (Lynham, 2002; Pentland, 1999; Storberg-Walker, 2007).

Methodology for the research was guided by Lynham's proposed approach to theory-building. Lynham (2002: 223) describes theory-building as: "the purposeful process or recurring cycle by which coherent descriptions, explanations, and

representations of observed or experienced phenomena are generated, verified, and refined.” Good theory building, she argues, “...should result in two kinds of knowledge: outcome knowledge, usually in the form of explanative and predictive knowledge, and process knowledge, for example, in the form of increased understanding of how something works and what it means.” Research endeavored to build a theory of both the present shortcomings of DTI evaluation practice and of potential practice improvements that could lead to more feasible, useful, and/or credible evaluations of training.

Lynham sets out two strategies for building theory, one being theory to research and the second research to theory. For this dissertation, the former strategy was used, defined by Lynham as “theory made explicit through the continuous, reiterative interaction between theory construction and empirical inquiry (2002: 227).” Reynolds (1971:144) delineates the stages of this strategy as follows:

1. Develop an explicit theory in either axiomatic or process description form.
2. Select a statement generated by the theory for comparison with the results of empirical research.
3. Design a research project to “test” the chosen statement’s correspondence with empirical research.
4. If the statement derived from the theory does not correspond with the research results, make appropriate changes in the theory or the research design and continue with the research.

5. If the statement from the theory corresponds with the results of the research, select further statements for testing or attempt to determine the limitations of the theory.

Drawing upon Reynolds's above delineation of theory-to-research strategy, Lynham (2002) describes the theory building process as follows:

...both knowledge of and knowledge about the phenomenon central to the theory are brought together through the theory-building process and are ordered according to the internal logic, or logic-in use, and informed imagination of the theorist. This continuous and iterative conversation in applied theory construction, between knowledge and experience of the phenomenon that is the focus of the theory, facilitates the accumulation of relevant and rigorous theoretical knowledge of the phenomenon in the experienced world and is the focus of the theory and the theory-building method itself. (p. 228)

Lynham then summarizes the stages of "research to theory" theory building as follows: (a) conceptual development, (b) operationalization, (c) application (d) confirmation or disconfirmation, and (e) continuous refinement and development (of the theory). In the case of this dissertation, the starting point for conceptual development was the literature review, which was intended to identify existing theories of how evaluation both is and ought to be done. With regard to the first research question (Q1), which asks about existing weaknesses in evaluation methodologies, once the literature review identified possible weaknesses, the DTI interviews were used, where possible, in

order to examine the extent to which these, or other, weaknesses are recognized by DTIs. In the case of the second research question (Q2), on key elements of an improved evaluation model which could address these weaknesses, the literature review was once again used as the preliminary basis for identification of possible improvements to currently used evaluative models. Subsequently, the DTI interviews were used to identify evaluative elements viewed by DTIs as good practice, in order to confirm or revise the theory of elements of an improved evaluative model. Analysis of the results of these two phases formed the basis for conceptual development of the personal benchmark evaluation system. Once an evaluation system was developed on the basis of evidence collected on both evaluation needs and potential solutions for those needs, a field test was used in order to operationalize, apply and, to the extent possible, confirm or disconfirm the personal benchmark system. However, as is explained in the limitations section below, the field test was only designed as a preliminary tool for confirming and disconfirming theory, and checking feasibility of the system's application. Further research would be necessary to validate the system, to explore its generalizability, and to refine and develop it.

The following sections of this chapter provide detailed descriptions of procedures associated with each of the phases of the dissertation research, as well indicating how each phase corresponds with Reynolds's and Lynham's stages of theory building.

Stage One: Conduct comprehensive literature review.

The literature review contributed to stage one in Reynold's stages of theory building - developing an explicit theory of training evaluation. In other words, the review

contributed to development of a theory of the existing weaknesses in training evaluation (Q1) and key elements of an improved evaluation model (Q2), and identified statements to be tested by empirical research in both the DTI interviews and the field test. The literature review was divided into three distinct parts. First, the literature on the intended aims and quality criteria for evaluation was reviewed. The purpose of this review was to set the standards by which the personal benchmark system could be evaluated. Second, Kirkpatrick's four-level model was presented as well as several newer evaluation models which expand upon Kirkpatrick's framework. Critiques of the Kirkpatrick model from academic literature were then presented. Third, alternative methods to Kirkpatrick were presented, in order to identify potential approaches for addressing evaluation weaknesses identified in earlier stages of the literature review. In all these stages, both academic and practitioner literature was included in the review in order to provide a comprehensive picture of both the weaknesses of existing models of training evaluation and potential key elements for an improved model. Data gathered in the course of the literature review were then organized in matrix form, identifying both weaknesses of present models (Q1) and potential key elements of the personal benchmark system (Q2) to be tested in later phases of the dissertation.

Stage Two: Training institute interviews.

The second component in conceptual development of a new theory of training evaluation was a series of interviews conducted with representatives of DTIs. This was necessary to develop an understanding DTI evaluation practice and needs, as almost all of the research cited in the literature review, was done in developed-country, corporate contexts. To date, very little has been written on evaluation practice and performance in

training institutes serving the developing world through the conduct of international courses. DTI interviews were thus used to map present practice, including collecting successful practice examples and exploring challenges and unmet needs.

Representatives of twenty training institutes that primarily conduct international or multi-organization courses for participants from the developing world were interviewed. These twenty institutes include most of the most prominent and reputable DTIs active today. As no comprehensive list of such training institutes exist, the author compiled a list of suitable training institutes through consultations with international donor organizations that finance training in and for the developing world. As a starting point, training institute participants in the ongoing High Level Forum of Development Training Institutes, which comprises many of the most reputable DTIs worldwide, were canvassed. In addition, major donors who finance DTIs such as the WBI and JICA were asked to suggest DTIs whose training courses are sufficiently focused on international development goals and priorities and who are sufficiently well-regarded to have attracted donor financing. In addition, only DTIs whose courses met the dissertation's course criteria (short-term, multi-organization courses aimed at workplace behavior change in international development-related fields) were selected. Efforts to identify, contact and interview suitable development training institutes continued on an ongoing basis until a total of twenty institutes were interviewed, of which five were in Europe, four each in North America, South America and Asia, and three in Africa. Those canvassed conducted training in a wide range of fields, including: legal reform, public administration, private sector development, public financial management, health and

welfare policy, agriculture, rural development, urban policy, gender, civil society and education. Appendix B lists the DTIs interviewed.

The interviews were completed between February 2009 and February 2010. Interviews were conducted on the basis of a structured interview protocol, in part by telephone and in part face-to-face, depending on availability. No substantial differences were found in the content or depth of interview material from these two interview modes (telephone/face-to-face). Senior representatives of the respondent organizations were interviewed for thirty minutes to one hour on evaluation practice. Interviews were conducted with the person responsible for evaluation in the training institute. In institutes without someone charged with this function, senior staff members with overall responsibility for the DTI's training programs were interviewed. English-language interviews were conducted by the dissertation author. Four interviews of Latin American training institutes were conducted in Spanish by Natalia Barrionuevo, who was hired as a Spanish-language research assistant for this purpose, using a translated version of the same interview schedule used for the English-language institutes. Ms. Barrionuevo then provided the dissertation author with translated interview transcripts. The purpose of the interviews was to map present M&E practice, get the perspectives of senior representatives of DTIs with regard to the usefulness and credibility of their monitoring and evaluation systems, and identify possible ways of improving practice, as well as the impediments to such improvements.

In accordance with qualitative interview strategies outlined by Patton in his book "Qualitative Evaluation and Research Methods" (1990), initial interviews were open-ended, using a general interview guide, outlining the set of topics to be explored.

However, after the first three or four interviews, a standardized open-ended interview protocol was developed that was used with all remaining respondents, including the four Spanish-language interviews. While question sequence or wording varied slightly in accordance with the course of the interview, the interviewers made sure to ask each question in the schedule using similar, if not exact wording, for each question. The interview protocol appears below (Figure 3.1). The interview format was semi-structured, with interviewers following up questions from the interview schedule with probing questions to obtain additional information.

Table 3.1: Interview Protocol for DTIs

1. Please describe briefly the training your institute does – course length, purpose, type of participants.
2. How do you plan training content?
3. Do you ask participants to submit information prior to attending the course on their work environments, needs and expectations from the course, or other information that can be used to shape course content?
4. Does your institute generally have any contact with participants' work supervisors or organizations either before the course (eg. to ask about needs, expectations) or after the course (eg. to ask about the usefulness of course content or provide guidance)?
5. Do you feel that your institute is able to improve the training it does over time? If so, what are the main sources of information which you use to improve training?
6. Who does evaluation in your institute? A separate evaluation unit? Course planners? Other?
7. Does your institute have a separate budget for evaluations?
8. Please describe what forms of evaluation you do of training?
9. What sort of useful information to you generally get out of your evaluation?
10. How, if at all, does your organization collect information on the extent to which participants have learned in courses, and the extent to which they were able to apply what they learned on the job in a relevant manner?
11. Are there types of evaluative information which you are at present unable to gather that you feel would help your institute improve training results? If so, what information would be useful to you and why are you unable to gather this information?
12. Do you at any point in time during the planning, implementation or evaluation of training, gather information about the specific work contexts of participants, such as: support of managers and colleagues for training, available resources for implementing changes suggested by training, availability of follow-up support structures at the workplace.
13. Are there specific elements of your evaluation regime that you would like to improve? If so, what?

The decision to opt for qualitative interviews based on open-ended questions, rather than a close-ended, more easily quantifiable survey was made in order to facilitate gathering of more in-depth, richer information on both the ways in which training institutes do evaluation and their perspectives on how evaluation should and could be improved. As Marshall & Rossman (1989) note, the strengths of qualitative interviewing over more quantifiable survey instruments include delving into complexities and processes, exploring why certain practices do not work, and investigating innovation. Similarly, Patton (1990) suggests that qualitative methods are most appropriate when research is focused on the process, implementation or development of a program or its participants; when detailed-in depth information is needed; when there is a focus on unique qualities exhibited by individual subjects; or where the intent is to understand the program theory. As the major aims of this study conformed in most part with these criteria, qualitative interviews based on an interview schedule was chosen as the best way to obtain rich descriptions of DTI evaluation practice, solicit elements of best practice and develop a deeper understanding of unmet evaluation needs and the challenges inherent in addressing these needs.

There are clearly, however, certain risks and limitations involved in using qualitative interviews as a form of evidence. First, information obtained through self-reports of interviewees may be less accurate than information obtained by empirical observation or measurements (Marshall and Rossman 1989: 83). However, any form of direct observation was infeasible, given the diverse geographic location of development training institutes. A second problem with qualitative interviews is the possibility that interviewer bias will affect results, either through the interviewer asking questions in a

leading way or during the analysis phase of the research. To mitigate this possibility, a standardized open-ended interview schedule was used for all but the initial few interviews. In addition, in accordance with Guba and Lincoln's guidance (1989: 181), interviewers used informal "member checks" in the course of the interviews, with interviewers repeating back to interviewees their understanding of subject responses. Furthermore, several of the interview questions asked for similar information from different perspectives as a further check on the thoroughness of the process. For example, survey respondents were asked in separate questions to describe what forms of evaluation they conduct of training and, in a subsequent question, to describe how they collect information on the extent to which participants have learned in courses, and the extent to which they were able to apply what they learned on the job in a relevant manner.

Once completed, information collected in the DTI interviews was fed into to the matrix first constructed at the end of the literature review. An additional column was added to the matrix, detailing evidence from the DTI interviews on present DTI evaluation needs and practice, as compared with evaluation needs and solutions identified in the literature review (Table 4.1). For example, where the literature review found that most training institutes do not conduct level three or four evaluations due to lack of time and resources, the DTI interviews were used to provide evidence with regard to whether this is also the case in training institutes targeting developing country participants. In addition, where additional training evaluation needs or innovative solutions to training evaluation needs identified in the literature review were found in the DTI survey, these were also added to the matrix.

Stage Three: Development of a preliminary theoretical model and field test.

On the basis of information gathered both through the literature review and the training institute interviews, a new theoretical model for evaluation of multi-country or multi-organizational training in international development contexts was developed. To facilitate theory building, a final column was added to matrix used to summarize the literature review and DTI interviews (Table 6.1). This column stated what elements of the personal benchmark system were used to address evaluation needs identified in earlier phases of the research.

Once the initial model was developed, organizers of the two field test courses were consulted with regard to useful course-specific questions for the two subject courses (see below). The author endeavored to develop a model which would meet both learning and accountability evaluation needs identified in the literature review and interviews of training institutes and address feasibility, usefulness and credibility/construct validity issues associated with evaluation models presently enjoying widespread use. This model was then operationalized and applied, to use Lynham's theory-building terminology, in two field tests.

It should be noted, that the field test was not intended to prove the model's benefits or lack thereof, but rather to provide information for further assessment and refinement of the model. The decision to limit the purpose and scope of the field test to refinement of the model through two field test courses, rather a more wide-spread test with the purpose of proving its benefits, was based on the following assumptions: First, the difficulties associated with feasibly collecting accurate, attributable data on training

impact, as discussed in the literature review, rendered it impossible to prove at reasonable cost whether the dissertation evaluation model generated more accurate data on training outcomes and impact than existing models. Even if a control group had been administered more widely used types of evaluation questionnaires and discrepancies were found between participant responses in the test and control groups, it would not be possible to empirically establish which group's results were a more accurate reflection of reality without doing prohibitively costly on-site baseline and post-course measurement of participant behavior change.

Second, the diverse nature of training topics, goals, target populations, cultural and institutional contexts, length, and didactic methods and so forth severely limit the generalizability of findings. In order to develop a fuller understanding of the benefits and shortcomings of the personal benchmark system presented in this dissertation, it would be necessary to test it repeatedly in a wide variety of contexts. This is beyond the scope of this dissertation. In consequence, while the field test was used to better understand issues that might arise in application of the new dissertation evaluation model and to spot potential issues with or unforeseen advantages to the model, it was not used to provide generalizable evidence on its utility or credibility.

The model as field tested included three questionnaires: one administered on the first day of class, one administered at the end of course, and one administered three months after course completion. The purpose of the field test was not to validate the personal benchmark system, but rather merely to gather preliminary data on the feasibility of the system and needed refinements. Full validation of the system was considered beyond the scope of this dissertation. Further research is needed to gather

data both on this applicability and generalizability. The full system as developed is presented in Chapter Four of this dissertation.

Field test subject courses.

The personal benchmark system was field tested on two two-week courses sponsored by MASHAV (Israel's bilateral aid agency) held concurrently in September 2009 at the Mount Carmel Training Centre in Haifa, Israel. One of the courses, on Gender and Local Governance, was given in cooperation with UN-Habitat, and drew heavily from the Habitat sourcebook for trainers on Gender in Local Government. The second course, on Support Systems for Microfinance, was given in cooperation with the Organization of American States' Young Americas Business Trust. Both courses were conducted in English and included about 20 participants from countries throughout the developing world. The following are brief descriptions of the courses reviewed. Each course was organized and accompanied by experienced course staff at the MCTC, although most course content was delivered by guest lecturers. With a few exceptions, MASHAV provided full tuition, room and board for the course. Course participants were expected to finance or find financing for their own plane tickets to Israel.

MASHAV-MCTC-UN Habitat course on Gender and Local Government: The course's stated aim was to strengthen gender mainstreaming in local government through training of local government trainers and other relevant persons. Its objectives included: increasing understanding of why gender is important in local government; recognizing gender biases in government and obstacles to the advancement of women; familiarizing participants with existing tools for gender

and local governance training; enhancing confidence to deliver gender training to diverse groups and exposing participants to Israeli models for gender mainstreaming. Course participants included local government trainers from government-affiliated training institutes and NGOs based in developing countries worldwide. The complete course description appears in Appendix C.

MASHAV-MCTC-OAS-YABT course on support systems for microenterprises:

The course objectives were defined as threefold. First, broadening knowledge of support systems for small-scale entrepreneurs as a tool for regional and local development; second, familiarization with the support systems for micro- and small- entrepreneurship in Israel; third, analysis of the position of young and older entrepreneurs at times of global economic crisis. The course targeted entrepreneurs, representatives of governmental institutions and small and medium enterprise (SME) support agencies, non-governmental organizations (NGO)s dealing with small business support and development, women's business associations, and employees of UN organizations, International Financial Institutions and other national and international agencies dealing with entrepreneurial development. The complete course description appears in Appendix D.

Implementation of field test.

As noted above, the purpose of the field test was to provide information on both problems with and benefits of implementation of the system in order to both assess the feasibility of the model and provide insight into further improvements that could be made

to the personal benchmark system. Since utility and feasibility, as defined in Chapter Two of this dissertation, are subjective concepts rooted in perceptions of those administering evaluations and using the data generated by them, interviews with the organizers of the two subject training courses were conducted at several points in the field test, in order to elicit their views on the utility and feasibility of the model. In addition, the researcher's own observations as to the success of the model in addressing identified evaluation needs were recorded.

Stage Four: Assessment and review of implications.

The final phase of this dissertation, as presented in Chapter Six reviewed findings from the literature review, DTI interviews and field study in order to assess the likely merits and limitations of the training evaluation model to be proposed. This phase corresponds with Lynham's stage of confirmation and disconfirmation of the theory (2002). Analysis of the field test results followed Reynolds's (1971:144) guidance to select statements generated by the theory for comparison with the results of empirical research and then to design a research project to "test" the chosen statements' correspondence with empirical research. These questions for analysis are delineated in Chapter Five.

After critically reviewing the field test results on the basis of the pre-defined questions listed in Chapter 5, the dissertation revisited the three quality criteria of utility, feasibility and credibility, analyzing the extent to which the personal benchmark system successfully addressed shortcomings of standard practice identified in the literature review and DTI interviews. As part of this analysis, recommendations were made for

improvements to the personal benchmark evaluation system as field tested, and for further needed research. Finally, needed conditions for application of the model in DTIs were discussed.

Potential Limitations

For the dissertation's discussion on potential limitations of the research, the author has opted to use Lincoln & Guba's (1999) conditions of trustworthiness rather than the more commonly used framework of threats to validity as enunciated in Campbell and Stanley's "Experimental and Quasi-Experimental Designs for Research" (1963). This section first presents Lincoln & Guba's argument against Campbell and Stanley, as well as the alternative criteria for "truth value" of research which they propose. It then explains why the author opted for this framework of analysis over Campbell and Stanley's. Finally, it presents an analysis of the limitations of this research based on Lincoln & Guba's criteria for truth value.

Lincoln & Guba's framework was developed for field studies in naturalistic settings, as an alternative to Campbell and Stanley's framework. Lincoln & Guba argue that criteria such as internal and external validity, reliability and objectivity are appropriate only from an ontological and epistemological perspective of "naïve realism" – that is, the belief that there is an absolute truth which can be arrived at through carefully constructed scientific inquiry. Where the researcher sees reality as a social construct, and the role of the scientist as capturing subjective constructions of reality rather than arrive at "objective truth", criteria such as objectivity, reliability and validity take on very

different meanings than those intended by Campbell and Stanley. For example, Lincoln & Guba (1999) deny the possibility of establishing the internal validity of experimental findings if one holds by a non-rationalistic paradigm of scientific knowledge:

On the assumption of a single, tangible reality, that an investigation is intended to unearth and display, the ultimate test of internal validity for the conventional inquirer is the extent to which the findings of the inquiry display an isomorphism with that reality. But the determination of such an isomorphism is in principle impossible, for, in order to make it, the inquirer would need to know the nature of that ultimate tangible reality a priori. But it is the precisely the nature of that reality that is at issue. If one already "knew" it, there would be no need to mount an inquiry to determine it. (p.402)

Instead, Lincoln & Guba argue that naturalistic inquiry is based on the assumption of multiple constructed realities, where there is no ultimate benchmark of truth. In consequence, in place of internal validity, they recommend the criterion of credibility, as noted in Chapter Two of this dissertation. They argue that the researcher's responsibility is not to present as accurate as possible a description of objective reality but rather to demonstrate that he or she has adequately represented the multiple constructions of reality being studied, or, in their words, the researcher must show that, "the reconstructions...that have been arrived at via the inquiry are credible to the constructors of the original multiple realities" (1999: 402).

Lincoln & Guba also argue against the concept of reliability by asserting that it is impossible to replicate experiments in naturalistic (as opposed to laboratory) settings and obtain similar measurements. Citing the well-known parable of the river that cannot be crossed twice because one is inevitably crossing over different water, they argue that, in naturalistic enquiry, the vast diversity of the human experience makes replicability and thus "reliability" impossible. Instead, they propose the concept of dependability, which involves developing and depicting an understanding of factors of instability and design-related drivers of change (p.406).

Finally, they take a very different view of external validity than Campbell and Stanley, arguing that the natural world, unlike world of laboratory experiments, is too complex for the researcher to be able to specify precisely in which cases or under what conditions study findings might be generalizable, or, in Lincoln & Guba's terms, transferable. They claim that it is the prerogative of the person seeking to apply research findings elsewhere, and not of the original inquirer, to determine the applicability of the research to his or her case. The researcher's responsibility is merely to provide a sufficiently rich description of his or her working hypotheses, together with a description of the time and context in which they were found to aid those seeking to determine whether the research is applicable to their case (1999: 420).

This dissertation has adopted Lincoln & Guba's criteria for trustworthiness of research for two reasons. First, the author's ontological and empirical perspective is more closely in line with that of Lincoln & Guba than that of Campbell & Stanley. Second, and perhaps more importantly, the author believes the framework to be more appropriate for analysis of the data collected. First, the research was conducted in a naturalistic

setting rather than an experimental one, through analysis of two field tests. Campbell and Stanley see little value in these sorts of case studies, which they argue, "have such a total absence of control as to be of almost no scientific value (1963: 6)." Second, as detailed in Chapter Two, since empirical measurement of participant workplace performance and behavior change due to training is prohibitively expensive for international courses, the dissertation methodology is necessarily based on gathering subjective accounts of training results rather than empirically confirmable proof. As a result, the dissertation research was much more in line with Lincoln & Guba's representation of the researcher's responsibility as adequately representing the multiple constructions of reality existent within a given case. Below, the dissertation methodology and its limitations are discussed using Lincoln & Guba's four criteria for establishing the "trustworthiness" of research.

Credibility and dependability. Lincoln & Guba's standards for credibility, their replacement criterion for internal validity, are based on establishing the study subjects' confidence in the "truth" of the findings of a particular inquiry. Credibility is thus from the perspective of the subjects of the research, and not of outside readers. Among the techniques Lincoln & Guba propose for increasing the likelihood that credible findings will be reached are: prolonged engagement with the test group, in order to deepen the researcher's understanding of the subject groups culture; developing trust; and testing for misinformation introduced by distortions either of the self or of the respondents. These proposed techniques could only be very partially applied in the research. Lincoln & Guba's standards for credibility examine the criterion from the point of view of the subjects of the inquiry. In the case of this research, there were three relevant subject

groups: a. interviewed representatives of DTIs, b. participants in the two courses that were the subjects of the field study and c. course coordinators for these two courses who were asked to provide their assessment of the dissertation evaluation methodology. Of these three groups, the researcher only had prolonged engagement with the course coordinators, with whom she has had a working relationship for several years in different contexts. In the case of the DTI representatives, in some cases, the only engagement with the researcher was in setting up a time for the interview and in conducting the interview. In the case of the course participants, the researcher attended the opening and closing of one of the courses, on Gender and Local Government and attended most of the session of the second course, on Support Systems for Microenterprises. However, in neither of these cases, nor with regard to the DTI interviews, can it be plausibly argued that the researcher's engagement with study subject was prolonged enough to develop trust or reduce the possibility of respondent misinformation.

To the contrary, both in the case of the DTI interviews, and in the case of training participant responses to the field test of the dissertation evaluation method, there are clear reasons why respondents may have felt it in their interest to provide distorted accounts to the researcher. In the case of the DTI interviews, it is possible that respondents may have wanted to portray the work of their institutes in as complementary a light as possible. Moreover, their own perceptions of the credibility of the information obtained from their evaluation protocols may be distorted. In the case of the training course participants surveyed as part of the field test evaluation protocol, it is quite possible that their satisfaction or gratitude at receiving a scholarship for a course in Israel may have

influenced their assessment of the course's impact on their work. This issue is discussed at greater length Chapter Five's discussion of the field study implications.

Lincoln & Guba also propose triangulation of research methods and/or sources to enhance credibility. In the DTI interviews, this could possibly have entailed two types of activities: interviewing multiple respondents in each DTI or comparing evidence from the DTI interviews with other studies with regard to training institute evaluation practice. This study opted for the latter form of triangulation. To this end, interview results were matched against results of other studies done of training institute evaluation practice. However, the author did not triangulate findings about individual training institutes interviewed by speaking with multiple informants within each DTI. While it is possible that multiple interviews may have yielded slightly more nuanced or richer information on DTI practice, the author did not feel that, given the interviews' intended purpose of providing an overview of DTI evaluation practice and needs, this higher level of data credibility was necessary. Data from the DTI interviews was reported in aggregate form only (with the exception of good practice examples), so that any variations in descriptions of DTI evaluation activity that may have come from interviewing more than one DTI representative would be unlikely to significantly alter findings. Instead, efforts were made to interview the person in the training institute most likely to have the most information of DTI practices and needs. In addition, the dissertation evaluation model was designed on the basis of three sources of input: the literature review, DTI interviews and feedback from field test organizers. By using three separate sources of input, the author of this study hoped to enhance the credibility of her findings with regard to the dissertation evaluation model.

An additional method of establishing credibility is that of member checks. According to Lincoln & Guba, member checks are the "most crucial" technique for establishing credibility. Member checks involves either formally or informally checking data, analytic categories, interpretations and conclusions with members of those stakeholder groups from whom the data were originally collected (Lincoln & Guba, 1999: 418). In the DTI interviews, and in gathering assessments of field study course coordinators toward the evaluation methodologies used, member checks were carried out by interviewers in the course of the interviews, by repeating key points to interviewees for confirmation.

In sum, through triangulation, member checks and prolonged engagement with course organizers, the author of this study hoped to enhance the credibility of findings. Having said that, none of these methods as applied in the research entirely mitigated the challenges to credibility inherent in a study largely reliant on self-reports of respondents whose responses might be influenced by external incentives and incomplete information. In consequence, it is important to note that, while the study endeavored to establish the theoretical likelihood that the personal benchmark system could enhance the usefulness, feasibility and credibility of DTI training evaluation, it could not prove that this model overcomes these limitations. As mentioned earlier, further research would be necessary in order to validate the system and better understand its benefits and limitations.

Transferability. As noted above, Lincoln & Guba's concept of transferability differs greatly from that of external validity in that they see the duty of the researcher as merely providing a rich enough description of the research process and aims enable someone interested in making a transfer to decide whether this may be possible (1999:

420). Their rejection of the concept of external validity for studies in naturalistic settings is based on the contention that generalizability can only be ensured in transferring experiments from one controlled, laboratory environment to another laboratory environment with similar controls.

Similarly, the author of this study acknowledges that lessons on the implementation of the dissertation methodology in the field study training programs are not necessarily applicable to training conducted at other training institutes, in other countries, on other topics, with training participants of different academic or professional levels. Further testing of the model in a range of different contexts is necessary in order to develop a richer picture of how the personal benchmark system might be used by potential users.

Dependability. Lincoln & Guba's third criterion of trustworthiness is meant to take the place of reliability. Lincoln & Guba argue for dependability, or "instrumental unreliability", suggesting that the researcher merely include in the account of his research an understanding of factors of instability within data collected that could change the results of the study depending on when it was administered. In the case of this dissertation, the main concern with regard to instrumental unreliability was that training participants involved in the field test may have responded in different ways to a post-course evaluation had it been administered at different times. Specifically, it is plausible that had the post-course questionnaire been administered two years after course completion rather than four months, it would have yielded substantially different results. With the course experience still fresh in the minds of participants, response rates could be

higher, and responses more detailed, than had participants been polled years later. This point is further addressed in Chapter Five's discussion of the field test and conclusions.

Summary

This study explored existing weaknesses in present models of training evaluation for assessing the contribution of training to organizational and institutional capacity in international development contexts. On the basis of this analysis, a new model was developed for evaluation of DTI training aimed at workplace behavior change, with the ultimate aim of helping target organizations and institutions better achieve their objectives. The field test did not attempt to provide objective, empirical proof of either the results of training courses assessed in the field study or, consequently of the relative ability of the dissertation evaluation methodology to provide accurate, empirically provable data on the results of training courses in the field. Instead, due to the nature of the data which could be feasibly collected as part of the field test, this study aimed only to present the attitudes of research participants – including DTI representatives, course organizers, course participants and the dissertation author, and on that basis to provide a theorized model for improved training evaluation that may be replicable to other international training courses targeting participants from developing countries.

There are two primary potential limitations to this research. The first principal limitation relates to transferability of the model. The results of two case studies do not provide sufficient data to the reader to determine whether the dissertation evaluation model would be easily transferable to any given context. Further field studies would be

useful in order to build up a greater body of evidence on this question. Second, the author acknowledges the imperfect reliability/dependability of the field test, based on the fact that results may be influenced by the timing at which the post-course questionnaire was distributed.

Chapter Four: DTI Interviews and Building of the Conceptual Model

This chapter presents results of the second and third phases of research outlined in Chapter Three. Specifically, it first presents results of the interviews on Development Training Institute (DTI) policies, practices and preferences. Then, it outlines the theory building process used to develop the dissertation's personal benchmark evaluation system on the basis of data collected through the literature review.

Development Training Institute Interviews

The DTI interviews revealed many themes common to a wide range of DTIs. There were no notable differences in DTI concerns between regions or even between developed and developing countries. Due to the small number of DTIs interviewed from each subject field, it was not possible to identify any differences in DTI attitudes based on the subject matter of training provided. However, generally, DTIs training on largely technical skills, such as public financial management, expressed greater satisfaction with their current evaluation regimes than those teaching "softer" skills and subject matters whose applications are more context-specific, like rural development or public administration.

As noted in Chapter Three, the DTI interviews were intended to collect information on three topics: a) present DTI course planning and evaluation practice, b) unmet DTI planning and evaluation needs and obstacles to meeting those needs, and c) good practice examples. The table in Appendix E presents a full schedule of interview responses.

Present DTI planning and evaluation practice.

Needs assessment.

Most DTIs do not consult with either training participants or their workplace organizations in designing training. When asked how training content was planned, the expert knowledge of course organizers was cited by all of the training institutes interviewed. In addition, 5 of the 20 DTIs interviewed used local coordinators to help them shape training content – either from local training institutions or relevant government bodies. Only four training institutes responded to this question by citing any form of direct needs assessment of beneficiary organizations or of training participants – in all but one case this was done through pre-course participant application forms which asked participants or their supervisors about their workplace functions and needs. However, when asked in a subsequent question whether they ask participants to "submit information prior to attending the course on their work environments, needs and expectations from the course or other information that can be used to shape course content", nine DTIs stated that their pre-course participant questionnaires include questions on participant needs and circumstances, two DTIs used pre-tests of participant knowledge and one DTI did a baseline survey of participants' workplace organizations.

End of course participant satisfaction questionnaires.

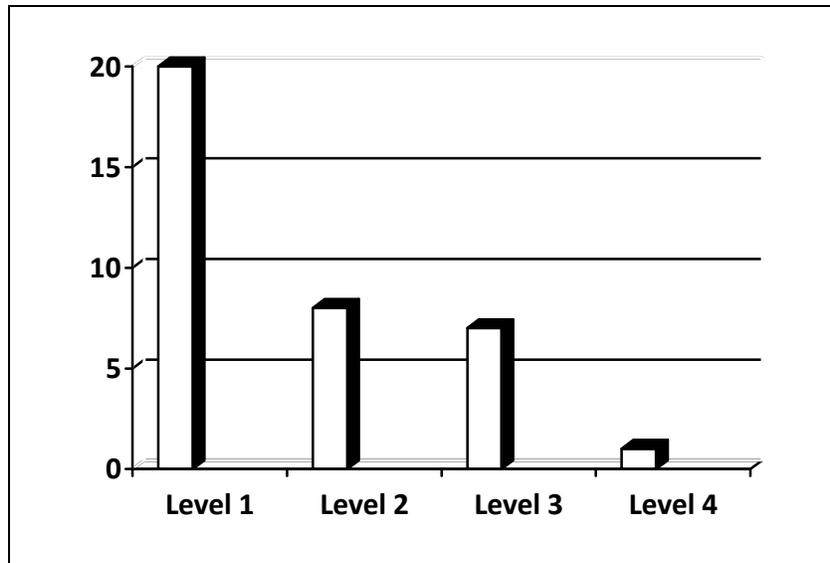
All interviewed DTIs administer end-of-course participant satisfaction questionnaires (Kirkpatrick's level one), generally for all participants. Many institutes found that questions on the quality of lecturers and materials, the mix between lecturers and practical work, and the balance of time accorded to various topic areas enabled them

to learn how courses could be improved. Two institutes, however, said that they conduct end-of-course evaluations primarily to satisfy donors. These institutes claimed that they had neither the time nor manpower to process and analyze level one questionnaires, which were done largely on a pro-forma basis.

Post-course evaluations.

Unsurprisingly, while all institutes administer some form of end-of-course evaluation questionnaire, far fewer administer higher level evaluations. Three training institutes reported doing formal learning (level two) evaluations through participant tests, which in two cases included pre-tests of knowledge as well as post-tests. In addition, five training institutes stated that they were able to measure participant knowledge through end-of course projects and presentations. Seven stated that they do at least some form of workplace performance outcomes evaluation (level three) for at least part of their participants – almost exclusively in the form of written or electronic questionnaires. One institute also reported doing impact assessments (level four) occasionally.

Figure 4.1: Prevalence of Levels of Evaluation

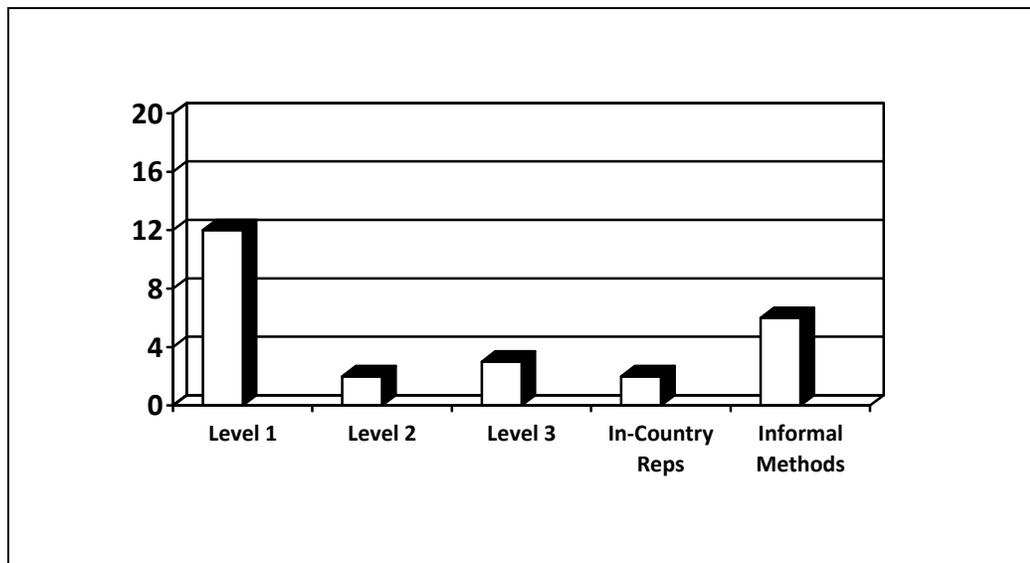


Evaluative learning.

When asked what sources of information they use to improve training, twelve DTIs cited information from end-of-course participant satisfaction questionnaires on training topics and didactic methods, two institutes cited level two evaluations, two cited level three. Two of the institutes engaging in public financial management also canvass coordinators from beneficiary countries and/or institutions. A higher percentage of DTIs cited informal methods of information gathering as being among the main sources of data used to improve training over time than cited formal evaluative material. The importance for DTI evaluative learning of informal sources of information -- that is, information gathered not in a structured, purposeful manner but through ad hoc contacts -- was one of the key findings of the DTI interviews. Specifically, some DTIs cited case studies presented by students during the course as an important method of gathering information

on realities on the ground, and ongoing informal contacts with past participants through websites, alumni forums, return participants and informal telephone and email contact after the course as an important source of information for improving course content over time. For example, when DTI representatives were asked whether they gather information on the extent to which participants learned in courses and were able to apply what they learned on the job in a relevant manner, only two organizations cited level three evaluations, while six other organizations cited informal methods of information gathering.

Figure 4.2: How Do DTIs Learn?



The main source of informal information was participant feedback given in informal conversations during and after the course, cited by eleven DTIs. Others spoke about feedback from course organizers (two DTIs), feedback from participants' workplace organizations gathered informally by course organizers and other DTI personnel (1) and use of a comment box in the DTI (1). Several training institutes claimed that the information gathered in these informal ways was more useful in

understanding course results than the scores gathered as part of a formal evaluation process. Some suggested that informal conversations with past participants provided a much richer, more nuanced and useful picture of course effects than the responses which participants gave in evaluation questionnaires. In a second question, when asked what sort of useful information they get out of evaluations, only three DTIs cited something other than information gathered from level one evaluations. These three training institutes all cited level two information on participant learning. None of the DTIs cited in this context their level three evaluations as providing them with useful information.

Supervisor feedback.

Few training institutes have any substantive contact with the workplace supervisors of participants. Six training institutes ask for supervisor input in application questionnaires, although generally this involves supervisors "signing off" on participant application forms rather than contributing their own perspective. Three training institutes at least occasionally speak to supervisors post-course to get feedback or provide guidance.

Evaluation methodology.

All pre-course participant needs assessments and level three evaluations were done through written surveys. One DTI, as is discussed more at length in the good practices section, has a policy of phoning dissatisfied level three survey respondents in order to understand in greater depth the source of dissatisfaction. None of the training institutes surveyed interview participants or their supervisors by phone either pre- or post-course as part of a formal planning or evaluation regime.

Who does evaluation and how is it funded?

Six of the twenty DTIs surveyed had a separate budget for evaluations. Four DTIs had a separate evaluation unit. In most training institutes, however, evaluations were conducted by course organizers. Five of the surveyed DTIs used independent outsiders to conduct periodic evaluations. DTIs with separate evaluation units, unsurprisingly, seemed to have put more thought into methods for improving evaluation than those that relied exclusively on course coordinators. However, the importance of informal contacts between course organizers and participants, both during and after the course in gathering useful information on course improvements, suggests that it is useful that course organizers play a central role in conducting any form of evaluation with DTI learning as its goal.

In sum, DTIs, like their Northern, corporate counterparts (as demonstrated in Chapter Two), almost exclusively focus on end of course participant satisfaction questionnaires in order to both provide them with information to improve future activities and to assess the quality of past activities. These questionnaires were found to be particularly useful in helping training institutes decide on needed changes in lecturers, course materials, didactic methods and topics. Conversely, few training institutes use any formal evaluation methods in order to gauge the impact of training on participants' work or upon organizational achievements. Of those that did believe that they were able to determine the extent to which training impacted workplace performance, all cited informal methods of information gathering such as online chat rooms and ad-hoc contacts between course organizers and former course participants.

Unmet DTI needs.

When asked what types of evaluative information DTIs are at present unable to gather, but would help improve training results, the majority of DTIs referred to information on participant use of learning at work and obstacles to applying learning. Sixteen DTIs cited information on workplace use of training as the main type of evaluative information they would like improve. Four DTIs cited their desire to develop stronger participant networks post-course, inter-alia, as a means of gathering better information on post-course use of training. In addition, three DTIs cited more contact with supervisors post-course and three cited better information on learning. Finally, two DTIs cited a desire for better needs assessment, and the same number spoke of the need for improved level one questionnaires.

The three most commonly cited constraints to improving evaluation regimes were the cost of evaluations, the difficulty getting a clear, accurate picture of workplace use of training through questionnaires and the low response rates on level three questionnaires. Several DTIs commented on their need for better level three evaluation techniques to overcome these constraints. As one training institute noted, there is a trade-off between the cost of an evaluation and the accuracy of survey responses. It is not generally cost feasible to invest the necessary resources to get a clearer picture of training transfer to the workplace. In this context, several DTIs cited the almost universal tendency of participants surveyed to assess courses positively, whether due to the fact that dissatisfied participants are less likely to submit questionnaires or because participants do not respond frankly to survey questions. One DTI representative interviewed was of the opinion that nationality of training participants affects their ratings of training success,

another cited the problem of gathering honest information from participants when training is being provided as a free good. There was a sense among several interviewees that face-to-face methods such as focus groups and participant observation are far superior to written surveys for gathering credible data on workplace application of training, and yet only written surveys are feasible in terms of cost and donors tend to prefer written surveys for their quantifiability.

Other obstacles noted to improving evaluations included lack of support for greater time and cost investments in evaluations on the part of DTI funders and management. For example, one DTI explained that their funders are only willing to fund the training itself, and not pre-course needs assessment or post-course evaluation, even though these tools would be useful in order to improve training. In addition, two training institutes discussed difficulties with processing, analyzing and using the amount of evaluative information they gather. These two training institutes both have separate evaluation units that are responsible for evaluations of all their training programs.

In a separate question, when asked what elements of their present evaluation regimes DTIs would like to improve, eleven respondents cited more level three evaluations, with higher response rates, providing more accurate and detailed pictures of on-the-job use of training. In addition, two DTIs spoke of problems processing the amount of paper and data generated by their present evaluation procedures so that information gathered is not used.

Good planning and evaluation practice.

A number of good practice examples were revealed in the interviews. Among the training institutes that conduct level three evaluations, some expressed the belief that they were able to get higher response rates to these evaluations by maintaining ongoing relationships with training participants, generally through online communities or listserves. For example, the Latin American and Caribbean Institute for Economic and Social Planning in Chile frequently has the same participants returning for further courses, enabling them to learn over time about the needs of participants and their success at applying learning. The Justice Studies Centre of the Americas (CEJI), also in Chile, invites participants to organize replicate courses in their home countries, which also provides CEJI with the possibility of getting a much clearer picture of how suitable learning material is for conditions on the ground in participants' countries. The Institute of Rural Management (IRM) in Pakistan, in an attempt to ensure higher response rates for level three questionnaires, asks participants at the end of training to draft a handwritten letter to themselves detailing how they intend on using what they've learned. This letter is then dispatched to trainees a few months after the course along with a stamped, pre-addressed envelope, and trainees are asked to report on their progress towards their goals. Another interesting practice of the IRM is to place follow-up phone-calls to course participants who did not respond to level three questionnaires, in order to ascertain why they did not respond and to discuss their attitudes toward the course. This is done on the assumption that dissatisfied participants are less likely to respond to questionnaires than satisfied ones, and yet their perspective is necessary to better identify sources of dissatisfaction that might be addressed by the institute. Finally, IRM's level three

questionnaires ask for supervisor feedback as well as participant feedback as an important source of information on course usefulness.

The Administrative Staff College of India (ASCI) assigns a mentor for every trainee half-way through the course. The mentor is responsible to give continued support to the trainee, both during and after the course, including through participation in an online group with mentors, experts and training participants. At the end of their courses, ASCI also writes a report to each participant's superiors on what was covered in the course in order to try to engage managers in processes of change. In addition, ASCI tries to strengthen the link between learning and application of learning by having each participant prepare a project in-course, and then, in a supplementary session two months after the course, report on progress toward implementation of the project. In this supplementary session, participants present their progress to an advisory group whose role is to make useful suggestions. According to ASCI, this supplementary session is an important source of information on how to improve training. In addition, ASCI identifies the most promising "champions" of change among training participants and then provides them with ongoing support in their workplaces. This is done in order to support change, where possible.

Finally, the International Development Law Organization (IDLO) noted the importance of the involvement of DTI training program staff in determining evaluation questions and tools. While IDLO has a separate evaluation unit that is now working to introduce significant improvements into the organization's evaluation regime, the evaluation unit sees its role as facilitating an institute-wide process engaging all relevant IDLO stakeholders. In particular, the IDLO representative emphasized that they see

programming staff as the primary clients of evaluation and thus believe they should be the primary determinants of what information should be gathered as well. In addition, the evaluation unit hopes that the institute-wide process will help enlist the support of IDLO staff across the institution for changes to their present evaluation regime.

Several of the other major Northern-based capacity development institutions, including, World Bank Institute (WBI), Internationale Weiterbildung und Entwicklung gGmbH (INWENT) and the United Nations Development Program (UNDP) have recently been working on developing results frameworks for the planning, monitoring and evaluation of human capacity development programs. These results frameworks for the most part build on the Kirkpatrick model of results monitoring by elaborating specific questions that should be addressed in planning and evaluating training results. More in-depth consideration of these frameworks, however, is beyond the scope of this dissertation for three reasons: First, these frameworks are still in early experimental stages so that it is difficult at this point to draw conclusions with regard to their utility or feasibility. Second, these frameworks all demand very detailed needs assessments and extensive monitoring and evaluation, which are often not feasible for the short-term, relatively inexpensive training courses that are being examined in this dissertation. Third, for the most part, they provide guidance with regard to *what* is to be monitored and evaluated, rather than *how*. Consequently, they present few innovative tools and mechanisms can be used to collect required data. As a result, while these various results frameworks have the potential to provide good practice ideas in the future, they were not deemed to be particularly relevant for this research.

In sum, the innovative practice ideas on evaluation methodology that were forwarded were most generally concerned with raising response rates of formal level three questionnaires and developing informal methods of gathering data on post-course. No specific innovative practices were identified with regard to the format or content of formal evaluation mechanisms.

DTI interview summary and analysis.

The interviews revealed a number of patterns. The one aspect of their evaluative regimes which training institutes would most like to improve is that of information on workplace use of training and on workplace environmental factors influencing implementation of training. Constraints to improving post-course evaluations (level three) include lack of money, time, and feasible methods for gathering accurate information. To the extent that training institutes believe themselves to be capable of gathering information on workplace contexts or implementation of training, this tends to be through informal contacts with participants and their workplace organizations during and after the course, as well as information gathered in in-class case studies, rather than through their formal evaluation regime. Few of the institutes interviewed administer formal post-course evaluations on a regular basis. Conversely, all training institutes administer formal end-of-course participant satisfaction questionnaires and all but one interviewed believe that information on training content and topics gathered through these questionnaires is useful for improving future training courses. After better post-course evaluations, the second most frequently cited need was a desire to improve their capacity to conduct participant needs assessments.

The interviews also suggested that, while separate evaluation units can be helpful in processing evaluative information gathered, training program managers/course coordinators should play an active role in the process. This is for two reasons. First of all, informal information gathering generally is done by the course coordinator, who is the most likely DTI staff person to maintain contact with participants post-course. Removing the evaluative function from the hands of course coordinators makes it more difficult to take advantage of these informal mechanisms of information gathering. Moreover, removing evaluation from the purview of those administering the courses can result in a disconnect between information collectors and consumers, contributing to greater difficulty with processing and use of evaluative information gathered. Having said that, separate evaluation units can play an important function in setting up a thoughtful evaluation regime, monitoring administration of evaluations by programming staff, and analyzing overall trends.

The DTI interviews confirmed certain evaluation needs first identified in the literature review, and contributed to a deeper understanding of the specific needs of DTIs and possible avenues for solving these needs. The literature critiques the feasibility and of collecting data on workplace participant behavior training outcomes and impact and the lack of construct validity inherent in the practice of using end-of-course evaluations as proxy indicators of training use in the workplace. However, while the challenge of better post-course data collection was identified by DTIs as their most prevalent need, there is little extant research on how to improve the feasibility of credible post-course data collection on training transfer. To the contrary, the more innovative approaches to training evaluation presented in the literature review did not aim to better document

results, but, rather, to add elements such as process and context evaluation to results evaluation. While these latter elements are important for understanding the reasons behind training results, and for using evaluations to improve training performance, none of these enhanced models of training evaluation actually help improve the quality of data on results. Thus, the literature on training evaluation has not as of yet provided a sufficient response to the DTI challenge of feasibly collecting more credible data on higher-level training results through low-cost evaluation methods.

The DTI interviews reveal, conversely, one possible avenue to enhance the feasibility of post-course evaluations of participant behavior change through increasing the participatory nature of evaluation. As noted, low response rates to post-course evaluation written or electronic questionnaires is recognized by DTIs as a substantial hurdle to such evaluations, but some DTIs found that continuing dialogue with participants after they have returned to their workplace has a positive impact on response rates of post-course questionnaires. Moreover, the literature review suggests that enhancing the participatory nature of evaluation may have other positive effects. Specifically, the IDRC has suggested that greater beneficiary ownership of capacity development program evaluation may enhance the quality of data collected, and the work of Brinkerhoff and Stufflebeam, among, others, suggests that giving training participants and their workplace managers a greater role in needs assessment and program evaluation may increase the relevance and impact of training, enhance evaluative learning and improve the performance of trainees upon return to the workplace.

The DTI survey also suggested a way to enhance the usefulness of data gathered. Several DTI representatives, when asked how they are able to learn about the quality of

their courses, cited the particular importance of informal methods of data gathering such as informal conversations with participants both during and after the course, and phone-calls with other representatives of participants' workplaces. However, while these informal contacts may generate a wide range of information that can be used to help understand and improve course results, such informal methods are not sufficiently rigorous to fulfill the accountability function of evaluation. Thus, it can be argued, that the challenge is how to use formal evaluation mechanisms in order to generate information which, on the one hand, is as rich as that obtained through informal conversations, and yet is sufficiently rigorous and quantifiable to meet the accountability expectations of donors and other stakeholders. These two principles - of increasing the participatory nature of evaluations, and formalizing the rich information on participants' experiences that at present is generally collected through informal methods – guided design of the personal benchmark system of evaluation.

Below, to complete this discussion of evaluation DTI interview results, the summary of evaluation needs from the literature first presented in Table 2.1 is revisited, with the addition of a third column on present practice in DTIs as documented in the interviews (Table 4.1, Column 2). In addition, an additional row has been added to the Feasibility section on the basis of DTI comments on the need to address low level three response rates.

Table 4.1 Summary of DTI Evaluation Needs

Identified Need (Literature Review)	Present Practice (DTI Interviews)	Theorized solution
Utility		
<p>1. Improved and more participatory participant/target organization/country needs assessment as the basis for program planning. (ADB, 2008; Brown and Gerhardt, 2002; Brinkerhoff, 2006; ECDPM, 2009; IEG, 2005, 2007; OECD-DAC, 2006:8; UNDP, 2002; IEG, 2007; UNDP, 2002, 2003.)</p>	<p>Needs assessment rarely done. Short courses generally planned without input from training participants or their workplace organizations.</p>	<p>1. Inclusion of a developmental evaluation phase. (Patton, 2006). 2. Increased beneficiary ownership of needs assessment. Use of participatory needs assessment and evaluation methods where clients, supervisors set personal training goals and report on their achievement. (Brinkerhoff & Apking, 2006; Brown & Gerhart, 2002; Earl 2001; Salas & Cannon-Bowers, 2001). This can also help improve training results through increasing participant “intentionality” – i.e. conscious focusing on their individual goals (Stufflebeam & Wingate, 2005).</p>
<p>2. Improved level 3 evaluations, including greater measurement of workplace environment factors contributing to training success, because workplace environmental factors account for the majority of the variance on training success. (Brown and Gerhardt 2002; Brown and Reed 2002; Holton 1996; and Brinkerhoff 1988, 2006.)</p>	<p>Level three evaluations rarely done by DTIs. Better/more level three evaluations is the evaluation need most commonly expressed by DTIs. Training planners and providers have very little information about the workplace context of trainees or constraints to transfer of training to the workplaces. Training does not take into account whether trainees have sufficient</p>	<p>Use of post-course evaluation forms to identify constraints to application of learning in the workplace. This information can then be used to find ways better support trainees in implementation of learning through training on related issues like change management and other accompanying interventions in the workplace. (Berthnal 1995; Holton et al. 1997; Richey 1992; Stufflebeam 2007; Warr, Bird & Rackham 1970).</p>

	support to use learning.	
3. Evaluation questions which provide increased variation in responses on training quality in order to provide richer information on training results (Fowler 1995).	Training institutes routinely report that virtually all of their courses rated highly by participants. Uniformly high grades do not provide trainers with an in-depth picture of what may be improved in courses.	Use of goal-free evaluation methods to identify a range of learning and training use goals and achievements, providing ratings for each. (Scriven, 1991; DeLange & Feddes, 2008).
4. Reinforcing partnership between management, trainers and trainers in order to ensure that actual organizational needs are addressed and to improve training results by ensuring that participants receive support from their managers for implementation of learning. Evaluation should be used to enhance partnership between training providers, training participants and their workplace managers in order to increase likelihood of positive training results (Brinkerhoff 2006).	Training managers have little contact with participants' workplace organizations.	Use of participatory needs assessment and evaluation methods where clients, together with their supervisors set personal training goals and report on their achievement (Brinkerhoff & Apking, 2006; Brinkerhoff & Montesino, 1996; Dubin et. al., 1974; Georgenson, 1982; Gill 1989; Hamblin 1974; Tannenbaum 1997).
5. Improved information on quality of training processes and their contribution to training results. (Brinkerhoff 1987; Mayne 1999; Patton 1986; Holton 1996; Brinkerhoff and Apking 2006).	DTIs generally satisfied with the quality of the process information gathered from level ones and other in-course satisfaction questionnaires.	Continued/increased focus in end-of-course and post-course evaluations on training management process related questions such as quality of lecturers and need for follow-up support. (Berthnal 1995; Bramley, 1996, Bushnell, 1990; Earl 2001, IEG, 2008; Stufflebeam 2007; Warr, Bird & Rackham, 1970).
	Feasibility	
6. To be feasible, evaluations should be inexpensive, relatively non- time-	Training evaluation commonly restricted to level	To be feasible, evaluation should be done primarily through written questionnaires (DTI Interviews).

<p>consuming evaluation methodologies that take into account training providers' lack of time and resources to evaluate training's workplace outcomes and impact. (Iverson 2003, 15; Morgan and Casper 2000; Tamkin et al, 2002:ix).</p>	<p>one and informal collection of information.</p>	
<p>7.</p>	<p>Improved response rates needed. Several DTIs cited problem of extremely low response rates on level three questionnaires.</p>	<p>Several DTIs suggested that maintaining formal or informal contact with participants after the course helps raise response rates to evaluation questionnaires.</p>
	<p>Credibility</p>	
<p>8. Better predictors of participant use of learning in the workplace. Level one and two evaluations are inaccurate predictors. (Bramley & Newby 1984; Lee & Pershing, 2002; Warr et. al. 1999). Level three (behaviour change) questionnaires measure attitudes with regard to utility rather than actual utility, participant self-reports often inaccurate (Atkins & Wood, 2002; Carless & Robert-Thompson, 2001; Iverson, 2003: 15).</p>	<p>Training institutes tend to use questions on course utility in level 1s as their sole evidence of utility. Level threes, where done, tend to ask only general questions on utility and relevance of courses. Responses generally very positive, regardless of course quality.</p>	<p>More focus on level three evaluations. Use of participatory methods of training evaluation to increase thoughtfulness of participant responses and "ownership" of evaluation process (Earl, 2001).</p>

Development of a Conceptual Model for Training Evaluation

As detailed in Chapter Three, the personal benchmark system for training planning and evaluation was developed through a theory-to-research strategy of theory building. According to Lynham's (2002: 228) explanation of this process, the first stage is conceptual development, bringing together knowledge about the phenomenon central to the theory and then ordering it, "according to the internal logic, logic-in-use, and informed imagination of the theorist." This process should result in a conceptual model which can then be field tested. This section presents the logical process by which the dissertation author developed her model of training evaluation, based on data on DTI training evaluation needs, constraints and possible innovative solutions gathered through the literature review and the DTI interviews. The author used the table of evaluation needs (Table 4.1) as part of her theory-building process, to organize collected data and develop possible evaluative solutions.

For purposes of clarity, the logical process used to develop the system is presented in reverse order. In other words, rather than starting with evaluation needs identified in research and then detailing the solutions to those needs proposed by the new model, the description below begins by presenting the elements of the new system proposed, and then will demonstrate how these elements are intended to address evaluation needs. Where possible, the ensuing description will reference the row numbers in Table 4.1 for each of the evaluative needs listed so that the reader can better follow the connection between the system's proposed elements and the needs detailed in that table. In addition, the reader may reference the actual survey questionnaires used in

the field test in order to better understand how these principles were operationalized in the field test.

Aims of the model.

The literature review and DTI survey identified two main training evaluation issues. First, DTIs need better information on post-course participant use of training and behavior change. The most cost-feasible method for doing so is distribution of post-course evaluation questionnaires to former participants, but these generally result in very low response rates and do not generate a sufficiently rich picture of how training affected participants and their workplace organizations. Second, a considerable amount of literature argued for the need for evaluation systems that are geared toward improving program planning and design rather than only proving program results.

To date, the broader evaluation systems involving elements such as pre-course workplace context and needs assessment, process evaluation and post-course context evaluation proposed in the literature have not been widely adopted by training institutes (Evans, 2007). In addition, there has not been a considerable amount of work done by theorists on improving the quality of post-course evaluations. On the basis of both the literature review and the DTI survey, this dissertation identified a number of basic principles that could inform an improved system: from the literature review, the possibilities of goal free evaluation, using more participatory methods, and integrating within evaluation systems planning, process and context elements were discussed. From the DTI survey, the potential benefits were identified of both more participatory methods,

and methods that enable the formalization of the type of rich data generally only gathered informally.

The personal benchmark system integrates all of these elements in order to address the two challenges listed above. The system mandates three written questionnaires to be administered to all training participants: one prior to the course, one at the end of course and a post-course questionnaire distributed several months-one year after their return to their workplace. The system enables participants to define for themselves the criteria upon which the course is evaluated through a series of overlapping questionnaires. The first questionnaire asks participants to set for themselves individual learning goals. These individual goals are then redistributed to participants at the end of course and they are asked to rate the extent to which they achieved their goals. They are then asked to set individual goals for use of training and then are given the opportunity, in a third, post-course questionnaire, to rate training use goal achievement and to discuss the obstacles to achievement of their goals. The following description will discuss the elements of this system in greater detail:

Enhancing the participatory nature of evaluations (Table 4.1 Rows 1, 4, 8, 9).

The personal benchmark system enhances the participatory nature of evaluations by asking participants to determine for themselves the parameters upon which they rate their training experience. In practice, this was accomplished in the following manner: On the first day of a training course, the model calls for the distribution of a pre-course questionnaire asking participants to define up to three individual learning goals, and to explain the ways that they hope to use this learning in their workplaces. The

questionnaires are not anonymous in order to enable participants to review their individual learning goals at the end of the course, as part of the next phase of the evaluation model. Accordingly, on the last day of course, participants are redistributed the individual learning goals which they defined for themselves at the beginning of the course, and are asked to rate the extent to which the course helped them meet each of their individual goals. In addition, participants are asked to name the three most useful knowledge or skill areas which they obtained from the course, and how they believe they will be able to use this knowledge or those skills on the job. These responses are then redistributed to participants in the final post-course questionnaire, three months to one year they have returned to their workplaces, and they are asked to rate the extent to which they were able to use on the job the knowledge or skills, as expected at the end of the course. This method is a variation on Scriven's Goal-Free method of evaluation (1991), which calls for evaluations which do not assess programs on the basis of pre-defined goals, and is intended to enhance the participatory nature of the evaluation by having participants themselves define the parameters by which the course is to be evaluated.

Enhancing the participatory nature of evaluation was chosen as an important element of the personal benchmark system to address a number of issues which were raised in the literature review and DTI interviews: First, as the review of capacity development studies in Chapter One indicated, higher levels of beneficiary ownership are associated with capacity development intervention success (ADB, 2008; Land, Hauck, & Baser, 2009; IEG, 2007; OECD-DAC, 2006:8, UNDP, 2002, 2003). Asking participants to set for themselves their learning and training use goals and to assess the achievement of these goals is intended to enhance participant perceptions of ownership of

the training process and thus contribute to the likelihood of training success (Brinkerhoff & Apking, 2006; Brown & Gerhart, 2002; Earl 2001; Salas & Cannon-Bowers, 2001). Moreover, asking participants to define up-front their learning goals and, subsequently, the ways that they intend on using learning on the job may increase intentionality – that is, conscious focusing of the participants on their individual goals in order to improve success rates (Brinkerhoff & Montesino, 1995; Stufflebeam & Wingate, 2005). Third, use of participatory methods is intended to increase response rates, once again by enhancing participant perceptions of ownership of the training process (Earl, 2001). Finally, having participants define for themselves the criteria by which courses are to be evaluated is intended to increase the credibility of responses. In other words, by having participants devote thought to their specific learning goals, and on-the-job training use goals, it was hoped to increase the thoughtfulness with which training participants fill in evaluation questionnaires and thereby to increase the extent to which evaluation scores are reflective of actual participant perceptions of training results (Earl, 2001).

In addition, use of participatory, goal-free methods of defining and rating achievement of training goals was intended to provide more useful data on learning and training use. In international courses, such as the ones that the personal benchmark system is designed to address, participants in any given course tend to have widely different backgrounds, workplace contexts and responsibilities, and thus, learning needs. This method of rating individually chosen goals was intended to give trainers a richer picture of what trainees actually gained from the course and how used learning in the workplace. Moreover, having participants give separate ratings for each of their learning and training use goals was intended to increase variation of participant responses.

Variation between ratings of different learning and training use goals aimed to give readers of the evaluation forms a more detailed, and thus more useful understanding of participants' attitudes on which course topics were most useful and/or most successfully delivered (Fowler, 1995, Table 4.1, Row 3). Thus, for example, this method could be used to calculate separate scores for different learning topics or use areas which could then be used to see which topic were comparatively stronger or better delivered.

Questionnaire one: Pre-course assessment (Table 4.1 Rows 1, 4).

Both the DTI interviews and the literature review identified a need for greater use of needs assessments in the design of training interventions. A review of the literature on capacity development suggested that capacity development and training frequently suffers from being insufficiently targeted to beneficiary needs due to lack of adequate needs assessment (OECD-DAC 2006:8, UNDP 2002; IEG 2005). In fact, as was revealed in the IEG (2008) study on training for capacity building, poor targeting of training to participant needs is one of the two primary reasons for training failure. The DTI interviews revealed that while most training institutes do not use information gathered from participants or their workplaces in order to develop course content, some training managers are able to use information gathered from participant application questionnaires on participants' work functions or learning needs in order to gain a deeper understanding of the needs of specific participants.

In order to address this general lack of knowledge on the specific needs and circumstances of training participants, the personal benchmark system incorporates a pre-course questionnaire asking participants to define their learning goals and detail how that

learning is relevant for their workplace functions. In addition, the pre-course questionnaire includes questions asking participants to describe relevant elements of their country and workplace contexts. This was intended to provide training organizers with information that could be used to help shape training content to address the specific needs of training participants.

For the field test, the decision was made to administer the pre-course questionnaire on the first day of course, after course organizers had presented the course program to participants, rather than prior to the start of the course. This was done for two reasons: a) to simplify evaluation procedures; and b) to raise the likelihood that participants would draft individual learning goals within the actual parameters of the course program, rather than choosing entirely unrelated or tangential personal goals (Iverson 2003, 15; Morgan and Casper 2000; Tamkin et al., 2002:ix).

Questionnaire two: Evaluation of training management processes (Table 4.1, Row 6) and learning results

The second phase of the personal benchmark evaluation system called for an end-of-course questionnaire to be distributed to participants on the last day of course. The DTI interviews indicated that all DTIs administer end-of-course participant satisfaction questionnaires, and most DTIs find these questionnaires to be useful. In particular, DTIs cited that participant satisfaction questionnaires produce particularly useful information on training management processes such as the quality of lecturers and course materials, the mix between lectures and more participatory forms of learning, the length of course

and the quality of facilities. These types of information are found by many DTIs to be more useful than ratings on course quality, which tend to be almost uniformly positive.

In accordance with this DTI feedback on post-course satisfaction questionnaires, the personal benchmark system maintains these training management process elements of standard level one questionnaires but modifies the elements dealing with training results. As noted above, participants are not asked to provide general ratings on the extent to which they believed that they have learned on the course but, rather, to assess achievement of their individual learning goals as cited in the pre-course questionnaire and asked to rate the extent to which they achieved their goals. This method is a way to quantify the qualitative information on learning goals that participants themselves provided. In doing so, participants are empowered to use evaluations to communicate the specific benefits or shortcomings of training for their needs and training evaluators receive a much more rich understanding of both what was learned and what was important to participants to learn than would be possible through standard, close-ended questions. In this way, the participant benchmark evaluation system attempts to capture in a quantifiable manner the type of nuanced information on training achievements that the DTI surveys suggested is normally only generated through direct, informal conversations with participants.

Once participants have completed the learning goal section of the end of course questionnaire, they are also asked to choose up to three useful knowledge and/or skill areas that they gained in the course and how they intend to use these things at work. This is intended to give DTIs a clearer picture of what elements of the course are most valued

by trainees and to prepare for the third post-course phase of the personal benchmark system.

Phase three: Post-course assessment of results and obstacles to transfer of training (Table 4.1, Rows 2, 5, 9, 10).

The DTI interviews revealed that none of the DTIs administer post-course assessments of implementation of training for all courses whether by email/mail survey, telephone or other methods. This stands in stark contrast to end-of course participant satisfaction questionnaires which are generally distributed by DTIs to all participants. Among the obstacles cited in the DTI interviews to better and more widespread post-course evaluations were low response rates to mail or email questionnaires, the perception that respondents to written questionnaires do not generally accurately represent their post-course experience with transfer, but rather provide positive ratings regardless of reality, and the prohibitive cost or time demands of other methods of evaluation like telephone interviews, observation or focus groups. And yet, as noted in the previous section, improving post-course evaluations was cited by a plurality of DTIs as the element of their evaluation regime that they would most like to improve. The literature review also suggested the particular importance of a post-course phase of evaluation. Research has indicated that the majority of variance on workplace implementation results of training is due to factors in the workplace environment, rather than didactic methods, instructional design, or other classroom variables (Brown and Gerhardt, 2002; Brown and Reed, 2002; Holton, 1996; IEG, 2008; Brinkerhoff, 1988, 2006.) Given the importance of workplace environment to training success, a third post-course phase of the evaluation, administered a minimum of three months after course

completion was made an integral part of the new three-stage evaluation model. The three-stage evaluation model calls for written questionnaires to be distributed by email or mail to all participants. Mail or email questionnaires were found, in the DTI interviews, to be the only time and cost-feasible method for post-course evaluation. Thus, the three stage model uses only written questionnaires, but endeavors to improve the level three response rates and utility and credibility of information gathered over commonly used evaluation questionnaires.

In the participant benchmark system, several months to two years after participants return to their workplaces from training, evaluators redistribute participants' individual end-of-course responses on the most useful things they gained in the course. Participants are then asked to rate the extent to which they were able to use learning in the ways they intended. Rating course achievements on the basis of participants' individual training use goals is intended to fulfill two purposes: first, as in the case of ratings on learning goals in the post-course questionnaire, this format was intended to raise response rates by giving participants a greater feeling of ownership of the evaluation process and to produce more thoughtful, nuanced and varied (and thus more useful) information on the ways in which training participants have used the knowledge and skills which they have acquired in course. In addition, participants are asked in the questionnaire a series of questions in order to determine whether various relevant elements in the workplace environment, such as availability of resources and support of managers contributed to or impeded transfer of training. These questions are a crucial part of the post-course questionnaires with the purpose of giving DTIs a clearer picture of

how environmental factors affect transfer in order to enable them to address such factors in future courses (Stufflebeam, 2007; Warr, Bird & Rackham, 1970).

Summary

This chapter detailed the results of the DTI interviews and then described the process used to build a conceptual model of training evaluation, based on data gathered in the literature review and DTI interviews. The following chapter will now detail the implementation and results of the field test of the personal benchmark system.

Chapter Five: Field Test of the Personal Benchmark System

In accordance with the model presented in Chapter Four, the field test involved three questionnaires: one administered on the first day of class, one at the end of course, and one three to four months after course completion. Both field tests followed a similar evaluation protocol. All questionnaires were designed in consultation with course organizers, who provided important input with regard to course content and the types of information which would be most useful to them. For this reason, there was a certain degree of variation between questionnaires in the two courses. Questionnaires were not anonymous in order to enable redistribution of participant responses to them at the end of course for the post-course questionnaire. Participants were informed that they had the right not to fill in the questionnaire if they were uncomfortable with its non-anonymity. No participant in either the course “opted out”. The following is a detailed description of the questionnaire design and rationale for the design, as well as findings from the field study.

Pre-course questionnaire.

The pre-course questionnaire was designed to provide course organizers with more detailed information on the home country contexts of training participants and their needs and expectations from training. In addition, some of the questions on participants’ home contexts were asked again at the end of course in order to track changing participant attitudes. Finally, participants were asked to designate up to three individual learning goals on the basis of which they would be asked at the end of the course to assess learning. The following is a question-by question presentation of the

questionnaires distributed at the beginning of the course. The actual text of survey questions is presented in italics. Where variation occurred between the questionnaires for the two courses, this too was noted. The full questionnaires, as distributed, appear in Appendices F and G. The following questions were asked in order to provide information to trainers on participants' country contexts and needs:

Gender course:

1. In your opinion, what are the three main challenges to gender mainstreaming in your country today?

Microfinance course:

1. In your opinion, what are the main areas in which microenterprises are in need of support in your country?

2. Are you aware of support systems for microenterprises in your own country? If so, what are they and do they address the challenges listed above?

3. In your opinion, what sort of improvements to existing support systems do you believe need to be developed to better support microenterprises in your country?

In addition, the following goal table (Table 5.1) was distributed to participants of both courses. Participants were instructed to fill in the two left columns of the table and to leave the two right columns blank until the end of the course.

Table 5.1: Learning Goal Table¹

What knowledge, skills, or other benefits would you most like to gain from the course? Please be as specific as possible. You may list up to three goals.	For each of the goals listed above, how do you expect to use what you have gained here once you have returned to your workplace?	Now that you have completed the course, please review the goals you set for yourself. For each goal, please tell us the extent to which the course met your needs. Please rate on a scale of 1-4 where 1 = did not meet my needs 4 = fully met my needs. If the goal you set was not relevant for the course, please select N/A	Comments
Goal 1:		1 2 3 4 N/A	
Goal 2:		1 2 3 4 N/A	
Goal 3:		1 2 3 4 N/A	

End-of course questionnaire.

The second phase of the evaluation was an end-of-course questionnaire. In the case of the Microenterprises course, the questionnaire was distributed on the second-to-last day of the course. For the Gender course, the questionnaire was distributed on the last day. This enabled the researcher to make modifications to the questionnaire on the basis of the experience of the first day. In addition to the non-anonymous questionnaire, and in keeping with normal evaluation practice in the MCTC, participants in both courses were asked for their detailed feedback in face-to-face feedback sessions and participants

¹ Table has been compressed here for purposes of readability. The actual questionnaires left more space for participants to fill in responses.

in the Microenterprises course were asked to fill out the anonymous, standardized, on-line, end-of-course participant satisfaction questionnaire used for all MCTC courses. The questionnaire included questions on various aspects of the training management process as well as detailed feedback forms for each lecturer in the course. While there was some concern that participants would experience “evaluation burnout” and give insufficiently thoughtful, detailed responses to the dissertation field test’s post-course questionnaire, this did not seem to be the case.

The questionnaires administered as part of the personal benchmark system were also not anonymous in order to enable redistribution to participants at later phases. Participants were reminded of their right to “opt out” if they were uncomfortable with this. No participants declined to submit the questionnaire. The following questions were asked in the end-of course questionnaire:

First, to track participant learning and shifts of attitude over the period of the course, one question about country context asked in the pre-course questionnaire was repeated in the post-course questionnaire:

Microfinance course:

1. On the basis of what you’ve learned here, what sort of improvements to existing support systems/new support systems do you believe need to be developed to better support microenterprises in your country?

Gender course:

1. In your opinion, what are the three main challenges to gender mainstreaming in your country today?

Subsequently, two standardized questions were asked of participants in both courses:

2. To what extent was the course relevant for the particular circumstances of your country?

This was a closed-ended, quantitative question, with participants being asked to rate the course on a scale. In the case of the Microfinance course, a scale of 1-4 was used. In the case of the Gender course, participants were asked to choose between the options: Not relevant, slightly relevant, mostly relevant and extremely relevant.

3. Please suggest up to two improvements that could be made to enhance the success of this course.

This open-ended question was meant to generate richer descriptive data on participants' attitudes with regard to the course.

Participants were then redistributed the individual goal tables that they filled out on the first day of course and asked to rate in the two right columns of the table the achievement of their learning goals, as defined by them on the first day of the course in the two left columns of the table. As in the case of question two of the end-of-course questionnaire, a numeric rating scale was used for the Microfinance course and a descriptive rating scale was used for the Gender course. In the Gender course, the answer choices offered for the learning goal table question on whether the course met participants' individual learning needs were as follows: not at all, partially met my needs,

mostly met my needs, entirely met my needs, not applicable: course content did not address this topic.

Finally, a second “utility” table was distributed to participants (Table 4.3). The table asked participants to name the three most useful things which they gained in the course and how they intended to use these things at work. As in the case of the individual learning goal table, this table was not anonymous in order to make it possible to redistribute individual responses to participants in the third phase of the evaluation protocol.

Table 5.2: Training Use Goal Table

Please share with us how useful the course will be/was for you.²

Of the knowledge/skills which you acquired in the course, which do you believe will be most useful to you back on the job? You can chose up to three knowledge/skill areas.	How do you believe you will use the knowledge/skills in your work?
Knowledge/skill area #1:	
Knowledge/skill area #2	
Knowledge/skill area #3	

In the case of the Gender course, in order to mitigate any possible biasing effect due to the non-anonymity of the questionnaire, participants were given the option on the utility table page of asking that questionnaires only be read by the researcher, with results being reported to course organizers in aggregate form only. Participants were also

² Questionnaire is condensed to fit the page here. Participants were given more space for responses.

informed verbally of this option, and reminded of their right to “opt out” of the non-anonymous questionnaire. None of the participants requested either of these options.

Phase Three: Post-course questionnaire.

The final phase in the field test was redistribution via email of participants’ individual utility tables so that they could rate the extent to which they were able to use learning as anticipated, and report on how transfer of training was affected by factors in their workplace. The questionnaire was administered to participants approximately four months after course completion. Participants were sent the post-course questionnaire by email and, in order to simulate common practice in training institutes for level three questionnaires, they were sent one email reminder only if they did not submit the questionnaire before the deadline.

Below, the full text of this questionnaire is reproduced in Table 5.3:

Table 5.3: Post-Course Evaluation Questionnaire

1. Have you been able to use the knowledge, skills or contacts you gained on the course have been useful to you since returning to your workplace? Please give specific examples of any ways in which the course has contributed to your ability to achieve your work-related goals.

2. At the end of the course, you were asked to share with us the three things you acquired in the course that you thought would be most useful to you back on the job, and how you expected to use these things in your work. For each of the knowledge/skill areas you listed please tell us the extent to which you agree or disagree with the following statements, where 1 = do not agree at all and 5 = fully agree. Choose N/A if the statement is not applicable to the knowledge/skill area you selected.

	Knowledge/skill area 1	Knowledge/skill area 2	Knowledge/skill area 3
Your response in the end-of course questionnaire	<i>Participant's response from the end-of course questionnaire reproduced here.</i>	<i>Participant's response from the end-of course questionnaire reproduced here.</i>	<i>Participant's response from the end-of course questionnaire reproduced here.</i>
I have been able to effectively use the knowledge and skills I acquired.	1 2 3 4 5 N/A	1 2 3 4 5 N/A	1 2 3 4 5 N/A
The knowledge/skills I have acquired have changed the way I think about issues at work.	1 2 3 4 5 N/A	1 2 3 4 5 N/A	1 2 3 4 5 N/A
I frequently use the knowledge/skills I acquired in the course.	1 2 3 4 5 N/A	1 2 3 4 5 N/A	1 2 3 4 5 N/A
I received in the course the practical skills needed to use this knowledge/skill	1 2 3 4 5 N/A	1 2 3 4 5 N/A	1 2 3 4 5 N/A

area in my work.			
I have enough support from my managers to use this knowledge/skill area in my work.	1 2 3 4 5 N/A	1 2 3 4 5 N/A	1 2 3 4 5 N/A
I have enough support from my colleagues to use this knowledge/skill area in my work.	1 2 3 4 5 N/A	1 2 3 4 5 N/A	1 2 3 4 5 N/A
I have the necessary resources/funds /equipment to use this knowledge/skill area in my work.	1 2 3 4 5 N/A	1 2 3 4 5 N/A	1 2 3 4 5 N/A
Comments – please give us specific details on your efforts to apply this knowledge/skill area. We are interested to hear specific examples of how you have tried to apply these areas and of any obstacles that you have encountered to applying them.			

3. Are there areas of learning in which it would have been useful for you to get expert follow-up support from the MCTC in order to help you use this knowledge in your workplace? If so, what support would have been helpful to you and why?

4. Are there other things that the MCTC could do to improve the relevance and usefulness of courses like yours?

Analysis of evaluation questionnaires.

Course organizers were asked for their opinions on the evaluation protocol and input on responses at several junctures during the field test. They were asked to review each of the draft questionnaires before distribution and to suggest improvements or changes to the questionnaire. They read all completed questionnaires, after each of the three phases, and were given averages of quantitative scores. Feedback was elicited from training organizers on field test results at two junctures. After receipt and review of the end-of-course questionnaire responses, each organizer was interviewed individually according to a structured questionnaire that can be found in Appendix H with regard to the utility of the information generated by each question. Amongst other things, course organizers were asked to rate the utility of responses to each question according to the following scale not very/somewhat/yes. Next, after receipt and review of post-course questionnaires, each organizer was once again interviewed, this time in an unstructured interview.

Field Test Analysis

In keeping with Reynold's (1971) guidance on theory-to-research processes, the field test results were analyzed by selecting statements generated by the model for comparison with the field test results. The following section presents each of the selected statements, in question form, and then brings evidence from a review of completed training participant evaluation questionnaires and from interviews with course organizers.

All observations noted below are that of the researcher, other than where specifically cited as opinions of the course organizers. As both utility and feasibility are largely subjective concepts, and course organizers were viewed as the primary consumers of evaluative information, course-organizer self-reports were seen as relevant evidence of the extent to which the evaluation protocol was useful and feasible. That noted, it is important to stress that the intention was not to generalize from the opinions of the two course organizers to other courses, in other DTIs, run by other course organizers. Further research would be necessary in order to check the extent to which the field test findings are applicable to other course evaluations. Instead, the dissertation adopted Lincoln and Guba's (1999: 420) concept of transferability. As noted in Chapter Three, to meet the criterion of transferability, the researcher's responsibility is only to provide a sufficiently detailed description of field test context and conditions, in order to aid future researchers to determine whether the system is relevant for their contexts.

How feasible was the personal benchmark system in terms of cost, time allocation, simplicity and resources?

Questionnaire distribution for the personal benchmark system involved no monetary costs other than that of printing questionnaires. Time feasibility was a more challenging problem, due to the fact that the two latter phases in the questionnaire asked participants to comment on their individual goals, as enumerated in the previous phase. This entailed preparing a separate questionnaire for each participant and, in the case of the post-course phase, sending each questionnaire individually to training participants. This was a time-consuming exercise which is likely to only be feasible in DTIs where there is sufficient support staff to devote the necessary time. Another option for

increasing the time feasibility of the model is administering the third phase of the questionnaire through a mail-out of copies of each participant's individual training use table. Alternately, if the first and second phase questionnaires were also administered electronically, electronic copies of participant's utility goal tables could more easily be forwarded via email to participants for the third phase of the evaluation protocol.

Computerization of the three-phase evaluation protocol could also simplify the process of filling in the questionnaires for respondents. One of the evident shortcomings of the evaluation protocol as administered in the field test is the complexity of the forms, in particular the learning goal and workplace use goal tables. For example, in the pre-course phase of the microfinance course, participants were distributed learning goal tables including four columns: two to be filled out pre-course (learning goals and relevance of these goals to work) and two to be left empty until after the course (rating achievement of learning goals and comments). Despite detailed written and verbal instructions by the researcher, and despite the fact that the columns to be filled in at the end of the course were shaded grey to emphasize the difference, one of the respondents filled in the grey columns at the beginning of the course. In order to simplify instructions, in the gender course participants were only distributed the first two columns at the beginning of the course and then the two ratings columns to be filled in at the end of the course were glued in later by the researcher. This was time-consuming. Computerization of the evaluation protocol could simplify this process and make administration of the evaluation model more feasible.

Did course organizers believe that information generated by the evaluation was sufficiently useful to merit allocation of time to reading participant responses?

The two course organizers differed in their opinions of the open-ended questions on country context and needs asked in the pre-course and end-of-course questionnaires. One of the course organizers was of the opinion that these questions were not useful enough to justify the time spent reading responses. The responses to these questions in the pre-course questionnaire, in her opinion, were very general and mostly reported information that she had known in advance. As for the post-course questionnaire, which asked a similar question on country needs in order to measure shifts in attitude during the course, the course organizer observed that she was able to see from comparing pre-course and end-of course responses a shift in the importance accorded by course participants to certain issues, in keeping with course content. However, she found the comparison to be only somewhat useful in relation to the time spent reading responses.

The organizer of the other course was of the opinion that had she received the information in the pre-course questionnaires one to two months prior to the course, they would have provided her with “excellent” information that would have enabled her to better adapt the course focus and contents to trainee needs. However, as the questionnaire was administered on the first day of course, at a time when few changes could be made to the program, it was only “somewhat” useful. As to the comparison between pre-course and post-course responses, the organizer of the other course found the information on shifts of attitude to be sufficiently useful to justify the time spent reading responses. In her words, “They changed their responses. They saw things here that they hadn’t thought before and wrote about them. At first, they wrote very broad categories of what doesn’t work and at the end they wrote specific things that need to be done. I learned that the

problems and needs are very similar in many places... I learned that they had learned in the course tools to deal with the problems.”

Both course organizers reported that the learning goal table and the training use table were sufficiently useful to merit time spent reviewing responses. Specifically, organizers of both courses found that the learning goal table provided “excellent” information on the extent to which participant expectations were in line with course content and contributed to the organizers’ ability to focus training content more on expectations. However, one noted the difficulty in finding time to read participant responses to the pre-course questionnaire at the beginning of the course. Both would have preferred the pre-course questionnaire to be distributed and collected one or two months prior to the course, both to make reading the forms more time-feasible and to facilitate adaptation of course contents on the basis of participant needs and preferences.

Both organizers agreed that participant end-of-course ratings of their individual learning goals accorded them a more nuanced, and thus more useful, understanding of what participants got out of the course than they would have gotten through standard level one questionnaires, which justified the time spent reading the responses. In the words of one of the organizers, “The information was excellent and the comments good. There was much more specific information than I would have had otherwise.” Both course organizers agreed that the fourth column of the learning goal table – an open-ended column asking for comments on achievement of learning goals should be optional, for those participants who desired to add written comments, rather than mandatory for all respondents. One noted that it would have been unnecessarily time-consuming to read all comments had everyone been asked to provide an open-ended explanation of ratings, and

the other course organizer responded that it was more efficient to leave the “comments” column as optional, because, in her words, “I don’t have to hear from people who have nothing to say.”

In addition, both course organizers were enthusiastic about the depth of information on the different ways that course participants had been able to apply learning at work and felt that the post-course questionnaire gave them a good understanding of how implementation was influenced by factors in the workplace context such as availability of resources, appropriate policies and managerial support.

In sum, both course organizers found the learning goal and training use tables to be an improvement over their standard evaluation regimes, providing sufficiently useful information to justify time spent reading participant responses. In particular, they were of the opinion that this method of results evaluation provided much more specific, nuanced information on participant needs and expectations from the course, and on learning and workplace transfer of training than the other forms of training results evaluation with which they were familiar. With regard to the open-ended questions on country context and needs, course organizers differed in the extent to which they found that these questions were sufficiently useful to justify time spent reading them. Finally, both course organizers believed that it would have been more useful to administer the pre-course part of the questionnaire before the course started. This could be feasibly accomplished by making the pre-course questionnaire part of the course application process.

To what extent did participants write detailed responses to open-ended questions, reflecting their investment of significant time and thought in the evaluation process?

Respondents provided lengthy, very detailed information in open-ended questions at all stages of the evaluation. For example, in the post-course questionnaire, all eleven of the participants in the Microfinance course provided specific details on how they had used learning at work. In the case of the ten post-course questionnaire respondents from the Gender course, eight provided specific examples of how they had used learning at work, one stated that she had not been able to use learning and explained why and only one did not give a detailed response to the question. Responses to other open-ended questions in all three questionnaires administered in the context of the dissertation field test were similarly detailed and specific. In order to give some indication of how detailed participant responses were, Table 5.4 provides a sample of participant feedback from the post-course evaluation forms. The research put all the post-course questionnaires in a pile and then selected every fifth questionnaire for inclusion in the below quotes in order to ensure that the sample was fully random. From each of the selected questionnaires, only the open-ended responses which provided substantive information on what learning was applied and how training could be improved were included. However, all such substantive information was reproduced in full in order to give a clear picture of the length and depth of responses. Identifying information was deleted to protect confidentiality.

**Table 5.4 Sample of Open-Ended Responses
in Post-Course Questionnaires**

Microfinance course:

Respondent 1:

“After the training course at Israel, a lot of changes were made inside my enterprise, following the knowledge acquired. In the areas of client management, financial management, advertising and marketing tasks and motivation of the employees have let the company to work better and to have better results than before Israel. Even the profit of the company has increased in almost 40% after the course.”

“It was hard for the enterprise to find motivation issues non-economical. Now we have the opportunity to share with the employees some chances like training, schedule flexibility among other issues improving the performance of each one.”

“The way of making advertisement has fully changed. Now we look for the client rapprochement through the clear message in our advertisement”

“The financial issues have improved the process we were following. Nowadays we manage these items more carefully and we register everything”

“ I personally think that the exercise of the business lab is the most useful because is more practical, so this activity should last more time and have more activities inside, in order to learn more from it, especially in the financial issues.”

“Me as an entrepreneur, personally think that the course should be divided in entrepreneurs and support systems workers. It is good to share experiences between these two kinds of people, but the process of support systems organizations is not a theme to take a lot of time for an entrepreneur but for a support system worker.”

Respondent #6

“I have been able to transmit the way the support systems in Israel works in the manufacturing society. I haven’t used all the knowledge acquired through except for the marketing course and the self-efficacy reading.... These courses have helped me be more accurate in my projects and in my actions.”

“The course in my opinion was a great introduction for most people who work at government dependencies or any other support organizations, but in my case which I am an entrepreneur I found interesting yet not that useful. I saw lots of opportunities and most of them were not taught in classroom. All support systems we say, it surely gave me ideas of how to implement them here in my country but sure they are for the long term.”

Table 5.4: Con't

“So in conclusion, the experience of being in Israel and travel and also have teachers who can explain to me at least a bit of how businesses are run in Israel and how the support systems works gave me some ideas but nothing solid. I would recommend more practical courses, with real case examples. You have lots of very interesting companies created and what we are most curious (I think) is how the inspiration and motivation and how the manage to pull up their business. I think the opportunity of being able to go somewhere else its mind and eye opening opportunity, so one of the things I would have loved to learn about is of the greenhouses and its technologies so I could be able to import, develop or apply something similar in my country.”

Respondent # 11

“The knowledge gained during the course has assisted in the sense that the Government of [my country] is about to implement a [program name] through the Ministry of X where I work. Main activities to be undertaken under this fund are training in vocational and technical skills as well as provision of business advisory services to entrepreneurs. This being the case the Ministry in collaboration with other stakeholders is currently developing modalities on how best to deliver support services. I shared my knowledge of the operations of the MATIs [Israeli Small Business Development Centers] there in Israel and people feel it’s a good model which could be modified to fit our situation.”

“The Government of [my country] established an institution called X Enterprise Development Institute to promote the development of small and medium enterprises. The institution has embarked on a project to establish Business development Centers similar to the MATIs in Israel. I was privileged to be invited to their planning meetings to enhance the operations of these centers. I shared with them some of the activities that these centers could be doing borrowing from the operations of the MATIs.”

“What I could have appreciated is a detailed training in specific areas as Business Proposal Writing/business plan, Business analysis, pricing, costing, cash flow analysis, basic business record keeping, financial management or general management of a business. I would seek further support in this because the Ministry requires master trainers in this area if the Youth Enterprise Development fund is to be successful.”

Table 5.4: Con't

Gender Course:

Respondent #16

After returning me have appointed on the position of the manager of the `X ` project, in whom I undoubtedly use skills which has received at you: easy communication with people, facilitation of the processes, planning trainings and others...and of course, important thing for me is the new contacts with new friends and colleagues from other countries of the world. “

“One of my responsibilities is the planning training courses for interweavers. After returning from Israel with new knowledge and new useful skills this process is easily for me.”

“This book [UNHABITAT sourcebook] contains much useful information and examples for trainers and I use them very often during my recent work:-situation tasks, tactical scheme, Some Examples of Capacity-Building Workshops, What Qualities are Necessary for Good Municipal Leadership, and etceteras”

“The decision of gender questions remains till now a mental problem in our country. Us, women respect, listen, but do not perceive seriously. Gender questions meanwhile interest women and female organization. That’s way we need the time for make really serious decision concerning, for ex. quotas ...and so on”

“Situation in [my country], slightly another, than in this country. It would be good to study examples from the were former Soviet Union`s countries.”

I often use which handouts you prepared for us, CD and I must mention than source book is book for reference for me.

“As my opinion, It would be better, if in the future you have more mixed group than there was we with four men☺. Gender issues must be well – known for both: for men and women. Also, I wish that part for mass media will be longer. One day for media communications really is not enough.”

Respondent #21

a. Since attending the course I have used to knowledge gained to do quite much at my work. One of the things I did on getting back was to have a facilitated discussion with colleagues who were much excited to get the notes and fresh information on gender governance.

b. I have also shared this with two local authorities and put a proposal to have a training session with elected counselors and chief officers in areas of gender mainstreaming following on agreed quarters nationally.

Table 5.4: Con't

c. I have facilitated at 2 meetings of NGOs that take forward Gender Equality and their main agenda was to find ways of lobbying the government and local authorities on implementing the gender quarters i.e. at least one third in all levels of employment. I have been elected to the Board of one of the NGOs to help them in articulating their gender work – I feel privileged and hope to do my best to help.

d. In November, our organization through the community programs I managed, worked with other organizations to observe the 16 days of campaign – saying no to violence against women. The lobby meetings took many dimensions and were well taken. It was something with a difference this time around and everyone felt we reached many people and hope the message too.

e. I have just submitted a sports initiative to my organization that will run parallel to a community citizens' programme that I manage. This initiative is meant to work with young people to embed gender equality from the grassroots. This community happens to have challenges in this area as it is male dominated and women and girls carry the burden of most of the work that goes into keeping the family together. I am excited about this initiative as it has attracted interest from agencies that could just help in keeping it sustainable and hopefully duplicated in other communities. I am so excited at this prospect.

f. I am in charge of Equal Opportunity and Diversity in my organization, and this learning has enabled me to come up with new experiences and areas of work and helped in my scorecard. I have many more examples of achievements this year to report on.”

“I have used this knowledge with colleagues at work and have added value to training content that we are using and this is encouraging because it shows me this information is well received.”

“It is my passion to work in the gender equality area and I have had the chance to use these skills in the field as I am in charge of community capacity building and I use much of the skills in this area.”

“I have just developed a new proposal for a sports initiative to help in getting gender mainstreaming and accepted from the grassroots. This has been well received by my organization and an organization has expressed the interest to work jointly with me and fund part of the way.”

The detailed nature of responses stands in sharp contrast to the dissertation author's experience with a level three questionnaire distributed the 2008 World Bank Independent Evaluation Group study of World Bank-financed training results. As detailed in Chapter Two, questionnaires were administered in face-to-face interviews with 548 former training participants from six developing countries worldwide six months to two years after course completion. Interviewers asked participants to rate the extent to which the World Bank-financed course which they had attended had changed the way they do their work and then probed for details on the way their work had changed. Only one out of six of respondents who stated that training had led to changes in the way they did their work gave specific examples when probed (IEG, 2008: 15).

To what extent did participant ratings of individual goals give course organizers a richer picture of course results than is normally obtained through the standardized questionnaires generally used in MCTC courses?

Both course organizers found the system of having participants select individual learning and training use goals to give them a substantially richer picture of participant needs and expectations and of course results. According to the course organizers, selection of individual goals enabled them to understand the areas of learning and use that were most important to each individual trainee, making it possible for them to see, *inter alia*, how differences in country contexts and type of work affected the needs and expectations of training participants, and the benefits accrued to individual participants through training.

Variation between ratings was also identified by the researcher as a partial indicator of the thoughtfulness of participants' responses, and of the utility of information gathered (Fowler, 1995). Where participants gave different ratings to each of their learning and training use goals, it was possible to see which topics participants found to be particularly informative, and which acquired knowledge/skills were most useful to them on the job. Variation between each participant's ratings of achievement of their individual use or learning goals was seen as one indication of response credibility. Where participants awarded lower ratings to some goals than to others, this was viewed as evidence that participants had answered the questionnaire at least somewhat thoughtfully, and had not given uniformly high marks as from gratitude or a desire to secure further training funding. Conversely, while variation in ratings given was viewed as partial evidence of credibility of responses, lack of variation could not be seen as evidence of lack of credibility of ratings, as the lack of variation might be a reflection of participants' belief that all goals were achieved to the same extent.

In the end-of course questionnaire on learning goals, 9 of the 16 respondents in the Microfinance course who listed more than one learning goal awarded different ratings for their goals, of which one third had gaps of more than one rating point (for example, one goal rated 4 and another 2). In the Gender course, of the 17 respondents that listed more than one learning goal, 13 did not give the same score to each of the goals listed, of which two varied by at least two ratings points. Similarly, in the post-course questionnaire, when asked whether they were able to effectively use knowledge and skills they acquired, seven of the eleven respondents from the Microfinance course rated use of different individual goals differently, in five cases with a gap of two or more ratings

points between individual goals. Eight of the ten Gender course respondents varied ratings of use goals by at least one point, and three of these by at least two points.

How were responses influenced by their non-anonymity of questionnaires?

One of the concerns in administering the three-phase evaluation protocol was the effect that the non-anonymity of questionnaires would have on the willingness of respondents to answer questions frankly and honestly. While there is no way to gauge the actual impact of non-anonymity, the following observations were made by the researcher: Participants were given the option in the end of course and post-course questionnaires of asking that their responses not be shared with course staff in their name. Of all respondents, only one chose this option. Moreover, all respondents were reminded at the beginning of each phase of the evaluation that their responses could not be submitted anonymously, and were informed that they had the right to opt out of participating in the three-phase evaluation field test, if they were not comfortable with the non-anonymous element of the evaluation. None chose to opt out on this basis.

In addition, the anonymous, computerized level one questionnaire administered at the end of the Microfinance course in addition to the personal benchmark questionnaire revealed similar scores. While the rating scale was different in the anonymous questionnaire than in the non-anonymous field test questionnaire (a five point scale rather than a four point scale), and thus ratings were not directly comparable, both yielded similar results. In other words, higher ratings were not awarded by participants to the course in the non-anonymous questionnaires than in the anonymous ones. Finally, all respondents offered criticism and suggestions for improvement in the end of course and

post-course questionnaires, suggesting that they were comfortable providing criticism even in a non-anonymous format. That being said, it is undeniable that non-anonymity can potentially reduce the credibility of responses. One way to mitigate this risk could be to administer the entire evaluation protocol by computer, giving each participant a numeric login code to access previous phases of the evaluation. However, this is only a partial solution as the open-ended elements of the evaluation could be used relatively easily in many cases to identify participants.

To what extent did organizers of the two field-study courses use the information generated by the pre-course assessments to help shape training content?

As noted in the analysis of question two, course organizers found pre-course information to be only somewhat useful in shaping course content. This was primarily because pre-course questionnaires were administered on the first day of course when they had limited time to review responses, and when it was too late to make significant modifications to course content. Both organizers were of the opinion, however, that had pre-course questionnaires been administered as part of the application process, the information received would have helped inform them of the specific needs and expectations of course participants which could have been used to shape course content.

Where course content did not cover subject matters listed by participants among their individual learning goals, is this reflected in participant ratings of achievement of learning goals by a low participating rating or selection of the “not applicable” option?

Course organizers were asked to review participant ratings of their individual learning goals to see whether in each case where the learning goal referred to material not covered in the course, the participants either chose the lowest rating option offered (course met learning needs "not at all") or opted for "not applicable". Both course organizers reported that in all cases where material was not covered in the course, participants chose "not applicable" or "not at all". In total, two ratings of "not applicable" were given by participants in the Microenterprises course regarding fulfillment of individual learning goals and two ratings of one ("The course did not fulfill my needs at all."). In the case of the Gender course, one participant chose the "not at all" option for one of his or her learning goals. According to the course organizers, low scores correlated with material not covered in the course and there were no cases in which participants gave high ratings on topics not covered in the course. Overall, the low number of "not available" or "not at all" ratings may be attributable to the fact that pre-course questionnaires were administered immediately after course organizers reviewed the course program with participants, so that their expectations were in line with course content. Had the pre-course questionnaire been distributed as part of the course application process, it is possible that a greater number of individual learning goals cited by participants would have been not relevant to course content. The correspondence of non-applicable ratings with topics not taught in the curriculum, was viewed as a partial indication of the credibility of responses.

Were response rates for the post-course evaluation questionnaires substantially higher than those of level three questionnaires administered to former MCTC course participants in the past?

Response rates for post-course questionnaires for both courses were above 50% - significantly higher than response rates to previous post-course electronic questionnaires that had been administered to former MCTC training participants, which have in the past been approximately 15% to 20%. In the Microfinance course, of the 18 training participants who had completed the end-of course questionnaire, it was possible to send questionnaires to 15. The email addresses of the other three candidates were incorrect. Of the 15 with valid email addresses, 11 responded, for a total response rate of 73%. In order to best simulate common practice in course evaluations, the researcher only sent out one reminder to participants who did not respond to the first request to fill out the post course questionnaire, and did not make follow-up telephone calls to participants to verify their email addresses.

The Gender course had a lower response rate than the Microfinance course, although still above 50%. Specifically, of the 20 people who filled out the end of course questionnaire, only one did not receive a request to fill out a post-questionnaire due to lack of a correct email address. Of the 19 who were sent out the post-course questionnaire, 10 responded, for a response rate of 53%. The discrepancy between response rates of the two courses may be attributable to the fact that the researcher was present in most of the sessions for the Microfinance course, but only in the sessions where questionnaires were distributed to participants in the Gender course. This may have fostered greater trust in and goodwill toward the researcher which may have, in turn, affected response rates.

To what extent did post-course evaluations provide a rich picture of how participant transfer of training was affected by their workplace environment?

In total, in participants' awarded a low score (1-3) to 15 individual training use goals in response to the question "I have been able to effectively use the knowledge and skills I acquired". Of these 15 low scores for individual training use goals, in 13 cases participants cited specific obstacles to implementation of learning. In all of these 13 cases it was possible to understand the primary obstacles to implementation through low scores (1-3) given to close-ended questions on specific obstacles (specifically: adequacy of skills received in the course, support of managers, support of colleagues and availability of sufficient funds/resources/equipment). In addition, in eight of these 13 circumstances, participants described in open-ended responses the specific obstacles they were facing to implementation of training. Some of these obstacles are evident in the examples of participant responses given above.

Summary

This chapter presented the conduct and the analysis of the field test. It found that the personal benchmark system had certain demonstrable advantages over standard training evaluation methods, as outlined in previous chapters. Advantages included the depth and length of participant open-ended responses, higher response rates, information on how workplace training obstacles affected training transfer of specific use goals, and the high degree of variation between participant ratings of various learning and training use goals. In addition, the field test identified several areas for further refinement of the model, including computerization of the forms and distribution of the pre-course questionnaire as part of the course application process, in order to provide the opportunity

to supervisors to give input and to course organizers to take into account participant needs in course organization and participant selection. The next section will now review these findings in the overall context of the dissertation research, linking them to the evaluation needs identified in the DTI interviews and the literature review, and analyzing them according to the three quality criteria of utility, feasibility and credibility identified in Chapter Two.

Chapter Six: Discussion and Conclusions

Introduction

This chapter will discuss the findings of the research, analyzing the extent to which the personal benchmark evaluation system successfully addresses the training evaluation needs and challenges identified in the literature review and DTI interviews, and identifying areas for further research. As noted in Chapter Two, the personal benchmark evaluation system was designed to improve the utility, feasibility and credibility of training evaluation through use of participatory methods whereby the training participant was asked to set his or her own learning and training use goals and then to rate achievement of those goals accordingly. The following analysis will revisit these three quality criteria. It will look at each of the evaluation shortcomings identified in the literature review and DTI interviews and listed Tables 2.1 and 4.1, analyzing the extent to which the personal benchmark system successfully addressed the shortcomings of standard practice. In addition, recommendations will be made for potential improvements to the model and further research.

Utility

As presented in Table 4.1, the following evaluation utility needs had been identified in the literature review and DTI interviews:

1. Improved needs assessment as the basis for program planning, based on greater beneficiary ownership of the assessment process (Brown and Gerhardt, 2002; Brinkerhoff, 2006).
2. Greater measurement of workplace environment factors contributing to training success (Brown and Gerhardt 2002; Brown and Reed 2002; Holton 1996; and Brinkerhoff 1988, 2006).
3. Increased variation in question responses on training quality in order to provide richer information on training results (Fowler 1995).
4. Reinforced partnership between management, trainers and training participants in order to ensure that actual organizational needs are addressed and to improve training results by ensuring that participants receive support from their managers for implementation of learning (Brinkerhoff 2006).

Improved needs assessment.

The pre-course assessment of the personal benchmark system provided detailed information on participants' individual learning goals and how participants intended to use learning in their workplaces. However, the utility of this information for course organizers was limited due to the timing of questionnaire administration. As noted above, in order to enable course organizers to take into account individual needs and preferences of participants in the course program, course organizers would have preferred to receive information several weeks before the course started. This could have been feasibly done if the first questionnaire was distributed as part of the course application process rather than on the first day of the course. A second advantage to early

administration of the questionnaires is that they could be used by training institutes to aid in participant selection.

One possible drawback to early administration of pre-course questionnaires is that individual course goals as formulated by participants might include many more goals that are not relevant to course content. In the field test, the first questionnaire was administered immediately following course organizer presentations of course content in order to ensure that participant individual learning goal selections would be relevant to course content. Further testing of the model, with early administration of the first questionnaire, would be necessary in order to determine how goal formulation would be influenced by pre-course administration of questionnaires. In addition, pre-course administration of the first questionnaire may not be feasible in some courses where participants are nominated by dispatching organizations and there is no pre-course application process.

Greater measurement of workplace environment factors.

As noted in the field test analysis above, the post-course questionnaire provided rich information on how workplace environment factors affect learning. As Foxon (1993) points out, this information can be used by training institutes to provide classroom-based support for problems encountered in implementation (such as including change management workshops in courses aimed at introducing significant change); modify participant selection strategies (such as training a critical mass of persons from each organization where lack of support of colleagues has been an obstacle to implementation, or only accepting to the course participants whose workplaces have

committed to making available the necessary resources for implementation); or partner with other donors to provide associated capacity support such as equipment purchase, technical assistance or policy support.

Greater variation on quantitative scores.

Course organizers were able to use quantitative scores on end-of-course and post-course questionnaires in order to get a better sense of the extent to which various course modules were useful to participants. However, in order to fully take advantage of this measure, it would be helpful to better quantify this variation. The field test generated average scores for various questions by first averaging between participant responses for each question and then calculating an average score for all participants for the question. In this manner, an average score for course learning was calculated from the end-of-course questionnaire and for course utility from the post-course questionnaire. Another possibility for quantitative analysis of scores would be to generate a separate quantitative score for each topic area. This would necessitate asking course organizers to review responses and sort them in a matrix according to course topic categories. In this way, average scores could be calculated for each course topic. In addition, by simply counting how many participants cited various training use goals in the post-course questionnaire, it would be possible to learn how successful the course was at achieving various learning aims. For example, if many course participants cited one particular course aspect as being particularly useful, and none cited another aspect of the course, this information could be used to determine which course units are most useful. This indirect method of measuring course utility might provide more credible information on the utility of various course units than closed-ended questions which list various course units and ask

participants to rate their utility. Further field testing would be needed, however, to establish whether this is actually the case.

Increased partnership between course organizers, course participants and their workplace managers.

As cited in the literature review, there is much evidence indicating that courses can be made more relevant and effective by better involving course participants and their managers in both planning and evaluation of training (Brinkerhoff and Apking 2006; Brinkerhoff and Montesino 1996; Dubin et al 1974; Georgenson, 1982; Gill 1989; Hamblin 1974; Tannenbaum 1997). The personal benchmark system tries to enhance participant feelings of ownership of the needs assessment and evaluation process by engaging course participants in the formulation of course learning and workplace use aims. However, the model, as field tested, did not involve the managers of training participants. In future applications of the system, managers could be included in the system by administering the first questionnaire as part of the application process, rather than on the first day of the course, and asking participants' superiors to provide input on learning goals. In addition, in the post-course questionnaire, supervisors could be asked to rate the extent to which participants succeeded in using material acquired or to provide their own comments on the course's workplace outcomes. Further research is necessary, however, to determine how feasible it would be to ask for supervisor input and how this might affect the credibility of participant responses. Specifically, there is no way of determining whether supervisors would actually be the ones to complete questionnaires and it is uncertain whether supervisors would be willing to devote time to meaningfully engage in the needs assessment and evaluation process. Moreover, participants might

be more reluctant to provide candid answers to questions on training use if their supervisors would be reading responses.

Feasibility

As noted in Table 4.1, the literature review and DTI interviews identified the need for more feasible methods of evaluation, particularly on course outcomes (level three). In particular, DTIs cited the issue of low response rates and lack of varied, in-depth information generated by level three questionnaires distributed electronically, on the one hand, and the cost infeasibility of any distribution method other than e-survey, on the other. Thus, the personal benchmark system endeavored to use participatory methods in order to raise evaluation response rates and the richness of data collected while maintaining the cost-feasible method of electronic questionnaires for level three data collection.

The personal benchmark system as field tested did successfully and substantially raise response rates from former level three questionnaires administered by the same training institute and generated particularly in-depth participant responses on application of training in the workplace. However, further field testing would be necessary in order to both ensure that this rise in response rates is attributable to the participatory nature of three-phase evaluation method and would be replicable in other contexts.

The variation between response rates of the two field tested courses suggests that factors other than just the three phase evaluation method might have influenced response rates. As noted above, substantially higher response rates on the Microfinance course

(73% as compared to 53% in the Gender course) could have been at least partly attributable to the fact that the researcher attended part of the sessions of the Microfinance course, whereas in the case of the Gender course, she only was present in the course in order to administer the pre-course and end-of-course questionnaires. This possibility may be in keeping with the observation made by several of the DTI representatives interviewed that participants are generally more willing to engage in an evaluative process post-course where they have maintained some form of contact with DTIs or individual trainers post-course. Where participants feel a greater personal connection with trainers or training institutes, they may be more willing to provide feedback. In addition, response rates might have been influenced by the fact that participants were informed in advance, during the course itself, that they would be receive a post-course evaluation several months after course completion, or by the fact that these questionnaires were distributed less than four months after course completion. In previous post-course surveys done by the MCTC, over a year had elapsed between course completion and distribution of level three questionnaires and participants had not been informed at courses that they would receive these questionnaires. Further research is needed to determine the factors contributing to higher response rates and to determine replicability in other contexts.

Credibility

As noted in Chapter Three, the personal benchmark evaluation system only endeavored to measure participant perceptions of training results, and not to empirically

measure actual learning and participant behavior change. As a result, accuracy was replaced by Lincoln and Guba's category of credibility (Fowler, 1995; Lincoln & Guba, 1989). In doing so, the author acknowledges that participant perceptions are not necessarily an accurate measure of actual results. However, the personal benchmark system is intended to address the fact that most training institutes do not have adequate resources to do more empirical methods of testing such as pre- and post- tests of learning, observation of participant workplace performance pre- and post-course or measurement of quantitative indicators of performance. As a result, credible and detailed representation of participant perceptions was chosen as a less desirable but far more feasible aim for training evaluation.

In keeping with this logic, table 4.1 documented the need for more use of post-course questionnaires in order to move away from the common practice of interpreting end-of course (level one) questions on training usefulness of training as a proxy indicator of whether training actually was useful to trainees upon return to their workplaces. The personal benchmark system, by mandating post-course evaluations for all courses, improves DTI access to information on actual training use, as perceived by respondents. In addition, the personal benchmark system was successful in eliciting detailed descriptions of training use.

Collecting input from participant supervisors might further enhance understanding of course results. The extent to which this would provide a different perspective from that of participants, and thus enhance credibility of information gathered, however, would need to be tested. It is also difficult to establish empirically how credibly participant responses reflected their actual attitudes toward course results. In particular, due to the

non-anonymity of responses, and the fact that most course participants benefitted from a free trip to Israel, it is possible that participants might have been reluctant to fully express criticisms of the course.

Three indicators would suggest, however, that responses were at least somewhat credible. First, the variation in participant ratings on learning and training use indicates that participants did give low scores to individual goals that were less fully achieved in the course. In addition, the fact that all participants who had formulated learning goals that were not actually covered in the course had rated these goals as not applicable suggests that participants chose ratings in a thoughtful manner, that was reflective of their actual experiences. Third, the three-phase evaluation method lends itself to a number of indirect methods for measuring course achievements, particularly in the post-course questionnaire. For example, as noted above, by counting the number of participants who cited various course subjects as having been particularly useful to them, it may be possible to develop an understanding of course strengths and weaknesses. Similarly, even where high ratings were given to course utility, where they were accompanied by low ratings on environmental factors affecting training transfer such as manager support or available resources for implementation of training, it could be possible to learn about the application and, hence, usefulness of training in the workplace. Finally, the heavy reliance on open-ended responses and the rich content of these responses in the field test, enhances the depth of data on participant's experience of training.

Contributions

This dissertation, through both the literature review and the DTI interviews, has contributed to the literature a better understanding both of shortcomings in presently used methods of training evaluation and DTI needs for improved evaluative data. The personal benchmark system provides possible solutions to many of the identified needs and shortcomings, and as such has contributed to the range of evaluation methods available to DTIs. In that manner, it facilitates and supports good capacity development practice in international development contexts. Chapter One's review of studies on capacity development practice noted five significant shortcomings in present practice:

1. Lack of sufficient needs assessment.
2. Lack of integrated, long-term, multi-level interventions addressing the full range of human, organizational and capacity constraints affecting achievement of organizational goals.
3. Need for greater beneficiary ownership of capacity development interventions.
4. Need for more long-term interventions.
5. Need for better and more consistent monitoring and evaluation of programs.

The personal benchmark system presents a feasible method for both increasing the quality and prevalence of needs assessments and improving evaluation of program outcomes. It does so through participatory methods that enhance beneficiary ownership of training by asking participants to define their learning goals and choose the learning and training use criteria by which training is evaluated and by adding a quantitative dimension to participants' qualitative descriptions of learning and training use. In

addition, the model both mandates and increases the utility and feasibility of post course evaluations for all courses.

Post-course evaluations are designed to generate detailed information on how participants' workplace contexts affect implementation of training. In doing so, these post-course evaluations can contribute to an understanding of where organizational capacity gaps exist that may hinder the transfer of learning to workplace practice. As noted above, where post-course evaluations identify common obstacles to implementation, it may be possible to adjust learning programs to address those obstacles. Finally, where it is not possible for training institutes to address obstacles to training transfer through their programs or participant selection, post-course evaluations can minimally provide an understanding of what complementary support is needed to enable training transfer. In this way, the system may contribute to an understanding among donors of the importance of integrating training within multi-faceted, longer term capacity development programs which can address capacity issues through a variety of means. In this way, greater use of post-course evaluation questionnaires that emphasize measurement of organizational context can be used, in the long term, to help shift attitudes towards the efficacy of stand-alone training, building awareness of the need for different, more complex, approaches toward capacity development. Table 6.1 now summarizes the training evaluation needs which were first presented in Tables 2.1 and 4.1, indicating the extent to which the personal benchmark system addresses each of these needs.

Table 6.1 Does Personal benchmark system Address Identified Needs?

Identified Need	Theorized Solution	Personal benchmark system
Utility		
<p>Improved and more participatory participant/target organization/country needs assessment as the basis for program planning.</p>	<p>1. Inclusion of a developmental evaluation phase. 2. Increased beneficiary ownership of needs assessment. Use of participatory needs assessment and evaluation methods where clients, supervisors set goals. This can also help improve training results through increasing participant “intentionality”.</p>	<p>Mandates 100% use of needs assessment based on consultation with participants. Field tested model could be improved by administering needs assessment as part of application process, including supervisor comments in needs assessment.</p>
<p>Improved level 3 evaluations, including greater measurement of workplace environment factors contributing to training success, because workplace environmental factors account for the majority of the variance on training success. DTIs presently rarely administer level three evaluations.</p>	<p>Use of post-course evaluation forms to identify constraints to application of learning in the workplace. This information can then be used to find ways to better support trainees in implementation of learning through training on related issues like change management and other accompanying interventions in the workplace.</p>	<p>Increases feasibility of 100% administration of level three evaluations. Level three evaluations generate information on how workplace context affects training transfer.</p>
<p>Evaluation questions that provide increased variation in responses on training quality in order to provide richer information on training results. Training institutes routinely report that virtually all of their courses are rated highly by participants.</p>	<p>Use of goal-free evaluation methods to identify a range of learning and training use goals and achievements, providing ratings for each.</p>	<p>Questionnaires generated substantial variation on ratings. Use of this variation to understand training impact can be further exploited by developing analytical methods based on coding responses and generating quantitative measures of</p>

		different types of goals.
Reinforcing partnership between management, trainers and trainees in order to ensure that actual organizational needs are addressed and to improve training results by ensuring that participants receive support from their managers for implementation of learning.	Use of participatory needs assessment and evaluation methods where clients, together with their supervisors set personal training goals and report on their achievement.	Model as field tested did not increase supervisor/manager input. Model may be improved by soliciting supervisor input on learning goals in the needs assessment phase and, possibly, also in the post-course evaluation phase.
Information on quality of training processes is an important indicator of DTI performance. DTIs generally satisfied with the quality of the process information gathered from end of course questionnaires and other in-course satisfaction questionnaires.	Continued use of level ones to gather information on the quality of training processes.	Personal benchmark system did not gather process information, but was used alongside the MCTC's standard level-one questionnaires, which, unlike the personal benchmark system's post-course questionnaires, were administered anonymously.
	Feasibility	
To be feasible, evaluations should be inexpensive, relatively non- time-consuming evaluation methodologies that take into account training providers' lack of time and resources to evaluate training's workplace outcomes and impact.	To be feasible, evaluation should be done primarily through written questionnaires.	Personal benchmark system administered entirely through inexpensive written questionnaires. Computerization could improve time feasibility of the model.
Improved response rates needed.	Participatory methods may be used to raise response rates.	Personal benchmark system significantly raised response rates.
	Credibility	
Better predictors of participant use of learning in the workplace needed. Level one and level two evaluations are inaccurate predictors. Level three (behaviour change)	More focus on level three evaluations. Use of participatory methods of training evaluation to increase thoughtfulness of participant responses and "ownership" of evaluation	Field test of personal benchmark system provides rich information on use of learning in workplace. However, further testing would be needed to determine the extent to

questionnaires measure attitudes with regard to utility rather than actual utility, participant self-reports often inaccurate.	process.	which this information provides an accurate picture of actual participant use.
--	----------	--

Thus, the personal benchmark system may improve the utility, feasibility and credibility of training evaluation. However, adoption of the personal benchmark system necessitates the willingness on the part of DTIs to somewhat shift their perceptions of the nature of evaluation. Without this shift, it is unlikely that DTIs will adopt the method or use it properly. Specifically, only training institutes which accept the following principles are likely to embrace the personal benchmark system:

Importance of qualitative information.

The personal benchmark system bases itself heavily on qualitative information. Certain quantitative measures have been suggested above to enhance analysis of course results, including calculating separate quantitative scores for different categories of learning or training use goals cited. However, an important contribution of the method to the learning function of evaluation is richness of information collected through open-ended questions. The model was particularly successful in eliciting in-depth responses. In order to benefit from the data generated through these open-ended responses, however, they must be read. Training institutes that are interested primarily in basic accountability and not in learning from their evaluations, or who downgrade the value of qualitative information, may not be willing to devote the greater time investment necessary in order to read in-depth responses, particularly in courses with a large number of participants.

Importance of course organizer role in evaluation. As noted in the DTI interviews, some DTIs rely on separate evaluation units or independent evaluators to administer evaluations, without involving course organizers. The personal benchmark system, however, assumes that course organizers are active participants in the evaluation process. The heavy reliance on open-ended responses necessitates that readers of evaluation questionnaires have in-depth understanding of course content. This is particularly true to the extent that DTIs wish to develop quantitative scores by course topic by using a matrix to sort individual goals into topic categories. In many cases, only course organizers are likely to have sufficient understanding of course content to be able to complete the matrix. This, in turn, necessitates that the DTIs ensure that course organizer workload leaves time for their engagement in evaluation.

Importance of adopting all three phases of model. Standard practice in DTIs today is to administer end-of-course questionnaires to all course participants, but to administer needs assessment questionnaires and post-course evaluations only selectively if at all. In contrast, the personal benchmark system necessitates that all three phases be administered in all courses, as each phase builds on the last. None of the DTIs interviewed do so today. Training institutes would thus have to commit to changes in practice to use the system. The field test suggested that personal benchmark system can be applied at minimal cost and with reasonable effort, if all three questionnaires are administered electronically. However, as mentioned above, time investment is needed to read and process evaluation results. Training institutes who view needs assessment and post-course questionnaires as necessary only in a representative sample of course may not

be willing to change their practice, regardless of the benefits documented in the course of this dissertation.

Implications for Further Research

The personal benchmark system of training evaluation is presented here as an exercise in “theory to research” theory building. In other words, a theoretical model of training evaluation was developed on the basis of a literature review and DTI interviews, after which areas for further refinement of the model were identified through two field tests. The field tests, as noted, were not intended to validate the system or provide generalizable evidence of its greater credibility, feasibility and/or utility, but rather, to provide information for further assessment and refinement of the model.

With regard to validation of the system, the decision to limit the field test in this manner was rooted in the cost and time infeasibility of comparing field test results with empirical data on training workplace behavior outcomes. In other words, in order to generate evidence on actual training outcomes, it would have been necessary to actually employ expensive methods of evaluation such as field studies involving participant observation or sophisticated tests of performance against baselines. These methods would have been prohibitively costly and time-consuming in the context of this dissertation. However, it might be feasible in other contexts, where empirical evidence of participant workplace behavior change is being gathered, to run a parallel test of the accuracy or credibility of participant self-assessments of change. Where empirical data on participant behavior change is available, it would be possible to administer the

personal benchmark system questionnaires to one group of training participants, and standard level three questionnaires to a control group, in order to determine whether the personal benchmark system generates more accurate information.

Moreover, there was no attempt made here to establish generalizability of the system's benefits to other training courses, run by other institutes, on other topics. Instead, according to Lincoln and Guba's alternate criterion of transferability, a better understanding of the applicability of the system in other contexts can only be established by further applications by training institutes who can then judge for themselves applicability to their needs. Broader conclusions with regard to the model's generalizability and value then could be drawn on the basis of accumulated experience of training institutes that might try to apply the model.

In addition to further research on the accuracy and generalizability of the system, further research can be used to further refine the system. The central idea of the personal benchmark system is that of overlapping questionnaires in which participants are asked to define for themselves learning and training use goals, and then are redistributed these goals in a subsequent phase and asked to provide ratings on achievement of their personal goals. There is much room for variation, however, with regard to the application. Different question wording and even different approaches to rating scales may be tried, as was done in the two field test applications – one of which used a numeric scale and the other a descriptive scale (“very much” “not at all” etc). Similarly, beyond the basic structure of the two learning goals and training use goal tables, it would be possible to experiment further with use of other overlapping open-ended questions between phases of the evaluation, such as the field test's questions on country needs and context in the pre-

course and end-of course questionnaires. In the end, variations on question wording and ratings scale, is largely a matter of the specific preferences and needs of DTIs using the method. However, accumulated experience may also lead to clearer determinations with regard to good practice and potential applications for the model.

Similarly, further research could be used in order to operationalize and explore two of the recommended revisions to the model on the basis of the field test. Specifically, the field test suggested that computerization of the model could significantly enhance its user-friendliness and feasibility. Moreover, the potential for using the personal benchmark system to better involve supervisors and workplace organizations in the training process needs to be further explored. In the conclusions of the field test, it was suggested that supervisors be asked to approve or comment on participant's individual learning goals in the pre-course questionnaire. Doing so could potentially increase supervisor involvement in the training process, with possible benefits for increasing ownership of target organizations over the training process and building supervisor support for and thus success of trainee transfer of training to the workplace. However, testing would be needed to determine whether supervisors would be likely to actually provide meaningful input to the process or would generally only "sign off" on the forms, perhaps without even reading them. The important issue of securing greater supervisor support for and involvement in capacity building processes involving training is one that has been discussed at length in Chapter Two and one which deserves further examination in any future tests of the model. Further research can also help develop new systems of quantifying the detailed qualitative information which these evaluations provide through various ways of coding responses and regression analysis.

In addition, this dissertation focused on the benefits of the personal benchmark system for achieving evaluation learning and accountability goals. However, in reality, the system is not only an evaluation system, but also a system for supporting instructional design and continuous improvement of course content. Understanding the possible benefits of personal benchmarking for these other aspects of training quality control was beyond the scope of this dissertation. However, it may be useful to explore these benefits in further research.

Finally, the concept of personal benchmarking – that is evaluations whereby quantitative scores are calculated on the basis of respondents' personal selection of goals and qualitative responses, is a unique concept that has not, to the knowledge of this researcher, been applied in other concepts. It is quite possible that personal benchmarking may have other useful applications outside of the context of training evaluation.

Conclusions

In sum, this dissertation has contributed to academic and practitioner literature in two ways: first, it has deepened our understanding of current development training institute (DTI) evaluation practice and unmet needs. Second, it has developed a new model of training evaluation which addresses some of the most significant utility, feasibility and credibility shortcomings of currently used models. Further research would be needed in order to develop a better understanding of the accuracy and generalizability of the model. However, adoption of the model by DTIs is likely to be

contingent not only on its success at addressing the evaluation needs cited by DTIs themselves, but also the willingness of DTIs to adopt a different approach to their evaluation regimes, entailing greater time investment in evaluation, greater involvement of course organizers in evaluation, and, finally, willingness to administer needs assessments and post-course evaluations for all courses.

Bibliography

Abelson, R. P. (1994). Opportunities in Survey Measurement of Attitudes. Questions About Questions. J. M. Tanur. New York, Russel Sage Foundation: 173-203.

Abma, T. A. (2006). "The Practice and Politics of Responsive Evaluation." American Journal of Evaluation 27(1): 31-43.

Alliger, G., S. Tannenbaum, et al. (1997). "A Meta-Analysis of the Relations Among Training Criteria." Personnel Psychology 50: 341-358.

Alliger, G. M. and E. A. Janak (1989). "Kirkpatrick's Levels of Training Criteria: Thirty Years Later." Personnel Psychology 42: 331-342.

American Evaluation Association (2004). Guiding Principles for Evaluators. Retrieved from: <http://www.eval.org/Publications/GuidingPrinciples.asp>.

Aquilino, W. S. and L. A. Losciuto (1990). "Effects of Interview Mode on Self-Reported Drug Use." Public Opinion Quarterly 54: 362-395.

Arthur, W. J. and W. J. Bennet (1998). "Factors that Influence Skill Decay and Retention: A Quantitative Review and Analysis." Human Performance 11(1): 57-101.

Asian Development Bank (2007). *Capacity Development in the Pacific*. Manila, Asian Development Bank.

Asian Development Bank (2008). *Effectiveness of ADB's Capacity Development Assistance: How to Get Institutions Right*. Manila, Asian Development Bank.

Atkins, P. W. B. and R. Wood (2002). "Self- vs Others' Ratings as Predictors of Assessment Centre Ratings: Validation Evidence for 360-Degree Feedback Programs." *Personnel Psychology* 55(4): 871-904.

Baldwin, T. T. and J. K. Ford (1988). "Transfer of Training: A Review and Directions for Future Research." *Personnel Psychology* 41(1): 63-105.

Bamberger, M. (2000). "The Evaluation of International Development Programs: A View from the Front." *American Journal of Evaluation* 21(1): 95-102.

Ban, C. and S. R. Faerman (1990). "Issues in the Evaluation of Management Training." *Public Productivity & Management Review* 13(3): 271-286.

Berthnal, P. B. (1995). "Evaluation that Goes the Distance." *Training and Development*: 41-45.

Birckmayer, J. D. and C. H. Weiss (2000). "Theory-Based Evaluation in Practice: What do We Learn?" *Evaluation Review* 24: 407-431.

Boesen, N. and O. Therkildsen (2003). *Capacity Development Evaluation Step 3: Draft Methodology for Evaluation of Capacity Development*. Copenhagen, Denmark, DANIDA.

Bogarert M.V.D., S. B., N.B. Bam (1981). *Participatory Evaluation of an Adult Education Programme*. *Participatory Research and Evaluation*. W. Fernandes and R. Tandon. New Delhi, Indian Social Institute: 172-182.

Bramley, P. (1996). *Evaluating Training Effectiveness*. London, McGraw Hill.

Bramley, P. a. A. C. N. (1984). "The Evaluation of Training Part 1: Clarifying the Concept." *Journal of European Industrial Training* 8(6): 10-16.

Brandenburg, D. C. (1982). "Training Evaluation: What is the Current Status?" *Training and Development Journal* 36(8): 14, 16-22.

Brandon, P. R. (1998). "Stakeholder Participation for the Purpose of Helping Ensure Evaluation Validity: Bridging the Gap Between Collaborative and Non-collaborative Evaluations." *American Journal of Evaluation* 19(3): 325-337.

Brinkerhoff, D. (2007). Capacity Development in Fragile States. Discussion Paper 58D. Maastricht, ECDPM.

Brinkerhoff, R. O. (1987). Achieving Results from Training. San Francisco, Jossey-Bass.

Brinkerhoff, R. O. (1988). "An Integrated Evaluation Model for HRD." Training and Development Journal 42(2): 66-68.

Brinkerhoff, R. O. (1989). "Using Evaluation to Transform Training." New Directions for Program Evaluation 44: 5-20.

Brinkerhoff, R. O. (2003). The Success-Case Method: Find Out Quickly What's Working and What's Not. San Francisco, Berrett-Koehler Publishers.

Brinkerhoff, R. O. (2006). Telling Training's Story: Evaluation Made Simple, Credible and Effective. San Francisco, Berret-Koehler.

Brinkerhoff, R. O. and A. M. Apking (2006). High Impact Learning. Cambridge, MA, Perseus Publishing.

Brinkerhoff, R. O. and M. O. Montesino (1995). "Partnership for Training Transfer: Lessons from a Coporate Study." Human Resource Development Quarterly 6(3): 263-274.

Brown, K. G. and M. W. Gerhardt (2002). "Formative evaluation: An integrative practice model and case study." *Personnel Psychology* 55(4): 951-983.

Brown, R. E. and C. S. Reed (2002). "An Integral Approach to Evaluating Outcome Evaluation Training." *American Journal of Evaluation* 23(1): 1-17.

Burgoyne, J. G. (1973). "An Action Research Experiment in the Evaluation of a Management Development Course." *Journal of Management Studies* get ref material from Hamblin: 8-14.

Bushnell, D. (1990). "Input, Process, Output: A Model for Evaluating Training." *Training and Development Journal* 44(3): 41-45.

Campbell, D. T. and J. C. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research*. Boston, Houghton Mifflin.

Carless, S. A. and G. P. Robert-Thompson (2001). "Self-Ratings in Training Programs: An Examination of Level of Performance and the Effects of Feedback." *International Journal of Selection and Assessment* 9(3): 217-225.

Carter, S. D. (2002). "Matching training methods and factors of cognitive ability: A means to improve training outcomes." *Human Resource Development Quarterly* 13(1): 71-87.

Carvalho, S. and H. White (2004). "Theory-Based Evaluation: The Case of Social Funds." 25(2): 141–160.

Catalanello, R. and D. L. Kirkpatrick (1968). "Evaluation Training Programs - the State of the Art." *Training and Development Journal*: 2-9.

Chen, H.-T. and P. H. Rossi (1983). "Evaluating With Sense: The Theory-Driven Approach." *Evaluation Review* 7: 283-302.

Clark, H. H. and M. F. Schober (1994). *Asking Questions and Influencing Answers. Questions about Questions.* J. M. Tanur. New York, Russel Sage Foundation: 15-48.

Clement, R. W. (1981). "Evaluating the Effectiveness of Management Training: Progress During the 1970s and Prospects for the 1980s." *Human Resource Management* 20(4) 20(4): 8-13.

Clement, R. W. (1982). "Testing the Hierarchy Theory of Training Evaluation: An Expanded Role for Trainee Reactions." *Public Personnel Management* 11: 176-184.

Cohen, D. J., L. C. Leviton, et al. (2006). "Online Diaries for Qualitative Evaluation: Gaining Real-Time Insights." *27*(2): 163-184.

Cole, G. E. (1999). "Advancing the Development and Application of Theory-based Evaluation in the Practice of Public Health." *American Journal of Evaluation* 20(3): 453-470.

Cooley, W. and W. Bickel (1986). *Decision-oriented Educational Research*. Boston, Kluwer-Nijhoff Publishing.

Cousins, J. B. (2004). *Crossing the bridge: toward understanding use through systematic inquiry*. *Evaluation Roots: Tracing Theorists' Views and Influences*. M. C. Alkin. Thousand Oaks, CA, Sage Publications: 319-330.

Cousins, J. B. and L. M. Earl (1995). *The Case for Participatory Evaluation: Theory, Reserch and Practice*. *Participatory Evaluation in Education: Studies in Evaluation Use and Organizational Learning*. J. B. Cousins and L. M. Earl. London, Falmer Press 3-18.

Cousins, J. B. and L. M. Earl (1995). *Participatory Evaluation in Education: What Do We Know? Where Do We Go?* *Participatory Evaluation in Education: Studies in Evaluation Use and Organizational Learning*. J. B. Cousins and L. M. Earl. London, Falmer Press: 159-180.

Cousins, J. B. and E. Whitmore (1998). "Framing Participatory Evaluation." *New Directions for Evaluation* 80: 5-23.

David J. Basarab, S. and D. K. Root (1993). *The Training Evaluation Process*. Boston, Kluwer Academic Publishers.

De Lange, P. and R. Feddes (2008). *General Terms of Reference "Evaluation of Dutch Support to Capacity Development"*. Hague, Netherlands, Ministry of Foreign Affairs.

DFID (2002). *Capacity Development: Where do we Stand Now?* London, DFID.

Dionne, P. (1996). "The Evaluation of Training Activities: A Complex Issue Involving Difficult Stakes." *Human Resource Development Quarterly* 7(3): 279-286.

Dixon, N. M. (1990). "The Relationship Between Trainee Responses on Participant Reaction Forms and Posttest Scores." *Human Resource Development Quarterly* 1: 129-37.

Dixon, N. M. (1996). "New Routes to Evaluation." *Training and Development*: 82-85.

Douglas Horton, A. A., et al. (2003). *Evaluating Capacity Development: Experiences from Research and Development Organizations around the World*. Ottawa, International Development Research Centre.

Dubin, S. S., M. Mezack, et al. (1974). "Improving the Evaluation of Management Development Programs." *Training and Development Journal*: 42-46.

Earl, S. and F. Carden (2002). "Learning from complexity: the International Development Research Centre's experience with Outcome Mapping." *12*(3): 518-524.

Earl, S., F. Carden, et al. (2001). *Outcome Mapping: Building Learning and Reflection into Development Programs*. Ottawa, Canada, International Development Research Centre.

Edelenbos, J. and A. v. Buuren (2005). "The Learning Evaluation: A Theoretical and Empirical Exploration." *Evaluation Review* 29: 591-612.

Erikson, P. R. (1990). "Evaluating Training Results." *Training and Development Journal* 44(1): 57-59.

Evans, D. (2007). *Are There New Approaches to the Evaluation of Learning Programs?* WBI Evaluation Series. Washington DC, World Bank Institute.

Facteau, J. D., G. H. Dobbins, et al. (1995). "The Influence of General Perceptions of the Training Environment on Pretraining Motivation and Perceived Training Transfer." *Journal of Management* 21(1): 1-25.

Faerman, S. R. and C. Ban (1993). "Trainee Satisfaction and Training Impact: Issues in Training Evaluation." *Public Productivity & Management Review* 16(3): 299-314.

Fetterman, D. M. (2001). *Foundations of Empowerment Evaluation*. London, Sage.

Fitz-enz, J. (1994). "Yes...You Can Weigh Training's Value." *Training* 31(7): 54-58.

Foddy, W. (1994). *Constructing Questions for Interviews and Questionnaires*. Cambridge, UK, Cambridge University Press.

Ford, J. K., M. A. Quinones, et al. (1992). "Factors Affecting the Opportunity to Perform Trained Tasks on the Job." *Personnel Psychology* 45(3): 511-527.

Fowler, F. J. (1995). *Improving Survey Questions: Design and Evaluation*. London, Sage Publications.

Foxon, M. (1993). "A Process Approach to the Transfer of Training: Part 1." *Australian Journal of Educational Technology* 9(2): 130-143.

Foxon, M. (1994). "A Process Approach to the Transfer of Training: Part 2." *Australian Journal of Educational Technology* 10(1): 1-18.

Fukuda-Parr, S., C. Lopes, et al. (2002). *Capacity for Development: New Solutions to Old Problems*. New York, NY, Earthscan.

Fukuyama, F. (2004). *State Building: Governance and World Order in the 21st Century*. Ithica, New York, Cornell University Press.

Garavaglia, P. L. (1993). "How to Ensure Transfer of Training." *Training and Development* 47(10): 63-68.

Georgenson, D. L. (1982). "The Problem of Transfer Calls for Partnership." *Training and Development* 36(10): 70-75.

Ghodsian, D., R. A. Bjork, et al. (1997). *Evaluation Training During Training: Obstacles and Opportunities. Training for a Rapidly Changing Workplace: Applications of Psychological Research*. M. A. Quinones and A. Ehrenstein. Washington DC, American Psychological Association.

Gill, S. J. (1989). "Using Evaluation to Build Commitment to Training." *New Directions for Program Evaluation*(44): 35-43.

Gist, M. E., A. G. Bavetta, et al. (1990). "Transfer Training Method: Its Influence on Skill Generalization, Skill Repetition and Performance Level." *Personnel Psychology* 43: 501 -523.

Group, N. C. Synthesis of Evaluations on Technical Assistance. Taastrup, Denmark, DANIDA.

Guba, E. G. and Y. S. Lincoln (1989). Fourth Generation Evaluation. London, Sage Publications.

Hamblin, A. C. (1974). Evaluation and Control of Training. London, Mc Graw Hill.

Holton, E. F. (1996). "The Flawed Four-Level Evaluation Model." Human Resource Development Quarterly 7(1): 5-19.

Holton, E. F., R. A. Bates, et al. (1997). "Toward Construct Validation of a Transfer Climate Instrument." Human Resource Development Quarterly 8(2): 95-113.

Horton, D., A. Alexaki, et al. (2003). Evaluating Capacity Development. Ottawa, Canada, IDRC.

Huczynski, A. A. and J. W. Lewis (1980). "An Empirical Study into the Learning Transfer Process in Management Training." Journal of Management Studies 17(2): 227-240.

IEG (2005). Capacity Building in Africa. Washington, World Bank.

IEG (2006). *Impact Evaluation - The Experience of the Independent Evaluation Group of the World Bank*. Washington, DC, World Bank.

IEG (2008). *Using Training to Build Capacity for Development: An Evaluation of World Bank project-financed and WBI Training*. Washington, World Bank.

Iverson, A. (2003). *Attribution and Evaluation in International Development: A Literature Review*. Ottawa, IDRC.

JICA (2006). *Summary of the Report Towards Capacity Development (CD) of Developing Countries Based on their Ownership: Concept of CD, its Definition and its Application in JICA Projects*. Tokyo, Japan, JICA.

Johnson, R. L., M. J. Willeke, et al. (1998). "Stakeholder Collaboration in the Design and Implementation of a Family Literacy Portfolio Assessment." *American Journal of Evaluation* 19: 339-353.

Joint Africa Institute Operations Evaluation Department (2008). *Evaluation of the Joint Africa Institute: Final Report*. Tunisia, African Development Bank.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.) Thousand Oaks, CA: Sage.

Kaufman, R. and J. M. Keller (1994). "Levels of Evaluation: Beyond Kirkpatrick."

Human Resource Development Quarterly 5(4): 371-380.

Kearns, P. and T. Miller (1997). *Measuring the Impact of Training and Development on the Bottom Line*. London, Pitman Publishing.

Keijzer, N. (2006). *Mapping of approaches towards M&E of Capacity and Capacity Development*. Maastricht, European Centre for Development Policy Management.

Keller, J. M. (1996). *Evaluating Diversity Training: 17 Ready-to-use Tools*. San Diego, Pfeiffer.

Kirkpatrick, D. L. (1968). *Evaluating Training Programs*. San Francisco, Berrett-Koehler Publishers Inc.

Knoweldge Advisers. (2006). *Training Operations Best Practices: Corporate Universities*

Kraiger, K., J. K. Ford, et al. (1993). "Application of Cognitive, Skill-Based and Affective Theories of Learning Outcomes to New Methods of Training Evaluation."

Journal of Applied Psychology 78(2): 311-328.

Kupritz, V. W. (2002). "The relative impact of workplace design on training transfer." *Human Resource Development Quarterly* 13(4): 427-447.

Land, T., Hauck, V. and H. Baser (2009). *Capacity development: between planned interventions and emergent processes Implications for development cooperation*. Policy Management Brief, No. 22. Maastricht.

Larbi, G., P. Jackson, et al. (2005). *Capacity Development Outcome Evaluation: Field-testing of the Methodology*. D. Copenhagen, Danida.

Lawrenz, F. and D. Huffman (2003). "How Can Multi-Site Evaluations be Participatory?" *24(4): 471–482*.

Le Compte, M. D. and J. P. Goetz (1982). "Problems of Reliability and Validity in ethnographic research." *Review of Educational Research* 52: 31-60.

Lee, S. H. and J. A. Pershing (2002). "Dimensions and Design Criteria for Developing Training Reaction Evaluations." *Human Resource Development International* 5(2): 175-197.

Levy, B. (2004). *Governance and Economic Development in Africa: Meeting the Challenge of Capacity Building*. *Building State Capacity in Africa: New Approaches*,

Emerging Lessons. B. Levy and S. Kpundeh. Washington, DC, World Bank Institute: 1-42.

Lincoln, Y. S. and E. G. Guba (1999). Establishing Trustworthiness. *Qualitative Research*. A. Bryman and R. G. Burgess. London, Sage. III: 397-444.

Locander, W., S. Sudman, et al. (1976). "An Investigation of Interview Method, Threat and Response Distortion." *Journal of the American Statistical Association* 71(354): 269-275.

Lopes, C. and T. Theilsohn (2003). *Ownership, Leadership and Transformation: Can We do Better for Capacity Development?* London, Earthscan.

Lynham, S. A. (2002). "The General Method of Theory-Building Research in Applied Disciplines." *Advances in Developing Human Resources* 4(3): 221-241.

Madavo, C. (2005). *Building Effective States, Forging Engaged Societies: Report of the World Bank Task Force on Capacity Development in Africa*. Washington DC, World Bank.

Malvin, J. H. and J. M. Moskowitz (1983). "Anonymous Versus Identifiable Self-Reports of Adolescent Drug Attitudes, Intentions and Use." *Public Opinion Quarterly* 47: 557-566.

Mangioni, T. W., Hingston, R., Barrett, J. (1982). "Collecting Sensitive Data: A Comparison of Three Survey Strategies." *Sociological Methods and Research*.

Marshall, C. and G. B. Rossman (1989). *Designing Qualitative Research*. Newbury Park, CA, Sage Publications.

Mathieu, J. E., S. I. Tannenbaum, et al. (1992). "Influences of Individual and Situational Characteristics on Measures of Training Effectiveness." *Academy of Management Journal* 35(4): 828-847.

Mayne, J. (1999). *Addressing Attribution Through Contribution Analysis: Using Performance Measures Sensibly: Discussion Paper Office of the Auditor General of Canada*.

McFarland, S. (1981). "Effects of Question Order on Survey Responses." *Public Opinion Quarterly* 45: 208-215.

Mertens, D. M., T. R. Berkeley, et al. (1995). *Using Participatory Evaluation in an International Context. Participatory Evaluation in Education*. J. B. Cousins and L. M. Earl. London, Falmer Press: 140-156.

Miles, M. B. and M. Huberman (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. London, Sage Publications.

Miller, R. L. and R. Campbell (2006). "Taking Stock of Empowerment Evaluation: An Empirical Review." *American Journal of Evaluation* 27(3): 296-319.

Morgan, P. (2006). *The Concept of Capacity*. Maastricht, Netherlands, ECDPM.

Morgan, P. (2006). *The Concept of Capacity: Draft Version. Study on Capacity, Change and Development*. Maastricht, Netherlands, ECDPM.

Morgan, P., T. Land, et al. (2005). *Study on Capacity, Change and Performance: Interim Report*. Maastricht, ECDPM.

Morgan, R. B. and W. J. Casper (2000). "Examining the factor structure of participant reactions to training: A multidimensional approach." *Human Resource Development Quarterly* 11(3): 301-317.

Naquin, S. S. and E. F. Holton III (2002). "The effects of personality, affectivity, and work commitment on motivation to..." *Human Resource Development Quarterly* 13(4): 357-376.

Negri, B. d., E. Thomas, et al. (1998). *Empowering Communities: Participatory Techniques for Community-Based Program Development*. Nairobi, Center for African Family Studies.

Newstrom, J. W. (1978). "Catch 22: The Problems of Incomplete Evaluation of Training." *Training and Development Journal*: 22-24.

Noe, R. (1986). "Trainees' Attributes and Attitudes: Neglected Influences on Training Effectiveness." *Academy of Management Review* 11(4): 736-749.

Noe, R., J. Sears, et al. (1990). "Relapse Training: Does it Influence Trainees' Post-training Behavior and Cognitive Strategies?" *Journal of Business and Psychology* 4(3): 317-328.

OECD (1992). *DAC Principles for Effective Aid*. Paris, OECD.

OECD (2005). *Paris Declaration on Aid Effectiveness*. Paris, OECD Development Assistance Committee.

OECD (2006). *The Challenge of Capacity Development: Working Towards Good Practice*. Paris, OECD.

Ofir, Z. and E. Bruning (2006). "Capacity for M&E: Moving Beyond Results-based Management." *Capacity.org*(29): 3.

Parker, B. L. (1986). "Summative Evaluation in Training and Development." *Journal of Industrial Teacher Education*(23): 2.

Patton, M. Q. (1982). *Practical Evaluation*. Beverly Hills, Sage.

Patton, M. Q. (1986). *Utilization-Focused Evaluation*. Beverly Hills, Sage Publications.

Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods*. Newbury Park, CA, Sage Publications.

Patton, M. Q. (1996). "A World Larger than Formative and Summative." *Evaluation Practice* 17(2): 131-145.

Pentland, B. (1999). "Building Process Theory with Narrative: From Description to Explanation." *Academy of Management Review* 24(4): 711-724.

Pershing, J. A. and J. L. Pershing (2001). "Ineffective reaction evaluation." *Human Resource Development Quarterly* 12(1): 73-90.

Phillips, J. j. and R. D. Stone (2002). *How to Measure Training Results*. New York, McGraw Hill.

Poulin, M. E., P. W. Harris, et al. (2000). "The Significance of Definitions of Success in Program Evaluation." *Evaluation Review* 24: 516-536.

Pulley, M. L. (1994). "Training 101: Navigating the Evaluation Rapids." *Training and Development*(September 1994): 19-24.

Reynolds, P. D. (1971). *A Primer in Theory Construction*. New York, Macmillan.

Richey, R. (1992). *Designing Instruction for the Adult Learner*. London, Kogan Page.

Richman-Hirsch, W. L. (2001). "Posttraining interventions to enhance transfer: The moderating effects of work environments." *Human Resource Development Quarterly* 12(2): 105-120.

Robinson, D. G. and J. C. Robinson (1989). *How to Link Training to Business Needs and Measure the Results*. London, Josey-Bass Inc.

Rockwell, S. K., L. Jha, et al. (2003). "Success Outcome Markers in Extension (SOME): Evaluating the Effects of Transformational Learning Programs." *Journal of Extension* 41(5).

Rodin, M. and B. Rodin (1972). "Student evaluations of teachers." *Science* 177: 1164-1166.

Rouiller, J. Z. and I. L. Goldstein (1993). "The Relationship Between Organizational Transfer Climate and Positive Transfer of Training." *Human Resource Development Quarterly* 4(4): 377-391.

Russ-Eft, D. and H. Preskill (2001). *Evaluation in Organizations: A Systematic Approach to Enhancing Learning, Performance and Change*. Cambridge, MA, Perseus Publishing.

Saari, L. M., T. R. Johnson, et al. (1988). "A Survey of Management Training and Education Practices in U.S. Companies." *Personnel Psychology* 41: 731-743.

Salas, E. and J. A. Cannon-Bowers (2001). "The Science of Training: A Decade of Progress." *Annual Review of Psychology* 52: 471-499.

Savedoff, W. D., R. Levine, et al. (2006). *When Will We Ever Learn? Improving Lives Through Impact Evaluation*. Washington, DC, Center for Global Development.

Scheirer, M. A. (2004). *Designing and Using Process Evaluation*. *Handbook of Practical Program Evaluation*. J. S. Wholey, K. Newcomer and H. Hatry. Hoboken, NJ, John Wiley & Sons: 40-68.

Schnoes, C. J., Virginia Murphy-Berman, et al. (2000). "Empowerment Evaluation Applied: Experiences, Analysis, and Recommendations from a Case Study." 21(1): 53–64.

Scriven, M. (1971). Objectivity and subjectivity in education research. Philosophical Redirection of educational research. L. G. Thomas. Chicago, University of Chicago.

Scriven, M. (1991). "Prose and Cons About Goal-Free Evaluation." Evaluation Practice 12: 55-62.

Smutylo, T. (2001). Crouching Impact, Hidden Attribution: Overcoming Threats to Learning in Development Programs. Draft Learning Methodology Paper, Block Island Workshop on Across Portfolio Learning.

Stammers, R. and J. P. . (1975). The Psychology of Training. London, UK, Methuen & Co Ltd.

Stokking, K. M. (1996). "Levels of Evaluation: Kirkpatrick, Kaufman and Keller, and Beyond." Human Resource Development Quarterly 7(2): 179-183.

Storberg-Walker, J. (2007). "Understanding the Conceptual Development Phase of Applied Theory-Building Research: A Grounded Approach." *Human Resource Development Quarterly* 18(1): 63-90.

Stufflebeam, D. (1997). *Strategies for Institutionalizing Evaluation Revisited: Abstract* University of Michigan Occasional Paper Series.

Stufflebeam, D. (2000). "The Methodology of Metaevaluation as Reflected in Metaevaluations by the Western Michigan Evaluation Center." *Journal of Personnel Evaluation in Education* 14(1): 95-125.

Stufflebeam, D. (2001). "Evaluation Models." *New Directions for Evaluation* 89.

Stufflebeam, D. (2004). "A Note on the Purposes, Development, and Applicability of the Joint Committee Evaluation Standards." *American Journal of Evaluation* 25(1): 99–102.

Stufflebeam, D. (2007). *Evaluation Theory, Models and Applications*. London, Jossey-Bass.

Stufflebeam, D. L. and L. A. Wingate (2005). "A Self-Assessment Procedure for Use in Evaluation Training." 26(4): 544-561.

Sullins, C. D. (2003). "Adapting the Empowerment Evaluation Model: A Mental Health Drop-in Center Case Example." *American Journal of Evaluation* 24(3): 387–398.

Swanson, R. A. and C. M. Sleezer (1987). "Training Effectiveness Evaluation." *Journal of European Industrial Training* 11(4): 7-16.

Tamkin, P., J. Yarnall, et al. (2002). *Kirkpatrick and Beyond: A Review of Models of Training Evaluation*. IES PDF Reports. Brighton, UK, Institute of Employment Studies.

Tannenbaum, S. (1997). "Enhancing Continuous Learning: Diagnostic Findings from Multiple Companies." *Human Resource Management* 36(4): 437-452.

Tannenbaum, S. and G. Yukl (1992). "Training and Development in Work Organizations." *Annual Review of Psychology* 43: 399-441.

Tascherau, S. (1998). *Evaluating the Impact of Training and Institutional Development Programs*. Washington, DC, World Bank.

Taut, S. (2007). "Studying Self-Evaluation Capacity Building in a Large International Development Organization." *American Journal of Evaluation* 28(1): 45-59.

Taylor, P. (2001). *10 Key Stages Towards Effective Participatory Curriculum Development*. Zurich, Helvetas.

Taylor, P., P. Clarke, et al. (2008). No Capacity, No Development: Draft Document. Sussex, UK, Institute of Development Studies.

Tracey, J. B., S. I. Tannenbaum, et al. (1995). "Applying Trained Skills on the Job: The Importance of the Work Environment." *Journal of Applied Psychology* 80(2): 239-252.

Training (1996). "Industry Report: Who's Learning What?" *Training* 53-65.

Turnbull, B. (1999). "The Mediating Effect of Participation Efficacy on Evaluation." *Evaluation and Program Planning* 22(2): 131-140.

UNDP (2006). Capacity Assessment Practice Note. New York, UNDP.

UNDP (2007). Supporting Capacity Development: The UNDP Approach. New York, NY, UNDP.

Wang, G. G. and D. R. Spitzer (2005). "Human Resource Development Measurement and Evaluation: Looking Back and Moving Forward" *Advances in Developing Human Resources* 7: 5-15.

Warr, P., C. Allen, et al. (1999). "Predicting three levels of training outcome." *Journal of Occupational and Organizational Psychology* 72: 351-375.

Warr, P., M. Bird, et al. (1970). *Evaluation of Management Training*. London, Gower Press.

Warr, P. and D. Bunce (1995). "Trainee Characteristics and the Outcomes of Open Learning." *Personnel Psychology* 48: 347-373.

Watson, D. (2006). "Embracing Innovative Practice." *Capacity.org*(29): 4-6.

Wentland, E. J. and K. W. Smith (1993). *Survey Responses: An Evaluation of their Validity*. San Diego, Academic Press, Inc.

Wexley, K. N. and T. T. Baldwin (1986). "Posttraining Strategies for Facilitating Positive Transfer: An Empirical Exploration." *Academy of Management Journal* 29(3): 503-520.

Whyte, A. (2004). *A Landscape Analysis of Donor Trends in International Development. Human and Institutional Capacity Building*, Rockefeller Foundation.

Wnuk, J. J. (1966). "Evaluation of Conceptual Training: A Step-by-Step Procedure." *Training and Development Journal*: 38-40.

World Bank (1979). *Structural Change and Development Policy*. World Development Report. Washington, DC, World Bank.

World Bank (1997). *The State in a Changing World*. World Development Report.

Washington, DC, World Bank.

World Bank (2004). *Global Monitoring Report: Policies and Actions for Achieving the MDGs and Related Outcomes*. Washington DC, Development Committee, World Bank and International Monetary Fund.

Yamnill, S. and G. N. McLean (2001). "Theories supporting transfer of training." *Human Resource Development Quarterly* 12(2): 195-208.

Appendix A: Summary of Development Capacity Building Studies³

Report	Looked At	Results	Targeting	Integration	Long-Term	Ownership	M&E
World Bank- IEG. Capacity Building in Africa , 2005.	World Bank financed capacity building activities in Africa.	World Bank capacity building support more relevant than in the past (more targeted to institutional capacity and building public demand) but still insufficiently effective in building sustained public sector capacity.	X	X	X	X	X
World Bank- IEG, Using Training for Capacity Building , 2008.	World Bank financed and World Bank Institute training activities.	Most training results in learning but only about half of training succeeds in achieving sustainable organizational/institutional capacity building	X	X	X	X	X
World Bank. Building Effective States, Forging Engaged	World Bank financed capacity building in Africa.	Capacity remains a binding constraint on development and poverty reduction.	X	X	X	X	X

³ The final five columns of this table track the following variables:

Targeting: The extent to which capacity building interventions target specific country, organizational and/or individual needs.

Integration: The extent to which interventions address the range of issues affecting capacity.

Long-Term: The extent to which capacity building programs are sufficiently long-term to build sustainable capacity over time.

Ownership: The extent to which there is sufficient country/target institution ownership of interventions.

M&E: The extent to which adequate monitoring and evaluation is incorporated into capacity building programs.

Report	Looked At	Results	Targeting	Integration	Long-Term	Ownership	M&E
Societies: Task Report of the World Bank Force on Capacity Development in Africa. 2005.							
OECD-DAC. The Challenge of Capacity Development: Working Towards Good Practice.2006.	General review of donor and partner country performance on capacity development	Adequate country capacity is one of the critical missing factors in current efforts to meet the Millennium Development Goals. Significant efforts are required to improve CD efforts.	X	X	X	X	X
UNDP. Ownership, Leadership and Transformation : Can we Do Better for Capacity Development? 2003	Assessment of the use of technical assistance for developing country capacity building, includes 56 case studies of capacity development initiatives.	Limited success – and sometimes downright failure – of development policies in recent years due largely to lack of country ownership, this in turn is related to lack of country capacity. “We can do better for capacity development.”	X	X	X	X	
UNDP. Capacity for Development: New Solutions	Assessment of the use of technical assistance for	Technical cooperation does not seem to be triggering transformative	X	X	X	X	

Report	Looked At	Results	Targeting	Integration	Long-Term	Ownership	M&E
to Old Problems. 2002	developing country capacity building.	processes of developing the knowledge, institutions and mobilization of society for development.					
JICA. Towards Capacity Development (CD) of Developing Countries Based on their Ownership. 2006	Analysis of JICA projects with capacity building goals.		X	X	X	X	X
ECDPM. Capacity development: between planned interventions and emergent processes: Implications for development cooperation. 2009	16 CD case studies involving different sectors, objectives, organizations and geographic locations. Research explored the nature of capacity, processes of endogenous CD, success	Conventional approaches to capacity development premised on planned technocratic interventions, fail to grasp the political, social and cultural dimensions of change that are intrinsic to sustainable outcomes. This accounts in part for the poor record of development cooperation in this field. CD must be viewed as an	X	X	X	X	X

Report	Looked At	Results	Targeting	Integration	Long-Term	Ownership	M&E
	factors for CD and the role of “outsiders”.	endogenous process.					
ADB. Effectiveness of ADB’s Capacity Development Assistance: How to Get Institutions Right. 2008	ADB CD interventions in Cambodia, Nepal and the Philippines.	ADB interventions have mixed results, with with no consistent patterns of success across sectors, countries, regions, or time. One-time, stand alone advisory technical assistance grants are seldom an effective tool for CD.	X	X	X	X	X
ADB. Capacity Development in the Pacific. 2007	Analysis by ADB and AusAid of CD case studies in the pacific region.	Few of the cases studied revealed a thorough assessment of capacity when planning projects. Many cases attributed "success" to a relatively solid understanding of the local context.	X	X	X	X	X
JAI. Evaluation of the Joint Africa Institute: Final Report. 2008	Review of JAI performance.	Difficult to assess performance due to lack of available documentation. 1. The themes and	X	X	X	X	X

Report	Looked At	Results	Targeting	Integration	Long-Term	Ownership	M&E
		<p>content of training events appear to have been often determined by supply-side considerations.</p> <p>2. No monitoring and evaluation of activities.</p>					

Appendix B: List of Interviewed Development Training Institutes

1. Administrative Staff College of India
2. African Institute for Capacity Development, Tanzania
3. Centre for Arab Women Training and Research, Tunisia
4. Center for Excellence in Finance, Slovenia
5. Centre for Training on Regional Integration, Uruguay
6. Geneva Centre for Democratic Control of Armed Forces, Switzerland
7. Independent National Legal Training Centre, Afghanistan
8. Institute for the Integration of Latin America and the Caribbean, Argentina
9. Institute of Rural Management, Pakistan
10. Instituto Interamericano para el Desarrollo Económico y Social, Inter-American Development Bank, USA
11. International Development Law Organization, Italy
12. International Monetary Fund Institute, USA
13. International Program for Development Evaluation Training, Canada
14. International Training Centre of the ILO, Italy
15. INWENT - Internationale Weiterbildung und Entwicklung gGmbH, Germany
16. Joint Africa Institute, Tunisia
17. Justice Studies Centre of the Americas, Chile
18. Latin American and Caribbean Institute for Economic and Social Planning, Chile
19. Mount Carmel Training Centre, Israel
20. World Bank Institute, USA

Appendix C: MCTC Gender and Local Governance Course Description⁴

The Golda Meir Mount Carmel International Training Center (MCTC), the Ministry of Foreign Affairs, Center for International Cooperation (MASHAV), in collaboration with the United Nations Human Settlements Programme (UN-HABITAT) jointly organize a training course on Gender and Local Governance for local government trainers and other interested parties.

The objective of the course is to strengthen gender in local government through training of trainers. The participants will:

- Increase their understanding of why gender is important in urban planning, land rights and housing, budgeting, economic development, media, urban environment and climate change, violence in cities as well as other urban development,
- Recognize gender biases in local governance and obstacles to the equal participation of women,
- Familiarize themselves with existing tools for gender and local governance training,
- Enhance their confidence to deliver gender training to diverse groups,
- Experience concrete examples of gender roles and gender mainstreaming in Israel during study visits.

⁴ Taken from MCTC course literature.

The programme of the workshop includes various themes relating to gender and governance, which are approached through different methods, as for example presentations,

sharing of experiences and learning by doing. As the course is designed as a training of trainers event, particular emphasis is placed on group discussion and experimental learning. The themes of the workshop include:

- Introduction to Gender training
- Participation of women in local government, and the role of quotas.
- Women/Gender in the Israeli Society
- Women and land rights/housing
- Gender and Urban Planning
- Gender Budgeting/Women and local government financing
- Women and local economic development
- Media and gender
- Women and the Urban Environment/Climate Change
- Violence against women/Safer cities

In addition the participants of the workshop will join two excursions: A two day tour to Southern Israel, including the Dead Sea, a Kibbutz in the Negev and a visit to meet Bedouin Women and a day tour to Jerusalem and Nazareth.

Who Should Apply:

The course targets local government trainers from government affiliated training institutes and NGOs. Anyone who provides training for elected councillors or local government staff may apply. Gender training experience is not required. Course language is English only.

Appendix D: MCTC Support Systems for Microenterprises

Course Description⁵

INTRODUCTION

Poverty eradication and equal opportunities for women and youth are among the United Nations' Millennium Development Goals (MDGs) for the 21st century.

*“Governments should ensure that poverty reduction is mainstreamed into all policies, ranging from a national macroeconomic strategy to local-level administrative actions. Particular attention should be paid to the creation of additional opportunities for decent work. Public investment and public institutions should endeavor to target the poor, particularly with regard to their expenditure on education, health and infrastructure.”*⁶

One of the main conclusions of the UNECE International Conference on “Reducing Barriers to Entrepreneurship and Encouraging Enterprise Development: Policy Options” was that: *“For successful enterprise development, it is crucial that the major stakeholders cooperate closely: i.e. central and local governments, operational and would-be entrepreneurs, enterprise associations and training centers for entrepreneurs. Cooperation with the interested partners provides Governments with essential feedback, helping them to establish a regulatory and institutional environment that should create*

⁵ Taken from MCTC course materials.

⁶ “The Millennium Development Goals Report ,Sha Zukang , *Under-Secretary-General for Economic and Social Affairs*, United Nations, 2008

incentives for entrepreneurship, thus encouraging technological change and promoting local economic development.”⁷

In this context, the current international workshop on “**Support Systems for Microenterprises**” is being conducted by MCTC as part of a long term agreement between YABT and MASHAV.

This workshop addresses one of the key issues related to *poverty alleviation* by encouraging entrepreneurship. Micro and small enterprises are acknowledged as important tools to create new opportunities and to generate additional sources of income. There is a recognized need for support tools and support systems that can give a holistic package to a young person who either wants, or- in these times of economic crisis - is forced, to find some alternative income by starting his or her own business.

Micro-enterprises are more flexible and adaptable to rapid changes than bigger companies. However, small-scale entrepreneurs, especially at the early stages, often lack the necessary information and tools needed for running their businesses. They are also in need of updated data about the market and about the threats and opportunities around them. Planners have to become aware of the need to enhance knowledge and managerial skills to ensure the success and survival of new small business ventures and of the importance of creating appropriate frameworks that can offer support, guidance and managerial tools to those willing to take the risk of entrepreneurship.

⁷ *“Developing Entrepreneurship in the UNECE Region, Country Experience in Reducing Barriers to Enterprise Development, ,United Nations, New York and Geneva,2008,p.54*

Israel has been engaged in supporting small businesses for more than two decades now. The Small Business Development Centers (SBDCs) and Technological Incubators are professional centers that provide a wide range of services to entrepreneurs and business owners.

The Golda Meir Mount Carmel International Training Center (MCTC) has been sharing the Israeli experience with participants from all over the world and has successfully cooperated in entrepreneurial training and in the establishment of business support centers in former participants' countries. It is within this framework that the present international workshop is offered.

OBJECTIVES

Upon completion of the workshop, the participants will have:

- widened their knowledge of support systems for small-scale entrepreneurs as a tool for regional and local development
- become acquainted with the support systems for micro and small entrepreneurship in Israel including technological incubators
- discussed the position of young and older entrepreneurs at times of global economic crisis

ADMISSION REQUIREMENTS

The workshop is designed for 26 – 30 participants, including individual entrepreneurs, representatives of governmental institutions and SME support agencies, NGOs dealing with small business support and development and women's business associations, as well as employees and others recommended by international organizations such as

UNECA,UNDP,UN-HABITAT,ECOWAS,ILO,ICA,IDB,CABEL,ADB,ILO, other entities within the OAS and national and international funds dealing with entrepreneurial development.

Preference will be given to candidates recommended by the YABT.

Candidates are expected to have relevant qualifications in economics, business administration, regional planning, industrial management or related subjects, with a minimum of five years of work experience. Candidates working in rural areas and in remote regions of a country are encouraged to apply.

The program of the workshop will stress the interrelation in this field between government workers and small scale entrepreneurs. A maximum of 5 participants will be accepted from each country representing a balance between the number of applicants accepted from government institutions and the private sector.

WORKSHOP PROCEDURE

Participants will reside at MCTC and the studies will be intensive. Five days a week are normally devoted to classroom study and observation visits. Saturdays and Sundays will be spent on assignments or on excursions to places of historical, religious or general interest. The themes will be presented in the form of **lectures** and **workshops**, by experienced lecturers from both the academic world and the private and institutional sectors. A special **practical module** on organizing and running a business will be part of the program. The workshop will be held in English; translation will be used when necessary.

PROGRAM OF STUDIES

- Support systems in Israel
- Microcredit schemes
- Marketing strategies for small businesses
- Entrepreneurial Training – Business Laboratories
- Information and communication technologies (ICTs) and new opportunities for young entrepreneurs; the Internet as a source of information and advice
- Business networking
- Preparation of a project proposal

Study Visits

Participants will have the opportunity to meet Israeli microentrepreneurs in different settings and to visit development centers and other institutions involved in supporting microenterprise development.

Among them:

- Small Business Development Centers
- Technological Incubators
- The Tefen Industrial Park

Exchange of Professional Experiences

An important aspect of the Workshop is the exchange of experiences among its participants and several sessions will be devoted to this purpose. The rich knowledge brought by the students from different countries is a very valuable contribution both to the proceedings and to the future development of their peers’.

Final Assignment

At the end of the Workshop, participants will present to the plenum a practical project proposal or innovative idea related to the subject matter that could be implemented in their own settings.

CERTIFICATE

Upon completion of the Workshop the graduates will receive a certificate confirming their attendance and active participation.

Appendix E: Results of DTI Interviews⁸

DTI	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
Formal Planning and Evaluation Mechanisms																					
Pre-course participant consultation	X	X	X					X	X		X	X	X								9
Utility	X	X										X									3
Improve																					0
Pre-course supervisor consultation	X	X						X	X		X	X									6
Utility	X	X																			2
Improve																					0
Pre-test			X														X				2
Utility																					0
Improve																					0
Level 1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	20
Utility	X	X	X			X			X	X	X	X	X		X	X	X				12
Improve				X										X							2
Level 2		X		X													X				3
Utility		X															X				2
Improve			X				X					X									3
Level 3	X						X	X	X			X			X					X	7
Utility	X						X														2
Improve	X	X	X	X	X		X					X	X	X			X		X		11
Level 4															X						1

⁸ Letters were randomly assigned to training institutes in order to protect anonymity of responses.

Separate evaluation budget	X	X		X			X			X			X							6
Separate eval. unit	X			X									X					X		4
External evaluator					X		X		X	X			X							5
Evaluation by course staff					X	X	X	X		X	X	X	X							8
Evaluation Constraints																				
Money		X			X	X								X				X		5
Too much info/hard to process	X												X							2
Low response rate		X											X							2
Low DTI support		X			X									X						3
Time					X					X				X						3
Lack of know-how	X	X			X	X												X		5
Data accuracy							X					X						X		3

Table Key

Formal Planning and Evaluation Mechanisms:

Utility: This form of evaluation was cited as useful by respondents in response to question 2, 5, 9, 10 or 14 of the DTI survey.

Improve: This form of evaluation was cited in question 11 or 13 of the DTI survey as one which the DTI does not presently employ but would like to employ

in the future or one which the DTI employs but would like to improve or extend its application.

Pre-course participant consultation: DTIs ask participants, as part of their formal needs assessment or application process to submit information prior to attending the course on their work environments, needs and expectations from the course or other information that can be used to shape course content (question 3).

Pre-course supervisor consultation: DTIs contact participants' work supervisors or organizations before the course to ask about needs or expectations from course (question 4).

Pre-test: Participants complete knowledge tests either prior to or at the beginning of the course.

Level 1: End-of-course participant satisfaction questionnaires and/or feedback (question 8).

Level 2: End-of-course tests of participant learning (question 8).

Level 3: Post-course assessment of participant workplace use of training (question 8).

Level 4: Post-course assessment of impact of training on the capacity of participants' workplace organizations (question 8).

Post-course formal supervisor consultation: As part of its evaluation regime, DTI contacts participants' work supervisors or organizations after the course to ask about the usefulness of course content or to provide guidance (question 4).

Informal Planning and Evaluation Mechanisms:

Informal Level 2: Measurement of learning through mechanisms other than tests, such as class projects and presentations (question 8, 10).

Informal post-course participant feedback: Information on course or post-course use of learning gathered through contacts with participants not specifically for the purposes of evaluation, such as alumni networks or ad-hoc phone calls (question 8, 10).

Post-course informal supervisor consultation: Information on post-course use of learning through contacts with supervisors not specifically for the purpose of evaluation, such as ongoing dialogue with beneficiary organizations (question 8, 10).

Country/institution-wide coordinator input: Information on course content gathered through consultations with DTI country coordinators or other local institutional representatives not directly in contact with course participants (question 8, 10).

Course coordinator feedback: Information on course based on assessment of course coordinator, including through informal conversations with participants (question 8, 10).

In-course case studies: Information on participant needs and workplace context gathered through participant presentations of case studies from their own contexts (question 8, 10).

Evaluation Administration:

Separate evaluation budget: DTIs explicitly earmark part of their course or institution budgets to evaluation (question 7).

Separate evaluation unit: DTI has a separate evaluation unit responsible for administering evaluations and/or processing evaluative data (question 6).

External evaluator: DTIs use an independent evaluator, external to the DTI to do periodic evaluations of courses and/or workplace outcomes or impact of training (question 6, 8).

Evaluation by course staff: Course staff are responsible for administering evaluation (question 6).

Evaluation Constraints(question 11, 13): Constraints cited by respondents to making desired improvements to their evaluation regime.

Money: insufficient budgets for evaluation.

Too much information/hard to process: amount of evaluative data received is difficult for the DTI to process in a way that it can be useful for learning and accountability.

Low response rate: level three and/or four evaluation questionnaires do not get sufficiently high response rates to be seen as representative.

Low DTI support: Management and/or staff of DTI do not value evaluation or see the benefit in making improvements.

Time: DTI staff does not have the time to do more extensive evaluation.

Lack of know-how: DTIs believe themselves to be lacking in sufficient evaluative expertise or tools to make desired improvements to evaluation system.

Data accuracy: DTIs suspect data collected through evaluation mechanisms to inaccurately represent participants' experiences and training outcomes or impact.

Appendix F: Gender Course Pre- and End-of Course Questionnaires⁹

International Workshop on Gender and Local Governance

September 1-15,2009

Introductory Questionnaire

Welcome to the MASHAV/MCTC-UN-Habitat workshop on Gender and Local Governance. We would like to ask you to fill out this brief questionnaire. We will use this information to better understand your expectations of the workshop, to adapt - where possible - the workshop to your specific needs, and to track whether the workshop is meeting these expectations. At the end of the workshop, you will be presented with the goals you have formulated for yourself here and will be asked to provide feedback on whether the workshop met your expectations. In order to make this possible, you have been asked here to provide your name. Your responses, however, will be kept confidential and will only be seen by MCTC/UN-Habitat staff. Thank you for your help and we wish you a successful and enjoyable course!

Name _____

Organization/Position _____

1. In your opinion, what are the three main challenges to gender mainstreaming in your country today?
 - a.
 - b.
 - c.

⁹ All questionnaires have been condensed from the originally distributed questionnaires to fit the page. Original questionnaires left more space for responses.

Name _____

Please share with us your personal goals for the course.¹⁰

What knowledge, skills, or other benefits would you most like to gain from the course? Please be as specific as possible. You may list up to three goals.	For each of the goals listed above, how do you expect to use what you have gained here once you have returned to your workplace?	Now that you have completed the course, please review the goals you set for yourself. For each goal, please circle the response which best represents the extent to which the course met your needs.	Comments
Goal 1:		<ul style="list-style-type: none"> • Not at all • Partially met my needs. • Mostly met my needs. • Entirely met my needs. • Not Applicable: Course content did not address this topic. 	

¹⁰ Respondants were instructed to fill in the first two columns of the questionnaire only in the pre-course questionnaire and then were redistributed the questionnaire at the end of the course and asked to complete the last two columns.

Goal 2:		<ul style="list-style-type: none"> • Not at all • Partially met my needs. • Mostly met my needs. • Entirely met my needs. • Not Applicable: Course content did not address this topic. 	
Goal 3:		<ul style="list-style-type: none"> • Not at all • Partially met my needs. • Mostly met my needs. • Entirely met my needs. • Not Applicable: Course content did not address this topic. 	

International Workshop on Gender and Local Governance
September 1-15,2009
End of Course Questionnaire

The following is an evaluation questionnaire for the workshop you have just completed. We would appreciate you taking the time to fully answer each of these questions, as they will help us learn how to improve future courses. We will also use your responses in order to help us track how useful your participation in the course will be to achieving your personal goals. In two months from now, you will be asked to fill out a final follow-up questionnaire to report on your experience. For this reason, the questionnaire is not anonymous. All your responses here, however, will be kept confidential.

Name: _____

You may share my responses with course staff. Yes _____ No _____

1. In your opinion, what are the three main challenges to gender mainstreaming in your country today?

- a. _____

- b. _____

- c. _____

2. To what extent was the course relevant for the particular circumstances of your country? Please circle one of the following:

- Not relevant.
- Slightly relevant
- Mostly relevant
- Extremely relevant

3. Please suggest up to two improvements that could be made to enhance the success of this course:

- a.
- b.

[Note: Participants were redistributed their individual learning goal tables, as completed in the pre-course questionnaire and asked to fill out the right two columns. In addition, participants were asked to complete the following utility goal table:]

Name_____

Email_____

You may share my responses with course staff: Yes_____ No_____

What are the three most useful things which you gained from the course?	How do you believe you will use these things in your work?
1)	
2)	
3)	

Appendix G: Microfinance Pre-Course and End of Course Questionnaires

Support Systems for Microenterprises

September 1-15

Introductory Questionnaire

Welcome to the MASHAV-MCTC-YABT course on Support Systems for Microenterprises. We would like to ask you to fill out this brief questionnaire. We will use this information to better understand your expectations of the course, to adapt - where possible - the course to your specific needs, and to track whether the course meets these expectations. At the end of the course, you will be presented with the goals you have formulated for yourself here and will be asked to provide feedback on whether the course met your expectations. In order to make this possible, you have been asked here to provide your name. Your responses, however, will be kept confidential and will only be seen by MCTC course staff. Thank you for your help and we wish you a successful and enjoyable course!

Name_____

Organization/Position_____

Please discuss the present situation for entrepreneurs in your country.

1. In your opinion, what are the main areas in which microenterprises are in need of support in your country?
2. Are you aware of support systems for microenterprises in your own country? If so, what are they and do they address the challenges you listed above?
3. In your opinion, what sort of improvements to existing support systems/new support systems do you believe need to be developed to better support microenterprises in your country?

Name _____

Please share with us your personal goals for the course.¹¹

What knowledge, skills, or other benefits would you most like to gain from the course? Please be as specific as possible. You may list up to three goals.	For each of the goals listed above, how do you expect to use what you have gained here once you have returned to your workplace?	Now that you have completed the course, please review the goals you set for yourself. For each goal, please tell us the extent to which the course met your needs. Please rate on a scale of 1-4 where 1 = did not meet my needs 4 = fully met my needs. If the goal you set was not relevant for the course, please select N/A	Comments
Goal:		1 2 3 4 N/A	
Goal:		1 2 3 4 N/A	
Goal:		1 2 3 4 N/A	

¹¹ Respondants were instructed to fill in the first two columns of the questionnaire only in the pre-course questionnaire and then were redistributed the questionnaire at the end of the course and asked to complete the last two columns.

Support Systems for Microenterprises

September 1-15
End of Course Questionnaire

The following is an evaluation questionnaire for the course you have just completed. We would appreciate you taking the time to fully answer each of these questions, as they will help us learn how to improve future courses. We will also use your responses in order to help us track how useful your participation in the course will be to achieving your personal goals. In two months from now, you will be asked to fill out a final follow-up questionnaire to report on your experience. For this reason, the questionnaire is not anonymous. All your responses here, however, will be kept confidential and only seen by MCTC course staff.

Name: _____

1. On the basis of what you've learned here, what sort of improvements to existing support systems/new support systems do you believe need to be developed to better support microenterprises in your country?

2. To what extent was the course relevant for the particular circumstances of your country? Please rate on a scale of 1-4 where 1 is not relevant at all and 4 is entirely relevant. 1 2 3 4

3. Please suggest up to two improvements that could be made to enhance the success of this course:
 - a.
 - b.

[Note: Participants were redistributed their individual learning goal tables, as completed in the pre-course questionnaire and asked to fill out the right two columns. In addition, participants were asked to complete the following utility goal table:]

Name _____

Email _____

You may share my responses with course staff: Yes _____ No _____

What are the three most useful things which you gained from the course?	How do you believe you will use these things in your work?
1)	
2)	
3)	

Appendix H: Structured Course Organizer Assessment Questionnaire – Gender Course¹²

International Workshop on Gender and Local Governance

September 1-15,2009

End of Course Questionnaire

The following is an evaluation questionnaire for the workshop you have just completed. We would appreciate you taking the time to fully answer each of these questions, as they will help us learn how to improve future courses. We will also use your responses in order to help us track how useful your participation in the course will be to achieving your personal goals. In two months from now, you will be asked to fill out a final follow-up questionnaire to report on your experience. For this reason, the questionnaire is not anonymous. All your responses here, however, will be kept confidential.

Name: _____

You may share my responses with course staff. Yes _____ No _____

2. In your opinion, what are the three main challenges to gender mainstreaming in your country today?
 - a. .
 - b. .
 - c. .

Please review responses to this question in the pre-and post-course questionnaires.

This question had two goals:

¹² Questions asked the course organizer in the Microfinance Course were identical to those asked here in the Gender course, and thus, the Microfinance interview schedule has not been reproduced here. In both cases, course organizers were redistributed copies of the evaluation questionnaires and ask to respond to the assessment questions appearing in bold.

Goal 1 (pre-course questionnaire): to provide you with information about the specific country contexts and needs of participants so that they could be taken into account in course delivery.

Was this information on country contexts was useful enough to you to justify the time spent reading responses. (Circle one) Not very /somewhat / yes

Would the information would have been more useful to me if it had been obtained as part of the participant application form rather than on the first day of the course. No/Somewhat/Yes

Goal 2: To measure shifts in attitudes over the course by comparing pre-course and post-course questionnaire responses as a way of measuring shifts in attitudes. (Please read both pre-course and post-course questionnaires of each participant before responding to this question).

Was this information on shifts in attitude was useful enough to you to justify the time spent reading responses. Yes/ Somewhat / No

Comments

2. To what extent was the course relevant for the particular circumstances of your country? Please circle one of the following:

- Not relevant.
- Slightly relevant
- Mostly relevant
- Extremely relevant

3. Please suggest up to two improvements that could be made to enhance the success of this course:

- a.
- b.

Was this information on suggest course improvements useful enough to you to justify the time spent reading responses.

Yes/ Somewhat / No

Goal Table #1 (Column 1 & 2 administered pre-course, column 3 & 4 administered post-course).

Please share with us your personal goals for the course.

What knowledge, skills, or other benefits would you most like to gain from the course? Please be as specific as possible. You may list up to three goals.	For each of the goals listed above, how do you expect to use what you have gained here once you have returned to your workplace?	Now that you have completed the course, please review the goals you set for yourself. For each goal, please tell us the extent to which the course met your needs. Please rate on a scale of 1-4 where 1 = did not meet my needs 4 = fully met my needs. If the goal you set was not relevant for the course, please select N/A	Comments
Goal:		1 2 3 4 N/A	
Goal:		1 2 3 4 N/A	
Goal:		1 2 3 4 N/A	

In the pre-course questionnaire participants were asked to set their individual course goals and how they expected to use what they gained in their workplace (Goal Table 1: columns 1 & 2). This question had two goals:

Goal 1: Provide you with information on the extent to which participant expectations were in line with course content.

Did it give you sufficient information? Yes/Somewhat/No

Was this information useful enough to you to justify the time spent reading responses?

Yes/Somewhat/No

Goal 2: To provide you with information that you could use to adapt course content to specific needs of participants

Was this information useful enough to you to justify the time spent reading responses?

Yes/Somewhat/No

For both of these goals, would this information have been more useful to you for both if it had been obtained as part of the participant application form rather than on the first day of the course.

Yes/Somewhat/No

In the end of course questionnaire, participants were re-distributed their goal tables and asked to rate whether they had achieved their goals. (Goal Table 1 Columns 3&4). Asking participants to rate achievement of their individual goals was intended to give you with a more accurate and nuanced picture of what participants got out of the course. Please answer the following questions with regard to these columns:

Please review all of the ratings of individual goals done by course participants. In order to get a picture of the accuracy of participant ratings, please count how many course participants gave high marks (3&4) for achievement of their personal goals in topic areas not actually covered by the course.

Number of participants giving high (3-4) ratings to topics not covered in the course:

Number of participants that wrote n/a for goals that were not covered in course:

Did you find that this method of having participants select individual goals and rating achievement of these goals gave you more nuanced information of what participants got out of the course? Yes / Somewhat/No

Did it justify time spent reading responses? Yes/Somewhat/No

Participants were also given space to provide optional comments. Would you have preferred if participants were asked more specifically to elaborate on their ratings? Yes.Somewhat/No

Comments

Participants were also asked to fill in the following goal table. They will be re-sent this table two months after the course and asked to report on the extent to which they were able to use the knowledge that they received in course and if not, why not.

Goal Table #2 (administered post-course)

Please share with us how useful the course will be/was for you.

Of the knowledge/skills which you acquired in the course, which do you believe will be most useful to you back on the job? You can chose up to three knowledge/skill areas.	How do you believe you will use the knowledge/skills in your work?
Knowledge/skill area #1:	
Knowledge/skill area # 2	
Knowledge/skill area #3	

Was this information on course utility useful enough to you to justify the time spent reading responses. (Circle one) **Not very /somewhat /**
yes

Please elaborate.

From which of the questions in the post-course questionnaire do you feel that you learned the most? Please name up to two questions.

From which did you learn the least? Please name up to two questions.

How many times have you administered this course or similar courses before?

Additional Comments