

Running Head: A Method for Evaluating the *Fireline Leadership* Training

A Method For Evaluating The *Fireline Leadership* Training

Michael T. DeGrosky

Fort Hays State University

**A Paper Submitted in Partial Requirement for
LDRS 890 – Internship in Organizational Leadership**

Dr. Curtis Brungardt

**THIS MATERIAL CANNOT BE USED, IN WHOLE OR IN PART, FOR PUBLICATION,
PERSONAL OR COMMERCIAL GAIN OR DISTRIBUTION, WITHOUT THE WRITTEN
CONSENT OF THE AUTHOR.**

Abstract

Current evaluation practices will likely prove inadequate to obtain long-term support for the leadership development initiative of the National Wildfire Coordinating Group (NWCG) Leadership Committee and their sponsor agencies. The scale, expense and strategic importance of the *Fireline Leadership (FLL)* program suggest that the sponsor agencies must monitor this training on a basis that is methodologically and statistically defensible. This project developed a quantifiable and, where appropriate, a statistically supportable method for collecting and analyzing training related data to support the NWCG leadership initiative with the intent of validating that the *FLL* training is on track and providing a model or template for ongoing evaluation of, not only the *FLL* program, but the broader NWCG leadership curriculum. The results are an intermediate evaluation strategy of the *FLL* training program, that employs, in part, an evaluation measure of quasi-experimental design oriented to producing Level 3/Behavioral evaluation as described by Kirkpatrick (1996) as well as Tannenbaum and Woods. (1992) The evaluation strategy employs four instruments including an immediate post-course questionnaire; a delayed, post-course survey for course participants; a delayed, post-course survey for the supervisors of course participants; and a qualitative focus group protocol.

Table of Contents

	PAGE
TABLE OF CONTENTS	3
LIST OF FIGURES	5
LIST OF TABLES	5
ACKNOWLEDGEMENTS	6
CHAPTER	
I. INTRODUCTION	7
Background	7
Statement of the Problem	8
Purpose and Research Questions	9
II. LITERATURE REVIEW	11
Current Efforts to Evaluate the <i>Fireline Leadership Training</i>	11
Evaluating Training	13
Evaluating Leadership Training	14
<i>FLL</i> Learning Targets	16
Theoretical Considerations: Evaluation Strategy	20
Kirkpatrick	20
Measuring Return-On-investment (ROI)	40
Evaluating Training Using a Mix of Quantitative and Qualitative Methods	42
Additional Measures & Methods To Crosscheck Self-Reporting Responses	43
Using A Mix of Immediate Measures and Delayed Measures	44
Pretest/Posttest Evaluation Design	45
Control Groups and Baseline Data	46
Response-Shift and Controlling Response-Shift Bias	47
Stakeholder Based Evaluation Strategy	54
Keeping It Simple	56
Training Evaluation Strategy	58
Evaluation Instruments	71

Table of Contents

	PAGE
III. METHODOLOGY	72
Overview	72
Step 1: Literature Search: Fireline Leadership	73
Step 2: Literature Search: Training Evaluation	74
Step 3: Evaluate Data Sources	74
Step 4: Determine Key or Core Competencies	76
Step 5: Select/Design Evaluation Method	76
Step 6: Develop Data Collection Instruments	77
Step 7: Develop Quantifiable/Statistical Measures	77
Step 8: Small-scale Test	78
Sample and Limitations of the Study	79
Step 9: Revise Strategy/Instruments as Necessary	81
Step 10: Develop Procedure	81
Step 11: Deliver Test Results Procedure and Instruments	81
Step 12: Operational During the 2005 Training Cycle	81
IV. PRESENTATION OF FINDINGS	82
Introduction	82
Evaluation Strategy (Assessment of <i>FLL</i>)	82
Immediate Post-Course Attitude Survey	86
Delayed Survey Administered to Participants	90
Delayed Survey Administered to Supervisors	90
V. CONCLUSIONS AND RECOMMENDATIONS	96
Research Questions and Conclusions	96
Other Conclusions	107
Summary Of Conclusions	109
Limitations of the Study	111
Recommendations	112
Implications for Future Research	113
Final Discussion	116
VI. REFERENCES	117
VII. APPENDIX A.	122

List of Figures

FIGURE		PAGE
1	<i>FLL</i> Element/Competency	17
2	Leadership Committee Grouping: Communicating Vision and Intent	18
3	Leadership Committee Grouping: Application of Leadership Skills	19
4	Leadership Committee Grouping: Teambuilding	19
5	Leadership Committee Grouping: Detecting and Mitigating Error	19
6	Leadership Committee Grouping: Managing Stress & Other Human Factors	19
7	Leadership Committee Grouping: Decision Making	20
8	Leadership Committee Grouping: Ethics	20

List of Tables

TABLE		PAGE
1	Sample Size by Treatment Group	80
2	Survey Response by Treatment Group	80
3	Immediate Post-Course Attitude Survey	87
4	Participant Post-Training Alpha Coefficients	90
5	Participant Pre-Training Alpha Coefficients	90
6	Supervisor Post-Training Alpha Coefficients	91
7	Supervisor Pre-Training Alpha Coefficients	92
8	Paired Samples T-Test: Survey Administered to Participants	94
9	Paired Samples T-Test: Survey Administered to Supervisors	96

Acknowledgements

The author would like to gratefully acknowledge the people who gave their time, expertise and support to assist in this project. First, thanks to my advisor and major professor, Dr. Curtis Brungardt, for his overall guidance, ideas and insight into the project; and to Dr. Brent Goertzen for his counseling and assistance on sampling, surveying and statistical matters. I would also like to express my appreciation to Larry Sutton and Jim Cook, Co-Chairs, and the other members, of the National Wildfire Coordinating Group Training Working Team Leadership Committee; without whose cooperation this project would not have been possible. I also owe thanks to Lark McDonald and his staff for their generous collection and sharing of data. Finally, a special thanks goes to my wife Tami for her tolerance and support.

Chapter I

Introduction

Background

Problems associated with the practice of leadership contribute to accidents and fatalities in the work environment of wildland firefighters. Findings associated with investigations of the South Canyon and Thirtymile Fires, and the Wildland Firefighter Safety Awareness Study identified leadership as an issue requiring attention by wildland fire agencies. (Training Working Team¹, 2003) Consequently, the National Wildfire Coordinating Group (NWCG) Training Working Team chartered the Leadership Committee. The parent organization charged the Leadership Committee with promoting cultural change in the workforce, emphasizing the vital importance of leadership concepts and sound leadership practices in the wildland fire environment, and providing educational and leadership development opportunities. (Training Working Team², 2003)

Among its specific tasks, the Leadership Committee's charter directs them to recommend a framework for promoting the understanding and practice of sound leadership principles in the wildland fire workforce and, within that broad charge, recommend development of new or modified leadership training courses or on-the-job training as needed. (Training Working Team³, 2003) Consequently, the Leadership Committee established a distinct leadership curriculum, made-up of courses that are connected conceptually from the follower level to the organizational leader level. (Training Working Team², 2003) Within that curriculum, *Fireline Leadership (FLL)* provides critical basic training for entry-level career firefighters who are moving into

leadership positions, and furnishes emerging leaders with foundational concepts, enabling them to assume more leadership responsibility on the job. (Training Working Team⁴, 2003) Specifically, *FLL* is a weeklong training program designed to provide the background knowledge and skills required for effective crew level leadership.

Since 2000, approximately 6,000 people have completed the *FLL* training program. (Personal conversation with Lark McDonald, October 5, 2003; McDonald, 2001) Consequently, the *FLL* course provides a fundamental, foundation element of the overall NWCG leadership training curriculum, and is therefore of strategic importance to that organization and its member agencies. Through Mission-Centered Solutions, their private sector provider, the sponsoring agencies deliver the *FLL* program below market costs when compared to other available alternatives. However, when compared to other training typically delivered within the target agencies, *FLL* training remains relatively expensive, with tuition at roughly \$950 per participant. (McDonald, 2001) According to McDonald, (personal conversation with Lark McDonald, October 5, 2003), as sponsor agencies face uncertain budgets, agency training funds appear susceptible to reduction, and calls for cost justification are increasing.

Statement of the Problem

Given the strategic importance of this training, the scope of participation, the relative costs to the participating agencies, and the vulnerability of agency training funds; the sponsor organizations appear to have strong incentive to evaluate the *FLL* program to maximize the return on their substantial investment in this training. The Leadership Committee, is in fact, charged with establishing a mechanism for evaluating

the effectiveness of leadership training, with the objective being to accurately assess how the leadership training impacted job performance. (Training Working Team², 2003)

Current evaluation practices would likely prove inadequate to obtain long-term support for the leadership development initiative of the Leadership Committee and their sponsor agencies. The scale, expense and strategic importance of the *FLL* program suggest that the sponsor agencies must monitor this training on a basis that is methodologically and statistically defensible.

Purpose and Research Questions

This study endeavored to develop a quantifiable and, where appropriate, statistically supportable method for collecting and analyzing training related data to support the leadership initiative. The intent of evaluating the *FLL* training is to validate that the training is on track and provide a model or template for ongoing evaluation of, not only the *FLL* program, but also the broader leadership curriculum.

The purpose of the of this project is to craft an evaluation method that:

- Provides a systematic, quantitative approach to evaluation.
- Allows the Leadership Committee and its sponsor agencies to determine whether the NWCG leadership training is effective at emphasizing the vital importance of leadership concepts and sound leadership practices in the wildland fire environment.
- Uses the *FLL* training program as a pilot or test-bed for determining whether the NWCG leadership curriculum is bringing about change in participant performance on key leadership competencies forming the core of the programs.

- Assesses whether the NWCG leadership training curriculum is effectively promoting cultural change in the workforce by evaluating whether behaviors are extending into the organization beyond the training environment.

The research questions of this project include:

- What are the program goals or objectives of the FLL training, and can they be used as a basis for evaluating the training?
- What evaluation strategy and evaluation design will serve the FLL training curriculum effectively?
- Can easily obtainable and inexpensive self-report data, supported by other qualitative data collection, provide a sustainable source of feedback on the *FLL* training approximating Kirkpatrick's Level 1 (Reaction), Level 2 (Learning), and Level 3 (Behavior)?
- Is it practical to measure the efficiency of FLL training at Kirkpatrick's Level 4 (Results) or for return-on-investment (ROI)?
- Can one use a design that compares pre-training and post-training results without using a control group to evaluate the *FLL* training?
- What evaluation methods, techniques, measures or instruments will best evaluate the FLL training at Kirkpatrick's Levels 1, 2 and 3?

Chapter II

Literature Review

Current Efforts to Evaluate the Fireline Leadership Training

A review of available literature indicates little effort to evaluate the effectiveness of the *FLL* training beyond Kirkpatrick's Level 2 (learning.) McDonald, (2001) reported that the *FLL* program is being evaluated through a number of mechanisms on four levels of training evaluation, using Kirkpatrick's model for training evaluations. However, with the exception of established Level 1 evaluation, none of the program evaluations to-date has reached the rigorous level of design implied by Kirkpatrick's guidelines. Clearly, the *FLL* program has been evaluated using participant reaction, as mid-course and final end-of-course evaluations are administered in every *FLL* session. Beyond this Level 1 (reaction) evaluation, assessments of *FLL* have been few, and those that evaluators have completed have not been systematic, and are pre-experimental in nature, including field observation of participants and one limited study of the program's effectiveness, completed by Hillman and McDonald. (Hillman & McDonald, 2003.) Their study has been described as a "non-scientific" approach, but may be more accurately described as a pre-experimental, post-training comparison intended at obtaining Kirkpatrick's Level 2 (learning) and Level 3 (behavioral) evaluation. This study investigates a population sample from just one geographic strata of the overall *FLL* participant population, and strives to achieve multiple aims, measuring the effectiveness of the training as well as attending to other immediate information needs including participants' attitudes about the value/return on investment of leadership training and the desirability of continued contracted delivery. (Personal

conversation with Lark McDonald, October 5, 2003; personal conversation with Jim Cook, October 14, 2003, Hillman & McDonald, 2003)

Mission-Centered Solutions personnel have also pursued behavioral evaluation (Kirkpatrick's Level 3) by directly observing performance during on-the-job training and observation sessions in the work environment with previously trained *FLL* participants. These evaluation efforts have also taken a largely "non-scientific" approach, primarily collecting data from a small sample, representative of one functional stratum of the overall *FLL* population, and with little control over sample or observer bias. The intent of these efforts has primarily been to determine participants' retention of key course concepts and what forms of support would be most effective to maximize course effectiveness.

The results of assessment efforts to-date indicate that participants remain well satisfied with the *FLL* program and have witnessed short-term improvement in their own behaviors and performance. In addition, participants have also noted short-term improvement in the performance of others resulting from the training. Anecdotal evidence also exists suggesting that change may be occurring within the workforce resulting from the training program. (McDonald, 2001; McDonald & Shadow, 2003) However, specific to the *FLL* training, beyond Kirkpatrick's Level 1 (reaction) oriented evaluation, the evidence derives from pre-experimental, non-systematic evaluations of samples representing limited strata of the population and lacking statistical supportability.

However, though limited, the level of evaluation completed on the *FLL* training actually places the efforts of the sponsor agencies and their contract provider above

both the training evaluation efforts of many organizations, and of the larger NWCG training curriculum.

Evaluating Training

Training evaluation represents an obligation for the training organization, which is simultaneously very important and very difficult. Organizations must evaluate the effectiveness of their training if they hope to assess training outcomes and whether they are using their training resources wisely. Ultimately, training evaluation enables organizations to modify program design, delivery, and administration to improve program outcomes. (Brungardt, 1997; Cousins, 1995; Delahoussaye, 2001; Mann, 1997; Nickols, 2000) Speaking specifically of leadership education, Brungardt (1997) asserts, "Before we can progress with the development of comprehensive curricula and other forms of intervention, it is imperative that we create and utilize assessment and evaluate procedures to measure leadership growth. It is essential that we take a detailed look at the current status of leadership education and evaluate the impact these intervention programs have on participating students." The literature makes clear that researchers inquiring into training agree on the importance of evaluating training. (Brungardt, 1997; Alliger and Tannenbaum, 1997; Bee, 2000; Eseryel, 2002; Kirkpatrick, 1956; Santos & Stuart, 2003; Swierczek & Carmichael, 1985; Tannenbaum & Woods, 1992; Tyler, 2002) However, there is similarly strong agreement on the complexity and difficulty of evaluating training effectively. (Abernathy, 1999; Alliger & Tannenbaum, 1997; Santos & Stuart, 2003) In addition, both time and money for evaluating training and researching evaluation are often scarce. Consequently, in general, organizations don't adequately evaluate training, and despite substantial

training investments, organizations are slow to evaluate their efforts and many do not conduct extensive evaluation of their management training to determine whether the training provided is effective. (Mann, 1997; Saari, Johnson, McLaughlin & Zimmerle, 1988; Santos & Stuart, 2003; Shulha, Cousins & Bradley, 1997) For example, according to Tannenbaum and Woods (1992), many organizations “fail to close the loop through systematic evaluation, and consequently make many training decisions based on anecdotes, trainee reactions, hunches, or inertia.” The authors also indicate that actions based solely on participant response, anecdote and personal observation fail to effectively support training initiatives.

Some researchers believe that by evaluating training, trainers engage in a combination of research, training improvement, and internal marketing to demonstrate value. (Abernathy, 1999, Kirkpatrick, 1996) However, others disagree with these sentiments, asserting that the only important information to gather is what is best for the training organization, its employees, and its constituents, specifically the impact of the training that determines whether the training improved the employee’s performance. (Tyler, 2002) However, despite intent, according to Bee (2000) “...it is now generally accepted that training evaluation can encompass a range of activities all aimed at finding out how well we support people in their learning.”

Evaluating Leadership Training

Leadership training usually refers to learning activities geared toward a specific leadership role or job, and these learning activities are typically considered to be components of leadership education, and leadership training programs are extremely common in both the private and public sectors. (Brungardt, 1997; Delahoussaye, 2001)

The American Society for Training and Development reports that organizations spend tens of billions of dollars annually on corporate training and development, much of which might be classified as leadership training. In addition, the majority of the nation's largest corporations offer some type of leadership training for their employees." (ASTD, 2003; Brungardt, 1997) Paradoxically, there is surprisingly little research into the effectiveness of leadership training. However, despite the paucity of research into its efficacy, leadership training is generally thought to be effective. (Bass, 1990; Brungardt, 1997; Saari et al., 1988)

Answering how well a leadership development training program is working requires clear program goals, and clear program goals enable the ability to evaluate a leadership development program. (Bee, 2000; Brungardt, 1997; Delahoussaye, 2001; Mann, 1997, Pernick, 2001) Most often, these program goals are expressed in the form of training objectives. According to training evaluation researchers, the presence of clear training objectives is critical to good training evaluation. (Bee, 2000; Mann, 1997; Pernick, 2001; Swierczek & Carmichael, 1985) Program goals or training objectives are alternatively known as learning targets and student outcomes.

Brungardt (1997) studied the effectiveness of a university leadership education program, and definitions he used to describe program outcomes can inform this discussion and help clarify the use of leadership education outcomes terminology used here. In the context of his study, Brungardt defines "student outcomes" as "... the qualities, talents and skills the student has obtained as a result of participating in a leadership program"; "student attitudes" as "...the beliefs, opinions, reactions, and feelings a student has about leadership"; and "student knowledge" as "...the

participants' comprehension of the facts, issues, techniques, and theories of leadership.”

FLL Learning Targets

A review of the *FLL* program materials and existing literature reveals that the *FLL* training does not take a conventional objectives-based approach, a fact that, unmitigated, could complicate efforts to effectively evaluate the training's outcomes. Pernick (2001) recommends that, “If the learning objectives were not well formed at the outset, it is vital to make them so retrospectively. Work with people who understand what the learning event was trying to achieve and tease out, and agree with them, a set of well formed objectives.” For the purposes of this project, this dilemma was anticipated, and to achieve the aims described by Pernick (2001), a review of the extant literature identified 33 separate elements, the apparent areas of competence addressed by the *FLL* training, initially described as “competencies.” Figure 1. illustrates the 33 elements identified as the competencies that the *FLL* training addresses.

Figure 1. FLL Element/Competency

<i>Element/Competency</i>	<i>Identified By</i>
1. After Action Review (AAR)	Hillman & McDonald
2. Briefing Behaviors	DeGrosky Review
3. Character	DeGrosky Review
4. Communication Tools (not active listening)	Hillman & McDonald
5. Conflict Resolution	Hillman & McDonald
6. Decision-making	Hillman & McDonald
7. Ethics (Professional Ethics)	Hillman & McDonald
8. Expectations and Behavior	DeGrosky Review
9. Fear	DeGrosky Review
10. Goals and Objectives	DeGrosky Review
11. Problem Solving (see Performance Feedback)	Hillman & McDonald
12. Delegation	Hillman & McDonald
13. Error Management/Mitigation	Hillman & McDonald
14. Experience and Memory – Making Slides	Hillman & McDonald
15. “I Message”	DeGrosky Review
16. Leader’s Intent	Hillman & McDonald
17. Leader’s Role (duties, responsibilities, principles)	Hillman & McDonald
18. Leadership Environment	DeGrosky Review
19. Listening Skills	Hillman & McDonald
20. Performance Feedback (see Problem Solving)	DeGrosky Review
21. Power (Sources of)	Hillman & McDonald
22. Prioritization ¹	Hillman & McDonald
23. Situation Awareness (SA)	Hillman & McDonald
24. Self Development	Hillman & McDonald
25. Leadership Styles (“Situational Leadership”) ²	Hillman & McDonald
26. Stress Mitigation	Hillman & McDonald
27. Subordinate Development	Hillman & McDonald
28. Teamwork/Team Cohesion (cohesion in Error unit)	Hillman & McDonald
29. Teambuilding	Hillman & McDonald
30. Training	Hillman & McDonald
31. Traumatic Stress (Critical Incident Stress)	DeGrosky Review
32. Values	DeGrosky Review
33. Vision	DeGrosky Review

1 Not Apparent As a Learning Objective/Target of the Course

2 FLL Does Not Teach Theoretically Based “Situational Leadership” – “Leadership Styles”

The 33 elements or competencies identified fall across a spectrum that can be described as a mix of student outcomes, student attitudes and student knowledge as defined by Brungardt (1997). Given this, and the fact that the *FLL* does not take a conventional objectives-based approach, for the purposes of this project they will be referred to as the course's "learning targets." The findings of the literature search, pertaining to the identification and validation of the learning targets of the *FLL* training were supplemented with validation by, and consultation with, the Leadership Committee. In the process, one learning target was eliminated, and the Leadership Committee established the learning targets in groups that the Committee believed might be reliably measured through an evaluation process. Their groupings appear below in Figures 2 through 8.

Figure 2. Leadership Committee Grouping: Communicating Vision and Intent

<i>Leadership Committee Grouping</i>	<i>Learning Target</i>
<i>Communicating Vision and Intent</i>	
	Expectations and Behavior
	Briefing Behaviors
	Communication Tools - Not Listening
	Goals and Objectives
	"I Message"
	Leader's Intent
	Performance Feedback ³
	Vision ⁴

³ LC Not Sure This Fell Into The Competency Category

⁴ Identified by Hillman & McDonald, Reported by DeGrosky, But Not Addressed by LC

Figure 3. Leadership Committee Grouping: Application of Leadership Skills

Leadership Committee Grouping	Competency/Target
<i>Application of Leadership Skills</i>	
	Character ⁵
	Conflict Resolution
	Delegation
	Leader's Role – Duties, Responsibilities, Principles
	Leadership Environment
	Power (Sources Of)
	Self Development
	Leadership Styles – Situational Leadership ⁶

⁵ LC Did Not Feel This Was A Measurable Personal Competency

⁶ FLL Does Not Teach Theoretically Based "Situational Leadership" – Leadership Styles

Figure 4. Leadership Committee Grouping: Teambuilding

Leadership Committee Grouping	Competency/Target
<i>Teambuilding</i>	
	Subordinate Development
	Teamwork/Team Cohesion
	Teambuilding
	Training

Figure 5. Leadership Committee Grouping: Detecting Error and Mitigating

Leadership Committee Grouping	Competency/Target
<i>Detecting Error and Mitigating</i>	
	After Action Reviews
	Error Management & Mitigation

Figure 6. Leadership Committee Grouping: Managing Stress & Other HF

Leadership Committee Grouping	Competency/Target
<i>Managing Stress & Other Human Factors</i>	
	Fear
	Stress Mitigation
	Traumatic Stress ⁷

⁷ LC Saw This As A Process Accomplished By Others, Not A Personal Competency

Figure 7. Leadership Committee Grouping: Decision Making

<i>Leadership Committee Grouping</i>	<i>Competency/Target</i>
<i>Decision Making</i>	
	Decision Making
	Problem Solving
	Experience And Memory - Making Slides
	Prioritization
	Situation Awareness

Figure 8. Leadership Committee Grouping: Ethics

<i>Leadership Committee Grouping</i>	<i>Competency/Target</i>
<i>Ethics</i>	
	Ethics
	Values

Theoretical Considerations: Evaluation Strategy

For any training evaluation to be valuable, training criteria must be psychometrically sound, meaningful to decision makers, and must be able to be collected within typical organizational constraints. (Alliger & Tannenbaum, 1997; Eckert, 2000; Fitzgerald & Hammon, 1999) Ideally, training should be evaluated at multiple levels, using multiple methods, techniques and instruments and contribute to continuous improvement of the training. (Brown, 2003; Brungardt, 1997; Kirkpatrick, 1996; Warr & Catriona, 1999) According to Brungardt (1997), "Generally, the more methods, techniques, and instruments used, the more effective the evaluation process will likely be."

Kirkpatrick

One of the most comprehensive and widely referenced models of training evaluation is Donald Kirkpatrick's. (Kirkpatrick, 1956; 1996) According to Kirkpatrick (1996), the reasons for evaluating training programs are to:

- Decide whether to continue offering a particular training program
- Improve future programs
- Validate your existence and job as a training professional

The four levels of Kirkpatrick's evaluation model are as follows:

1. Reaction.

A measure of how participants feel about the various aspects of a training program, essentially measuring customer satisfaction. Training managers routinely employ "Level 1" evaluation and frequently base decisions about training on participants' comments. Training evaluators also measure participant reaction to assess participant motivation and interest in learning.

2. Learning.

A measure of knowledge acquired, skills improved, or attitudes changed due to training.

3. Behavior.

A measure of the extent to which participants change their on-the-job behavior because of training. Commonly referred to as transfer of training.

4. Results.

A measure of outcomes occurring due to training such as higher productivity, reduced costs, less employee turnover, and improved quality.

Kirkpatrick's guidelines for measuring each of the levels includes:

Level 1: Reaction

- Determine what you want to find out.
- Design a form that will quantify reactions.
- Seek honest reactions.
- Encourage written comments and suggestions.
- Attain an immediate response rate of 100 percent.
- Develop acceptable standards.
- Measure reactions against the standards and take appropriate action.
- Communicate the reactions as appropriate.

According to Kirkpatrick (1996), from an analysis of reactions, one can determine how well participants accepted a training program, and obtain comments and suggestions that will help improve future programs. The more favorable the reactions to a program are, the more likely trainees are to pay attention and learn the principles, facts, and techniques discussed. However, favorable reactions do not assure learning.

Level 2: Learning

- Where possible, use a control group (not receiving the training).
- Measure the learning of each trainee so that quantitative results can be determined.
- Use a before-and-after approach so that learning can be related to the program.
- Attain a response rate of 100 percent.
- As much as possible, the learning should be measured on an objective basis.
- Where possible, analyze the evaluation results statistically so that learning can be proven in terms of correlation or level of confidence.
- Use the results of the evaluation to take appropriate action.

It is important to determine objectively the amount of learning that takes place. From Kirkpatrick's perspective, learning is defined in a rather limited way: as the principles, facts, and techniques were understood and absorbed by trainees. Kirkpatrick does not concern himself with on-the-job use of the principles, facts, and techniques at this level of evaluation. Evaluation of learning is more difficult than evaluation of reaction. Evaluators must have a knowledge of statistics, requiring the assistance of a statistician to plan the evaluation procedures, analyze the data, and interpret the results. Level 2 evaluation requires a great deal of work to plan the evaluation procedure, analyze the data, and interpret the results.

Level 3: Behavior

- Use a control group, if feasible.
- Conduct a systematic appraisal of on-the-job performance on a before-and-after basis.
- Conduct a post-training appraisal three months or more after training.
- Subsequent appraisals may add to validity of the study.
- Survey or interview (one or more) trainees, their bosses, their subordinates, and others who often observe trainees' behavior on the job.)
- Choose 100 trainees or an appropriate sampling.
- Conduct a statistical analysis to compare before-and-after performance and relate changes to the training.
- Repeat the evaluation at appropriate times.
- Consider the cost of evaluation versus the potential benefits.

Evaluation of training in terms of on-the-job behavior is more difficult than reaction and learning evaluations, requiring a more scientific approach and the consideration of many factors.

Level 4: Results

- Use a control group, if feasible.
- Allow enough time for results to be achieved.
- Measure both before and after training, if feasible.
- Repeat the measurement at appropriate times.
- Consider the cost of evaluation versus the potential benefits.
- Be satisfied with the evidence if absolute proof isn't possible to attain.

Kirkpatrick (1996) cautions that evaluation becomes more difficult, complicated, and expensive as it progresses from level 1 to level 4, but also becomes more important and meaningful. According to Kirkpatrick, if the time, money, and expertise are available, it is important to proceed through all four levels without skipping any, or at least evaluate programs at as many of the four levels as possible.

Other evaluation researchers agree that evaluation becomes more complicated and expensive with each level. (Pernick, 2001; Tannenbaum & Woods, 1992) According to Pernick, (2001) “Generally, each succeeding level of evaluation increases rigor and cost. It is financially prudent, therefore, to consider the program’s success criteria. For example, if success is primarily measured by satisfied participants who assert their intention to apply what they have learned, a Level-1 evaluation suffices. Conversely, if the program must pay its way to stay alive, you should determine return on investment by tabulating and subtracting program costs from bottom-line indicators such as revenue enhancement or cost savings.”

In 1999, the American Society for Training and Development found that 45 percent of surveyed organizations only gauged trainees’ reactions to courses. (Sugrue, 2003) Overall, 93 percent of training courses were evaluated at Level 1, and 52 percent at Level 2. Only 31 percent of training courses were evaluated at Level 3, and 28 percent at Level 4. Since that time, the ASTD shows that the overall percentage of organizations employing Level 1 evaluations has fallen substantially (75%). However, little growth has occurred in the percentage of organizations employing Level 2, 3 and 4 evaluations. In 2002, 41% of the ASTD’s Benchmarking Service organizations evaluated their courses at level 2 and 21% using Level 3. (Eseryel, 2002; Sugrue, 2003)

As evidenced by the literature, Kirkpatrick's model remains the most influential and prevalent model for training and education evaluation. Commenting on the Kirkpatrick model, Santos and Stuart (2003) illustrate the conventional wisdom regarding the model when they say, "This model has been highly influential. According to a recent survey by the American Society for Training and Development, it is still the most commonly used evaluation framework among the Benchmarking Forum Companies. The model is also widely accepted in the field of industrial/organizational psychology..." Alliger and Tannenbaum (1997) reinforce the point when they report, "Perhaps unsurprisingly, Kirkpatrick's four-level model continues to be the most prevalent framework for categorizing training criteria."

According to Santos and Stuart (2003), "The widespread acceptance of conventional evaluation models has much to do with their simplicity and prescriptive appeal." Writing both with Janak and with Tannenbaum, George Alliger expresses this very sentiment as the strength of the Kirkpatrick model. (Alliger & Janak, 1989; Alliger & Tannenbaum, 1997) According to Alliger and Janak (1989), "The power of Kirkpatrick's model is its simplicity and its ability to help people think about training evaluation criteria. The Kirkpatrick model provides training evaluators with an easily adopted vocabulary and a rough taxonomy where none existed previously, which likely explains why it became so well known and remains widely used, including in evaluations of leadership education."

However, despite its many strengths, its influence, and its ongoing popularity, Kirkpatrick's model is not without its considerable flaws. Consequently, the Kirkpatrick model should not be regarded as the final word on training evaluation design.

According to Alliger and Tannenbaum (1997), “This simple taxonomy of training criteria became very popular in business and academia because it addressed a need to understand training evaluation simply yet systematically.” However, they go on to say that “The model’s simplicity is appealing but, as revealed in more recent work, this simplicity is also a liability.” According to Alliger and Tannenbaum, “Kirkpatrick’s model may never have been meant to be more than a first, global heuristic for training evaluation. As such it is well done.”

According to Alliger and Janak (1989), three problematic assumptions regarding Kirkpatrick’s model may be identified.

- First, that its levels are arranged in ascending order of information provided
- Second, that its levels are causally linked
- Third, that its levels are positively inter-correlated

These implied assumptions, appear widely in the literature, and when taken together, effectively establish a hierarchical model in the field of training evaluation. However, these implied assumptions have the potential to cause misunderstanding and over-generalization. (Alliger and Janak, 1989; Alliger and Tannenbaum, 1997)

Assumption 1: Arranged in an ascending order.

Kirkpatrick (1956, 1996) created a model of training evaluation comprised of four “steps” implying a procedural purpose. However, those steps appear to be arranged in an ascending hierarchy of information provided. Consequently, they are commonly viewed, both by practitioners and researchers, as “levels”, implying a qualitative difference between levels. The consequence of the ascending order arrangement permeates the literature. According to Alliger and Janak (1989), “This assumption has

the flavor of reasonableness, which may explain a nearly absolute lack of discussion about it in the literature.” However, this first assumption leads to a perception that “Level 4” is naturally the best and most desirable measure of training efficacy, because it is highest in the hierarchy.

Assumption 2: Each level caused by the previous level.

Both evaluation researchers and practitioners commonly describe the impact of training as being linked in a cause-and-effect chain in which “training leads to reactions, which lead to learning, which lead to changes in job behavior, which lead to results. (Alliger & Janak, 1989; Newstrom, 1987; Santos & Stuart, 2001) However, others, including Kirkpatrick himself, argue against such causal assumptions, recognizing that the cause-effect chain is often difficult to demonstrate, especially with regard to ultimate level evaluations. (Alliger & Janak, 1989; Kirkpatrick, 1956,1996; Santos & Stuart, 2001)

Causality is very difficult to prove or disprove, and negative correlation between Level 1 and Level 2 have in fact been found in classroom research. (Alliger & Janak, 1989) According to Alliger, “There is some reason to believe, then, that reactions may not be expected to cause learning (this is of course more likely to the extent that reaction measures are attitudinally and not behaviorally based.)” According to Alliger and Janak (1989) and Shelton and Alliger (1993), Correlations that do not support Kirkpatrick’s hierarchical model fail to do so because “noise” from other variables such as motivation, transfer context, and trainee attitudes intervenes.

Assumption 3: Levels positively inter-correlated.

In meta-analysis of previous training evaluation studies, Alliger and Janak (1989) found only 12 articles in which attempts had been made to correlate the various levels

advocated by Kirkpatrick. No relationship was found between reaction measures and the other three levels of criteria and relatively small correlations were found between learning and behavior and between behavior and organizational results. Previous studies by Clement as well as Noe and Schmitt also found limited support for Kirkpatrick's hierarchical model of training outcomes. (Alliger and Janak, 1989; Santos & Stuart, 2001) After analyzing the existing literature on the evaluation of training effectiveness, Santos and Stuart (2003) found that they could offer little support for the Kirkpatrick model if used solely as a post-course evaluation that traces a prescriptive cause-and-effect chain from training to organizational performance. Researchers arrive at different conclusions from the apparent lack of causal connections between evaluation levels. For example, Santos and Stuart (2003) conclude, "...the lack of causal connections between different levels of training outcomes implies that evaluation should be done at all levels because each level provides a different kind of evidence." Others come to a different conclusion, noting that, given findings of a lack of causal connection between levels, they have departed from Kirkpatrick. For example, Alliger and Tannenbaum (1997) "...augmented Kirkpatrick's typology to be less coarse, while maintaining a broad enough framework to facilitate generalized understanding of training evaluation results.

Experienced practitioners have rejected Kirkpatrick altogether. According to Phil Rutherford, the academic director of project management studies at the University of New England in Australia, "We've gone outside Kirkpatrick in that we don't look at how training is evaluated and recognized, but how the application of competencies in the workplace influences change – at all levels." Personally, I think training is only ever

evaluated at one level – Level 1. Anything above that is evaluating the application of certain skills and knowledge and, as we know, these could come from anywhere, not just training.” (Rutherford in Abernathy, 1999)

Recent research demonstrates that Kirkpatrick’s model, and the practical application of that model, includes considerable flaws, and should not be regarded as the only model of training evaluation design. However, the purpose here is not to discredit the Kirkpatrick model, as it still provides a well established and effective foundation for thinking about training evaluation, but rather to recognize that modification of the approach is possible and justified and that alternatives must receive consideration. According to Alliger and Tannenbaum (1997), “Although there are problems with Kirkpatrick’s model, just how best to think about training criteria is not clear. Nonetheless, the Kirkpatrick typology remains by far the most influential and prevalent approach among practitioners and, to some extent, researchers. For this reason it can still serve as a point of departure for communicating understandings about training criteria.”

Measuring Kirkpatrick Levels 1 and 2 (Reaction and Learning)

Some scholars regard reaction measures with disdain, and Hubbard (2001) and Tyler (2002) are typical of researchers who see evaluation strategies at Kirkpatrick’s Level 1 (reaction) as fundamentally flawed. According to Hubbard, (2001) “...most training participants are perfectly terrible judges of training effectiveness. They tend very strongly to evaluate a course based not on what they learned, but instead on how they liked the trainer.” From Tyler’s perspective, “Most people think they learned more than they did. It’s good to ask both employee and the supervisor about the amount of

learning that occurred. The trainee cannot possibly immediately know the answer at the course conclusion, and it is often better to ask the supervisor rather than the employee. (Tyler, 2002)

However, Kirkpatrick (1996) reminds us that there exists intent behind reaction-level measures when he says, "It's important to measure participants' reactions in an organized fashion using written comment sheets that have been designed to obtain the desired reactions" and "People must like a training program to obtain the most benefit."

The dichotomy between these two perspectives illustrates an important trend in the literature. While reaction measures get a bad rap, they are important, and need to be measured. According to Fitzgerald and Hammon (1999), "Self-report data deserves its "Smiley Sheet" moniker only when ignored, misrepresented or cavalierly dismissed." Fitzgerald and Hammon explain how practitioners have caused the aversion to reaction measures when they go on to say that "Smiley Sheet" data does not contribute to systematic improvement of efficiencies in the design and delivery of education and training activities – it sits in the corner aspiring to be coal."

Training evaluation researchers point out that the flaws associated with reaction level data frequently derive from poorly designed evaluation forms that are relatively easy to fix. (Hubbard, 2001; Kristiansen, 2004) According to Hubbard (2001), "Too often, the questions are vague or un-quantifiable, or the scoring mechanism is imprecise. The fix is simple: course evaluation forms that have questions that mirror the learning objectives." Recall that, according to training evaluation researchers, the presence of clear training objectives is critical to good training evaluation. (Bee, 2000; Mann, 1997, Pernick, 2001)

Like Kirkpatrick, much of the research literature supports the use of Level 1 (reaction) measures. For example, according to Alliger and Tannenbaum, (1997) “Reactions of trainees are extremely important for several reasons. First trainees may be considered one of the “customers” of training. As such, assessment of their satisfaction with training seems entirely in keeping with most current models of provision of organizational services. Second, whether training is liked could have substantial influence on such distant variables as later training attendance, “word of mouth” advertising, subsequent training funding, and so forth.”

However, according to Kirkpatrick, another post-training state merits evaluation: that being a measure of knowledge acquired, skills improved, or attitudes changed due to training, generally thought of as Kirkpatrick Level 2 or Learning. It should be noted that Kirkpatrick points-out that when principles and facts rather than techniques are taught, it’s more difficult to evaluate learning. (Kirkpatrick, 1996) This is important to note because, as mentioned earlier, the learning targets of *FLL* are a mix of principles, facts and techniques.

Measuring Level 2 (Learning) without testing is problematic, as evaluation at this level is most typically accomplished by testing declarative knowledge. However, the research literature shows that, not only is Kirkpatrick Level 2 important, but that it can be effectively measured using what are typically thought of as Level 1 (Reaction) measures. However, it must be noted that the literature supporting reaction level measurement of Learning (Level 2) envisions a measure specifically designed to do so, not a general predictability of learning extrapolated from typical reaction measures.

For example, Alliger and Tannenbaum (1997) express the perspective that, absent the opportunity for testing, other measures of learning are possible. According to these authors, “Usually, learning as a training criterion is indexed by results of traditional tests of declarative knowledge. Many forms of knowledge assessment however could fit under this label. Hence, we include three sub-categories of learning: learning that is assessed immediately after training (most common), knowledge that is assessed at a later time, and behavior/skill demonstration assessed immediately after training.”

Alliger and Tannenbaum (1997) as well as other researchers assert that reaction measures may successfully predict learning, and that practitioners can effectively evaluate at Level 2 (learning) using self-report, reaction measures if they are designed well. Alliger and Tannenbaum report that a meta-analysis by Kraus “...found higher correlations between attitude and reactions may be in fact more useful, for known theoretical reasons, for predicting on-the-job behavior than are typical measures.”

On the other hand, many training organizations are most interested in the extent to which individuals are motivated to apply the material they have learned. (Warr & Catriona, 1999) The literature shows that one may also measure the post-training motivation of participants to apply the learned material using self-reporting “reaction” instruments.

According to Alliger and Tannenbaum, (1997) “Several researchers have suggested that reaction measures that directly ask trainees about the transferability or

utility of the training should be more closely related to other criteria than would reactions measures that ask about “liking”. In one study, Alliger and Tannenbaum (1997) found that “As expected, utility and combined reactions correlated with on-the-job performance more highly than did affective reactions. Alliger and Tannenbaum also implemented a second kind of reaction variable by asking such questions as “To what degree will this training influence your ability later to perform your job?” “Was this training job relevant?” and “Was the training of practical value?” According to the authors, these questions attempted to ascertain the perceived utility value, or usefulness, of training for subsequent job performance.” (Alliger and Tannenbaum, 1997)

Alliger and Tannenbaum broke “reactions” into two components, “affective” and “utility” reactions. In their study of an open-learning first line manager training program, Warr and Bunce (1995) demonstrated that three forms of reaction are factorially distinct: reported enjoyment of the training, its perceived usefulness, and its perceived difficulty. Warr and Catriona (1999) also describe this perspective well when they report that, in their study of a two-day training course, they found that a well-designed reaction measures could effectively predict learning. According to Warr and Catriona, “An additional measure is therefore taken here at the end of training, recording motivation to transfer course material into a job setting. This is viewed as a reaction to training activities, although such a reaction clearly has a more proactive emphasis than the other three.”

Measuring Kirkpatrick level 3 (behavior).

A surprising trend in the literature is that some practitioners believe that Level 3 (Behavior) is beyond what they can and should measure regarding training

effectiveness. For example, Hubbard (2001) expresses this perspective when he says “How about the third measure – whether the skills acquired are used back in the workplace? This is a particularly emotional question for a lot of trainers, including myself. I suggest that a person is trained to do something when he or she demonstrates that he or she is able to do it. At that point, the trainer’s responsibility has been fulfilled. Whether the skill is applied back at the workplace is the responsibility first of the participant, and then of his or her manager.”

Still, many training organizations show a great interest in measuring the effectiveness of their training in terms of job performance after training. Kirkpatrick (1996) refers to this level of training effectiveness evaluation as “behavior.” Alliger and Janak (1989) refer to this level of evaluation as “transfer.” Regardless of how one characterizes the level of evaluation, the focus is on performance after an individual attends a training program. A review of the extant literature shows that Level 3 (Behavior) can also be effectively measured using self-reporting instruments. According to Warr and Catriona, (1999) as with level two, there is a need to record both prior and subsequent performance, and that is most often undertaken in terms of supervisors’ ratings of key behaviors before and after training. However, sometimes self-reports are obtained if information is unlikely to be available to a boss. (Warr & Catriona, 1999)

However, measuring Kirkpatrick Level 3 (Behavior) is made problematic by the issue of transfer of training back to the workplace. The research has shown that Behavior level measurement is dependent on training transfer, that training transfer is sensitive to the organizational “transfer climate”, and that the transfer climate in an organization is significantly associated with the extent to which learning is actually

applied. (Ripley, 2002; Ruona, Leimbach, Holton & Bates, 2002; Newstrom, 1987; Warr & Catriona, 1999; Santos & Stuart, 2001)

According to Warr and Catriona, (1999) “When supervisors and colleagues encourage and reward the application of course material, training is more likely to yield positive outcomes at work. In assessing level three (job “behavior” in Kirkpatrick’s terms), it is thus important to examine differences between trainees in the nature of their local transfer climate.” According to Newstrom, (1987) “Although trainees find the training useful, the organization may contain substantial impediments to its transfer. For example, previous research has indicated that the most dominant barrier to training transfer is the lack of positive reinforcement by trainees’ supervisors for the new skills being practiced.”

Measuring Kirkpatrick level 4 (results).

Measuring Level 4 (Results) is the much-discussed “Holy Grail” of training evaluation. Some theorists, such as Bee (2000) argue for Level 4 evaluation. According to Bee, (2000) “if learning evaluation is going to be taken seriously, it has to show that it makes a difference to the quality of the learning process and to the business outcomes.” However, there is another perspective, typically expressed by practitioners. Lorne Armstrong, vice president of the Pacific Center for Leadership expresses the opposite perspective when he says, “The real purpose of training is to improve performance of the organization. It is only when we get to Kirkpatrick’s fourth level that we start to measure the results of applying what was learned. Any good manager knows how her work unit is performing and whether the unit’s performance is

improving or not. She is also paid to make some well-informed judgment calls about what's causing the performance to change." (Abernathy, 1999)

On balance, the literature is inconclusive on the notion of measuring results (and even less conclusive on the efficacy of measuring return on investment or ROI.) From the practitioner literature, John A. Zondlo, a project leader in the training and development group at Los Alamos National Laboratory, asserts, "Attempting to measure results is not for the fainthearted. Although measuring training programs in terms of results may be the best way to measure effectiveness, Kirkpatrick himself points out the complications, especially for Level-4 evaluations." (Abernathy, 1999) The ASTD (Sugrue, 2003) reports that only 10% of the company's participating in its Benchmarking Forum (a group of large Fortune 500 companies and public sector organizations) engages in Level 4 evaluation. A review of the training evaluation literature suggests a number of problems with evaluation at Level 4 (Results).

A question important to this project is whether Kirkpatrick Level 4 evaluations are applicable to leadership development training. John Zondlo expresses an important concern with ramifications for this project when he says, "I don't believe that Level 4 is applicable to soft skills training. There are too many variables that can impact performance, other than the training itself." (Abernathy, 1999) Shelton and Alliger (1993) describe the conundrum that Level 4 evaluation of "soft skills" training can present. According to Shelton and Alliger (1993) organizations may avoid Level 4 evaluation simply because collecting and interpreting the data is more difficult and time-consuming than surveying trainees. On the other hand, the authors go on to say "But sometimes organizations try to conduct Level 4 evaluations when they're not

appropriate. A Level 4 assessment may not be the proper evaluation method for training that doesn't affect observable outcomes – for example, training that aims to change only attitudes. Such training isn't likely to show changes in organizational output.”

Organizational Constraints Can Limit Opportunities For Level 4 Evaluation

According to Alliger and Tannenbaum (1997), organizational constraints greatly limit the opportunities for gathering Level 4 data, and most training efforts are incapable of directly affecting results. In addition, Level 4 evaluations require data, and data collection can require a lot of time, effort, and expense. One considerable factor in regard to evaluating the *FLL* training is the longitudinal nature of level 4 evaluations. Brungardt (1997) points-out that “The results method is most often utilized in a longitudinal approach, where participants are surveyed several years after leaving the program.” According to Swierczek and Carmichael, (1985) “The main difficulty in doing longitudinal research is obtaining access to participants.” This concern is of importance to this project, as *FLL* participants come from a highly mobile workforce.

Consequently, the training organization needs to weigh the potential costs against the potential value of the results. (Manthei, 1997; Shelton & Alliger, 1993) Shelton and Alliger (1993) illustrate this point when they assert that “When specifying data for a Level 4 evaluation, you need to consider the amount of time it will take to collect the data and whether such data are available and accessible.” These authors also note that, “Data must be available on a timely basis. If it takes too long to evaluate the training, key people may have transferred to other jobs by the time the results are in.”

As reported earlier, Shelton and Alliger (1993) come to the conclusion that a Level 4 evaluation should be conducted only when it is likely that training will have a detectable effect on business results. According to these researchers, “Unless the training is linked to clear business outcomes, there is no reason for a Level 4 evaluation.”

It Can Be Difficult to Get Meaningful Measures at This Level

According to Warr and Catriona, (1999) “Not only is it difficult to obtain sound measures at this level (especially in comparison with a previous period), but identification of a single training activity as the cause of any changes observed is logically dubious. For those reasons, systematic level four evaluation has hardly ever been reported.” Shelton and Alliger (1993) reinforce and expand this point when they say, “If the training contains material that requires frequent changes, it might be difficult to get meaningful results from a Level 4 evaluation. Changes in the training content during training – and while the Level 4 study is in progress – could affect the validity of

the data.” This concern is particularly pertinent to the *FLL* training, given its history of frequent redevelopment and modification. (McDonald, 2001)

Kirkpatrick (1996) expresses a perspective common to the research literature when he asserts that, “Evaluation in terms of results is proceeding at a slow pace. In a few attempts, researchers have tried to segregate factors other than training that might have had an effect. In most cases, before-and-after measures have been attributed directly to training even though other factors might have been influential.” In addition, Level 4 evaluations should be conducted only when it is likely that training will have a detectable effect on business results. Our ability to identify a single training activity as the cause of any observed changes is dubious, as there are too many variables that can impact performance, other than the training itself. (Abernathy, 1999) This concern is also particularly important to this effort, as the agencies sponsoring *FLL* training are not likely to be able to measure changes in organizational output or business results.

Measuring Return-on-investment (ROI)

According to Long (1999), “Return-on-investment (ROI) is a key financial metric of the value business investments and expenditures. It is a ratio of net benefits over costs expressed as a percentage.” Generally speaking, measuring ROI revolves around two goals, reducing costs and increasing desired impact. (Pernick, 2001)

Measuring return on investment (ROI) represents a perennial topic in the training industry. A search using the keyword “ROI” in the ASTD TrainLit literature database yields 285 hits. The same search of all the documents in the ASTD knowledge base yields 1380 hits. (ASTD, 2004)

However, despite the desire to quantify the effectiveness of training and the tremendous expenditure of energy and effort committed toward that end, the topic involves a great deal of controversy in the field. First, some researchers and practitioners question the very concept of ROI for training. For example, Fred Nickols, former executive director of strategic planning and management services at the Educational Testing Service believes that there is no real ROI for training. According to Nickols, "In strict accounting terms, it is an expense, not an investment. Moreover, it is often an act of faith. And faith is confirmed by the consequences of our deeds, not by financial rewards." "If you want to measure training, determine the impact it is having on people, and through them, on processes, in short, operations. Then, with particular emphasis on supervisors, managers, and executives, determine the perceived value of those impacts. That will tell you what the training is worth." (Nickols in Abernathy, 1999)

From a research perspective, Santos and Stuart (2003) describe the difficulty in describing training efficacy in terms of financial results. According to these authors, "From an analytical and managerial perspective, there are major difficulties in finding measures of training effectiveness in terms of bottom-line results. Indeed, assessing the rate of return from training may be an unrealizable ideal. Companies are not in a position to carry out such an assessment, due to uncertainties over the benefits of training and because of the difficulty in accounting for its true cost."

Santos & Stuart (2003) go on to describe yet another concern, the unintended consequence of financially oriented evaluation causing a reduction in the amount of training offered. They assert that, "Evaluation strategies may, in certain circumstances, even prove self-defeating." Reinforcing this perspective, Ashton and Green (1996) note

“...it may not be worthwhile and could be misleading to draw up a balance of the advantages and disadvantages that can actually be measured. Such an accounting mentality could itself be the cause of low training, if training programs were obliged to demonstrate sufficient measurable return on investment.” Finally, Kirkpatrick (1996) describes the state of research into measuring ROI for training when he asserts that, “Eventually, we may be able to measure human relations training in terms of dollars and cents. But at the present time, our research techniques are not adequate.”

Evaluating Training Using a Mix of Quantitative and Qualitative Methods

The literature suggests a need to evaluate training using both quantitative and qualitative measures. According to Brungardt, (1997) “Although quantitative research is popular among behavioral scientists, evaluation in leadership education should expand its efforts in the qualitative domain. Focus groups and individual interviews, for example, could provide additional insight into program effectiveness.”

Swierczek and Carmichael (1985) describe efforts to evaluate a planning and organization workshop to determine whether training needs were accomplished both quantitatively and qualitatively using participant response measures alone. According to Swierczek and Carmichael, to obtain quantitative data, the evaluators assessed the quality of the workshop’s materials and instruction using a conventional survey administered immediately after the training. Participants were asked to rate sixteen skills on a five-point scale from “never use” to “always use”. In addition, the students were asked two open-ended questions, “What skills did you learn in this workshop?” and, “How will you apply these skills back at work?” The first question measured learning, the second, potential applications. The open-ended format of the questions

was intended to reduce response bias, a phenomenon discussed elsewhere in this paper. According to Swierczek and Carmichael, the design of the survey matched the skills taught in the workshop with skills that the participants identified in their answers to the open-ended questions. Brungardt (1997) described a similar approach mixing quantitative and qualitative measures in his efforts to evaluate a university leadership studies program using a survey including both quantitative statements on a scale and open-ended qualitative questions.

Swierczek and Carmichael (1985) describe the value of combining qualitative and quantitative measures when they say, "The approach presented here combines qualitative participant responses to open-ended questions with the quantitative measurement of training effectiveness gained through statistical analysis of pre- and post-test data. Using both approaches yields a more accurate assessment of the impact of training, enabling the evaluator to provide feedback, assess learning, measure transfer of learning to the work place, and ultimately, improve training programs. These are, after all, the goals of evaluating training, and both quantitative and qualitative analysis are necessary for accurate, comprehensive training evaluation." Swierczek and Carmichael (1985) acknowledge that methodological issues, such as response bias or unreliable measures may have impacted their results, but stress that their quantitative findings reflected a strong positive change compatible with their qualitative results.

Additional Measures & Methods To Crosscheck Self-Reporting Responses

According to Brungardt, (1997) "Evaluation systems should utilize additional measures and methods to cross-check self-reporting responses. Co-workers and other

students can provide verification or disconfirmation of behavioral changes.” Alliger and Tannenbaum (1997) reinforce the perspective that effective evaluation may require more than self-report measures from participants at the conclusion of the training. According to Alliger and Tannenbaum “...trainees are not the only customers of training, nor is the conclusion of training necessarily the optimal time to collect reaction data. Trainees may not always be the most important or best judges of training effectiveness. Future research should evaluate the value of gathering utility-type data from supervisors of training participants and other business leaders. Do their observations confirm the value or utility of the training?” Given these perspectives, this project considered the possibility of surveying supervisors and subordinates in addition to training participants.

Using A Mix of Immediate Measures and Delayed Measures

The literature also lends considerable support to using a combination of immediate measures and delayed measures. For example, Alliger and Tannenbaum, (1997) report that training measures seem highly reliable, but that measures conducted immediately after training have slightly higher reliability than more delayed measures of the same criterion. On the other hand, Alliger and Tannenbaum assert, “By gathering data 1,3, or 6 months after training, trainees will have experienced whether the training was in fact useful, and should be in a better position to judge the utility of the training. Future research should examine when best to collect reaction data.” This dichotomy expressed by Alliger and Tannenbaum suggests that there exists a trade-off between reliability and linking reaction data to transfer/behavior.

Pretest/Posttest Evaluation Design

The evaluation of many training courses is limited to an immediate post-course questionnaire, as is the current case with the *FLL* training. Review of the extant research literature reveals that, to the extent that organizations do evaluate their training, evaluation forms administered after trainees participate in a program are the primary method used. (Sugrue, 2003; Alliger & Tannenbaum, 1997; Saari et al., 1988) However, the research also indicates that used alone, post-course reaction forms are likely inadequate. Commenting on the prevalence of participant reaction forms, Saari et al. (1988) note that the widespread use of reaction strategies runs counter to the ongoing emphasis that psychologists and economists place on the need for systematic training evaluations.

Extensive research into the subject has found that it is generally preferable to measure training outcomes in terms of change from pretest to posttest, rather than merely through posttest only scores. The concept is that before-and-after measurement designs attempt to control biases in evaluation design and are used best for measuring whether or not learning has occurred. (Shelton & Alliger, 1993; Swierczek & Carmichael, 1985; Warr & Catriona, 1999)

However, evaluators using experimental pre-post evaluation designs struggle to isolate the influence of extraneous variables (factors other than the training) on pre- and post-training changes. Reinforcing that particular point, Santos and Stuart (2003) write, "The effectiveness of a training program can also be influenced by events prior to training as well as post-training activities."

Control Groups and Baseline Data

Consequently, experimental evaluation designs and some quasi-experimental designs using pre-post measures employ a control group to control evaluation biases. (Campbell & Stanley, 1963; Caporaso, 1973; Cresswell, 2003; Shelton & Alliger, 1993) According to Shelton and Alliger (1993) “A good way to measure the effects of extraneous factors is to compare the results of a control group with the results of the trainee group. If you can’t use a control group, you should establish some baselines with which to compare post-training results.” Pernick (2001) reinforces both the value of baseline data and the use of control groups when he says “Gather baseline data for participants and comparison groups. Although it is tempting to forgo compiling baseline data, you should resist the temptation. The ability to show before-and-after change is a powerful argument that training has made a difference, especially when you have a control group that did not receive the training. Use a control group and, if possible, randomly assign people to training or no-training conditions. To avoid being accused of turning the organization into a laboratory, refer to the control group as the comparison group”

However, one can use a design that compares pre-training and post-training results without using a control group. By eliminating the control group requirement, evaluators can collect data more easily because the sample size is smaller, and non-control group design is likely to be less expensive than a control group design. However, this approach does not eliminate the possibility that pre- and post-training changes are due to variables other than training. (Shelton & Alliger, 1993 and Alliger and Janak, 1989) Accordingly, Shelton and Alliger (1993) recommend, “If you can’t use

a control group, you should establish some baselines with which to compare post-training results.”

Response-Shift and Controlling Response-Shift Bias

As already mentioned, evaluators routinely use self-report instruments to evaluate training programs. In the interest of controlling evaluation biases, training organizations commonly employ a pretest/posttest evaluation design, particularly when evaluators are trying to identify behavioral change. However, while widespread, this procedure is potentially problematic, including problems of internal validity. (Cantrell, 2003; Eckert, 2000; Mann, 1997; Manthei, 1997; Mezoff, 1981; Rohs, 2002; Rohs, Langone & Coleman, 2001)

While the literature shows that this methodology is both common and well accepted, it is also well known that self-report measures risk a threat to internal validity when evaluators use a pre- and posttest design. One problem encountered when measuring the impact of a training intervention when using a traditional self-reporting pre/post test design is the threat to internal validity known as “response-shift bias.”

According to Cantrell, (2003) “Response-shift occurs when a respondent’s internal metric or frame of reference is changed during the time between the pretest and the posttest, due to the effects of a training program or other intervention. Howard and his colleagues found that when self-reports are used to evaluate change after a training program, participants use an altered set of scale units to classify themselves. This disruption of the internal metric used for the pretest compared to that used for the posttest poses a threat to the internal validity of the instrument, which Howard labeled as a response-shift bias.”

The problem of response-shift bias arises when training organizations use self-report measures to enable participants to judge their own ability using measures taken before and after training. Researchers report in the literature that this response-shift bias has been responsible for disappointing evaluation results obtained by leadership and management trainers, because it conceals the real value of the training effort. (Eckert, 2000; Mann, 1997; Mezzoff, 1981; Rohs, 2002; Rohs, Langone & Coleman, 2001)

According to Rohs (2002) and Rohs, Langone and Coleman (2001), "Several studies have documented the "response shift bias" phenomenon as a source of contamination of self-report measures that result in inaccurate pretest ratings. Consequently, comparisons of pretest with posttest ratings would be confused and distorted, yielding an invalid interpretation of training effectiveness. The consequence is that the training organization, armed only with conventional pre-post data, would be disappointed to see that trainees' skills appeared to remain static or even worsen after training. (Mann, 1997; Manthei, 1997; Mezzoff, 1981; Rohs, 2002; Rohs, Langone & Coleman, 2001)

Ironically, it is the affective change that we desire from training that can confound our ability to effectively evaluate the training provided to cause the intended change. For example, one of the desired results of leadership development training is to change the participants' understanding of leadership principles and skills that typically drive both the objectives of the training and the criteria by which the training is assessed. The paradox of the response-shift phenomenon is that successful training that meets its goal of improving understanding; will alter the participant's perspective when they evaluate

themselves. (Cantrell, 2003; Manthei, 1997; Mezoff, 1981; Rohs, 2002; Rohs & Langone, 2001; Warr & Catriona, 1999)

For example, describing his work specific to leadership development training, Rohs (2002) describes response shift bias and its impact on training evaluation as follows. “For example, a participant might feel at pretest that they are “average” leaders with “average” leadership skills. The program changes their understanding of the skills involved in being a leader; after the workshop they understand that their level of functioning was really below average at the pretests. Suppose they improved their leadership skills as a result of their participation in this leadership development program and moved from below average to average with respect to their new understanding of leadership. Then their pretest and posttest ratings would be average. If we do not consider that these new ratings are based on different understandings of the dimension of leadership, we might erroneously conclude that they had not benefited from the leadership program. Whenever such shifts in understanding occur, conventional self-report pretest-posttest designs are unable to accurately gauge the impacts of these programs.”

When using self-reporting in pre/post tests the evaluator assumes that the metric a respondent uses when completing both the pretest and the posttest remains the same for both points in time. Therefore, response-shift bias poses a threat to the internal validity of the instrument. The literature shows that evaluators cannot assume that the participants’ internal standards for evaluating training will remain constant either during the training experience or from one testing period to the next. (Cantrell, 2003; Rohs, 2002)

According to Rohs (2002), “To compare pretest and posttest scores, a common metric must exist between two sets of scores. In using self-report measures, educators and practitioners assume that a person’s standard for measurement of the dimension being assessed will not change from pretest to posttest. If the standard of measurement were to change, the posttest ratings would reflect this shift in addition to the actual changes in the person’s level of functioning.” Consequently, comparisons of pretest with posttest ratings would be confounded, yielding an invalid interpretation of the effectiveness of the program.

Controlling response-shift bias with retrospective pretests.

Though the research literature points-out potential problems with internal validity when using traditional pre/post evaluation design, studies also show that, the training organization can make efforts to counter response-shift bias, and by doing so, can get more accurate and realistic results describing the effectiveness of their training. (Cantrell, 2003; Ingram, Staten, Cohen, Stewart & deZapien, 2004; Mann, 1997; Manthei, 1997; Rohs, 2002; Rohs, Langone & Coleman, 2001) The literature shows that, to compensate for response-shift bias, numerous researchers and practitioners have employed “retrospective pretests” as an alternative to the conventional pretest. Retrospective pretests involve participants rating themselves again after their posttest rating, but as they perceived themselves to be before completing the training. (Cantrell, 2003; Ingram et al., 2004; Mann, 1997; Manthei, 1997; Nowack, 1991; Pratt et al., 2000; Rohs, 2002; Rohs, Langone & Coleman, 2001; Umble, Upshaw, Orton & Mathews, 2000)

Because the retrospective pretest and the posttest are taken at the same point in time, respondents' internal metric remains the same, and the pretest and the response-shift bias have been removed. The retrospective pretest measure is commonly termed the "then" measure and some researchers and practitioners alternately refer to this type of evaluation as a "post-then" or "then-post" design. (Nowack, 1991; Mann, 1997; Rohs, 2002; Rohs, Langone & Coleman, 2001; Umble et al., 2000) Because the "then" rating and post rating are made in close proximity, it is more likely that both ratings will be made from the same perspective and the evaluation will remain free of response shift bias. Much of the research involving retrospective pretests has been done in the field of psychology and has focused on areas such as assertiveness training and leadership skill acquisition, so the extant research relates to *FLL* training well. (Cantrell, 2003; Manthei, 1997; Mezzoff, 1981; Pratt et al., 2000; Rohs, 2002; Rohs, Langone & Coleman, 2001; Umble et al., 2000)

According to Cantrell (2003), "Respondents are first asked to complete the posttest, which serves as an anchor or benchmark of where they currently perceive themselves to be relative to the construct of interest. Then they are asked to take the test again and answer by recalling how they were functioning just prior to the start of the program. In one study, Manthei (1997) reported that "As has been found in previous educational research comparing the counselors' retrospective pre-test scores with their post-test scores provided a more valid assessment of the impact of training than simply comparing pre-test scores with post-test scores." According to Mann (1997), in most studies, comparative analyses favored the then-post change scores as an objective indicator of change.

Rohs (2002) conducted a study with direct implications for this project, conducting a study to evaluate the effectiveness of a leadership training program and the effects of response-shift bias on outcomes using a self-report measure in a leadership development training program attended by County Extension Agents. Rohs (2002) found that the “then-post” evaluation design provided more significant change data than did the traditional pre/posttest design, indicating a response shift had occurred. According to Rohs, the differences in his evaluation findings suggest that the educational benefit of leadership training such as the program he evaluated may be underestimated when using the traditional pre-post evaluation design.

Rohs, Langone and Coleman (2001) had previously conducted a similar study in a nutrition education program with similar results. According to the authors, “Results from this study support earlier studies and recommendations that when self-report measures are employed to assess attitudinal consistency or behavior intentions across time and when a change in the standard of measurement or level of understanding is anticipated, it is advisable to collect “then” data in place of or in addition to traditional pretest measures.”

The literature suggests that when the fundamental goal of a training program is to change the participant’s understanding of essential concepts, pre/posttest designs may suffer from response shift bias and inaccurately gauge the impact of the training. The retrospective pretest design can avoid the problems of response-shift bias by employing a consistent metric by which participants assess themselves, thereby providing a better assessment of program outcomes. (Cantrell, 2003; Ingram et al.,

2004; Mann, 1997; Manthei, 1997; Pratt et al., 2000; Rohs, 2002; Rohs, Langone & Coleman, 2001)

However, the retrospective pretest design is not without its difficulties. Rohs, Langone and Coleman (2001) report that a disadvantage of the “then/post” is the lack of understanding of its use on the part of the participant completing the self-report measure. According to the authors, directions must be clear and concise to obtain valid data. Consequently, like others, Rohs, Langone and Coleman (2001) recommend that then/post measures be used with caution, and recommend that the training organization take additional steps to collect and integrate other qualitative follow-up measures such as observations documenting change.” Mann (1997) also asserts “The “post-then” measure should be used in addition to, not instead of, the pre-training measure. This would allow the difference between pre- and post-then measures to be examined which may provide information or insights about aspects of the training such as the appropriateness of the level or the accuracy of information given to trainees prior to the event.” (Rohs, 2002; Rohs, Langone & Coleman, 2001; Umble et al., 2000)

According to Cantrell (2003) “It must also be noted that while the use of retrospective pretests may reduce or eliminate response-shift bias, other forms of bias, such as desirability or effort justification, may be present or intensified by this methodology.” There are other means of reducing or eliminating response-shift bias. (Mann, 1997; Manthei, 1997; Pratt et al., 2000) According to Manthei (1997) response shift bias can be mitigated with a short time between pretest and posttest or by physically isolating the respondents. Mann (1997) asserts “Response-shift bias can be reduced by ensuring that participants at training events are given adequate information

prior to the event about what is to be covered by the course.” As previously mentioned, according to Swierczek & Carmichael, (1985) questions in an open-ended format may be used to reduce response bias.

Stakeholder Based Evaluation Strategy

Stakeholder-based evaluation represents an alternative evaluation method worthy of consideration. (Michalski & Cousins, 2001; Nickols, 2003) According to Nickols (2003), “The basic premise of the stakeholder approach is that several groups within an organization have a stake in training conducted for organization members and any effort to design, develop, deliver and evaluate training must factor in the needs and requirements of these stakeholder groups or the results of any subsequent evaluation are bound to fall short of expectations.” Typical training stakeholders include trainees, the trainees’ supervisors, sponsors, training developers, instructors, training managers, vendors, and the training community. (Cousins, 1995; Nickols, 2003)

Though the training evaluation literature shows that stakeholder-based evaluation has been fairly well developed, it is relatively rarely used in training evaluation practice when compared to other methods and essentially remains outside of conventional of training evaluation practice. (Michalski & Cousins, 2001; Nickols, 2003) According to Michalski and Cousins (2001), “The training evaluation and program evaluation literatures have developed largely in parallel, with few points of intersection. The general program evaluation literature includes a steadily developing discussion of stakeholder perspectives in terms of conduct, use, and impact of evaluation. However, despite this work and its emergent application in training evaluation research

stakeholder-based evaluation, for the most part, remains outside the mainstream of training evaluation practice.”

As already discussed, the Kirkpatrick model remains the dominant paradigm for training evaluation, and consequently, most training evaluators continue to evaluate training from a single perspective such as participant reaction, learning or behavior; results in the workplace or return on investment. However, a small, but growing number of researchers and practitioners are showing an interest in stakeholder-based evaluation as an alternative to the Kirkpatrick model. This faction of training evaluation experts argue that given current organizational trends toward collaborative and collective work and limitations of the Kirkpatrick model, opportunities exist to improve training evaluation practice by focusing on obtaining multiple stakeholder perspectives. (Michalski & Cousins, 2001; Nickols, 2003) Michalski and Cousins suggest that a stakeholder-based model of training evaluation may uncover multiple views of training that evaluators should consider. Nickols (2003), referring to the Kirkpatrick model, goes so far as to suggest that the training community is committed to an approach to evaluating training, that after more than 40 years, has failed to capture the commitment and support of other important constituencies...” and that a stakeholder-based approach represents a better way. Nickols (2003) believes that because training is an endeavor with multiple constituencies, adequate training evaluation requires that one assess the satisfaction of all stakeholder groups in regard to what they receive from the training.

Michalski and Cousins (2001) describe a study in which they used qualitative interviewing as their principal means for data collection and analysis. This type of

interviewing for data collection is well-established use for field research in the social sciences, and is emerging as an application for learning-related research in technology-based organizations. (Cresswell, 2003; Michalski & Cousins, 2001) In their study, Michalski and Cousins (2001) conducted semi-structured interviews with 15 people from three stakeholder groups, including training sponsors, trainees, and training providers. The authors audio recorded their interviews, transcribed them verbatim, and coded the transcribed notes for later analysis. According to Michalski and Cousins (2001), by diverging from common training evaluation practice, by involving stakeholders, they were able to show the validity and value of obtaining multiple stakeholder views.

Keeping It Simple

While the literature encourages evaluators to use reliable and psychometrically sound criteria, evaluate training at multiple levels and use multiple methods, techniques and instruments; it is also loaded with advice to “keep it simple,” particularly because, as the level of evaluation goes up the complexities involved increase and complexity can erect a barrier to evaluation. Consequently, evaluators need to strike a balance between effective measures and an evaluation strategy that the organization can be reasonably expected to accomplish. (Abernathy, 1999; Eseryel, 2002; Manthei, 1997; Swierczek & Carmichael, 1985) The challenge, as described by Manthei (1997) is to develop an evaluation design that is methodologically appropriate as well as practical. According to Manthei, “As evaluators, we are often faced with a choice between the strongest design and that which is practical in terms of time, resources, opportunity, and acceptability.”

In the literature, experienced training practitioners express the need for effective, but practical training performance measures well. For example, the organizational development manager of one of the world's largest pharmaceutical firms opines that, "Too often, trainers feel that they need to commission lengthy and often costly research projects to determine the value and effect of their training. I have found that for the majority of clients, simpler is better. A full-blown study that strives to determine statistical correlations to performance metrics or the number of standard deviation gains in tests is fine for the training researcher. For most clients, it is sufficient to explain how competencies from a model are related to performance metrics and that by going through the training, gains in participant performance can be directly tied to the training." (Furando in Abernathy, 1999)

As previously mentioned, as the level of evaluation goes up, the complexities involved increase, which may explain why many organizations only evaluate training at Kirkpatrick's levels 1 and 2, as is discussed elsewhere in this report. (Sugrue, 2003; Eseryel, 2002; Pernick, 2001) According to Pernick (2001), "Data needed at each level influences the kind of program activities one conducts. A well-planned evaluation minimizes the burden by building in practical and unobtrusive ways of gathering data." Abernathy (1999), citing Paul Bernthal, reminds us that training evaluators are not in a laboratory and that evaluators may never be able to isolate true cause-effect outcomes. According to Bernthal, what training evaluators need to do is "Find out what your internal customers want to know about training, then collect the data that will answer those questions. Don't waste your time measuring the wrong thing just to check a box in the evaluation hierarchy."

Along those same lines, Shelton and Alliger (1993) came to the conclusion that a Kirkpatrick Level 4 evaluation measuring results in the workplace should be conducted only when it is likely that training will have a detectable effect on business results. According to these researchers, “Unless the training is linked to clear business outcomes, there is no reason for a Level 4 evaluation.”

Pernick (2001) and Fitzgerald and Hammon (1999) reinforce the keep-it-simple perspective indicating that, because evaluation requires a level of time and expertise, the decision to conduct an evaluation at any level should be based on resource availability. Given organizational resources and demands on those resources, easily obtainable and inexpensive self-report data may represent the only sustainable source of ongoing feedback at more than Levels 1 or 2. Internal evaluations offer convenience and cost-effectiveness, but the obvious concern is whether internal evaluators possess the expertise and resources required to conduct the evaluation. (Abernathy, 1999; Eseryel, 2002; Fitzgerald and Hammon, 1999; Pernick, 2001; Shelton & Alliger, 1993)

Training Evaluation Strategy

Evaluating the effectiveness of leadership programs represents a daunting task, because of the kinds of skills leadership development programs endeavor to teach. Leadership training imparts values, character, influence, competence, credibility and other concepts resting firmly in very subjective terrain. The skills of leadership do not involve discrete behaviors that one can predict in advance, or the employ specific information for which one can easily test. Consequently, we resort to judging whether people’s performance improves in certain areas resulting from the training. In addition, unless one can demonstrate that on-the-job behavioral change results from the training

program, any link between the program and organizational objectives are circumstantial. Without evidence of changes in actual job behavior, one may easily argue that organizational effects may be caused by factors other than effective training.

(Tannenbaum & Woods, 1992)

All training evaluation can prove logistically and methodologically complex, and training organizations must always weigh cost and workload against potential benefit. According to Tannenbaum and Woods (1992) it is impractical to apply large scale, sophisticated evaluations in all instances. Conversely, it is foolish to assume that training works and to ignore evaluation. Kirkpatrick (1996) effectively demonstrates the logistical complexity of systematically obtaining Level 3 evaluation by suggesting a framework for evaluating training programs in terms of behavioral changes as follows:

1. A systematic appraisal should be made of on-the-job performance on a before-and-after basis.
2. The appraisal of performance should be made by one or more of the following parties (the more the better):
 - The participant
 - The participant's superior(s)
 - The participant's subordinates, and/or
 - The participant's peers or other people familiar with the participant's performance.
3. A statistical analysis should be made to compare before-and-after performance and to relate changes to the training program.

4. The post-training appraisal should be made three months or more after the training so that the participants have an opportunity to practice what they have learned. Subsequent appraisals may add validity to the study.
5. A control group (of people who did not receive the training) should be used.

Tannenbaum and Woods (1992) offer a useful alternative guidance to designing a training evaluation strategy. They advise consideration of three important characteristics including:

- The magnitude of the evaluation
- The research design employed
- The training criteria collected.

Each of these criteria, as explained by Tannenbaum and Woods, (1992) is briefly explained below.

Magnitude

An evaluation project can range in size from a small "pilot" test of the program based on a limited number of trainees to an ongoing, large scale evaluation based on thousands of trainees. As the magnitude of the evaluation increases, we can be more confident of the results. Evaluation projects that include multiple groups and locations will give a better indication of how well the training may be generalized. However, larger evaluations are clearly more complex and significantly more costly.

Research Design

Research designs range from simplistic "pre-experimental" designs, through increasingly sophisticated "quasi-experimental," and "experimental" designs. Two common pre-experimental designs used to evaluate training include the case study (i.e., examine one set of trainees without comparing them to any other group of employees), and the post-training comparison of two groups of employees (one group which received training and one which did not).

In the latter design, a failure to determine just how different the two groups are prior to training makes it impossible to determine the impact of the training. In general, it is difficult to interpret pre-experimental evaluations.

In an experimental training evaluation, employees are randomly selected and assigned to either a training or "control" group. The key feature of these designs is the randomization process, which is supposed to yield equivalent groups for comparison purposes. In addition, experimental training evaluations usually involve the collection of both pre-and post-training data from the trained and non-trained employees.

Quasi-experiments fall in the middle of the continuum. They are more rigorous than pre-experimental evaluations, but unlike true experiments, they do not use random assignment. Instead they use a variety of methods to enhance interpretability, including comparing pre-and post-training data from trainees to similar data collected from employees who are waiting to attend the training, or collecting pre-and post-training measures at several points in time to establish clearer base lines for comparison.

Training Criteria

Tannenbaum and Woods (1992) identify five types of training criteria (reactions, attitude change, learning, behavior and results) based on the work of Kirkpatrick, supplemented by Tannenbaum, Mathieu, Salas, and Cannon-Bowers (1990). Each criterion addresses different questions as follows:

1. Reactions--How well did the trainees like the program? What were their feelings about the training? Did they feel it was useful?
2. Attitude Change--Did the training modify trainees' beliefs, convictions, feelings, dispositions, or attitudes? For example, do they now think like a team? Are they aware of the importance of quality? Are they receptive to diversity in the workplace? Are they more confident that they can perform their jobs effectively?
3. Learning--What principles, facts, and techniques were learned, understood, and absorbed by the trainees? Can they recite new information? Can they verbalize new strategies, concepts, or approaches to performing a task?
4. Behavior--What changes in job behavior resulted from the training? Do trainees apply what they learned during training on the job? Do they perform their jobs differently?
5. Results--What were the tangible results of the program in terms of reduced costs, enhanced quality, or greater productivity?

As one moves upward from reactions to results, evaluation projects that measure the higher-level outcomes (such as behavior or results) should yield more compelling data about the organizational impact of training. For example, an evaluation project that shows increased productivity resulting from training will likely carry more weight than

one merely showing that trainees liked the course or had changed their attitude. It is usually more difficult, however, to collect data as you move toward the top of the scale.

Training criteria at the learning level or below can often be collected during training administration. Behavioral measures and measures of results require on-the-job, follow-up activities that make data gathering more difficult and that may elicit resistance. For simplicity, Tannenbaum and Woods (1992) present each training criteria as a distinct category, but like Kirkpatrick, (1996) the authors assert that it is often desirable to assess training effectiveness using multiple criteria.

Tannenbaum and Woods (1992) warn of the importance of any decision to collect particular training criteria measures, use a specific type of research design, and pursue a certain magnitude of evaluation. They use the phrase "evaluation strategy" to refer to the combination of criteria, design, and magnitude chosen. The advantages, disadvantages, and use of several evaluation strategies are discussed below.

No evaluation strategy.

One strategy is to forego any evaluation. The training is delivered but no effort is made to collect any information about whether the training was effective. The advantage to this strategy is that there are no cost or time commitments involved and it is totally unobtrusive, in that it does not divert the attention of the training staff or impose at all on training participants. The obvious disadvantage is that it provides no information with which to revise the training, plan for the future, or make decisions about training.

Reaction-only strategy.

Reaction-only evaluations employ a pre-experimental design. By definition, reaction data cannot be collected before training, so a comparison of pre-training to

post-training change is impossible. Furthermore, reactions cannot be collected from employees who have not been trained, so the use of a control group is impossible.

In a reaction-only strategy, trainees are asked at the conclusion of training what they thought about the training. Did they like it? Was it useful? Was the instructor effective? In this approach, no attempt is made to determine whether trainees learned anything or subsequently performed their job differently because they participated in the training. However, reaction-only evaluations continue to be the most prevalent form of training evaluation. The advantage to this approach is that it requires very little time and effort, and is an accepted practice which training participants will nearly always tolerate. It is relatively unobtrusive, and it gives participants the feeling that they have some input into subsequent training. In addition, positive reaction data can be an indication of trainee acceptance. Although a high level of acceptance does not imply that training objectives have been met, low acceptance can indicate a potential problem. Furthermore, to the extent that the evaluation form requests specific information about particular aspects of the training reaction data may provide some formative feedback to facilitate program revisions.

The primary disadvantage to the reaction-only strategy is that it provides little useful information with which to plan or to facilitate effective decision-making. While somewhat contradicted by other scholars, according to Tannenbaum and Woods, a growing body of research shows that trainee reactions may not be related to subsequent learning, behavior, and results and that, at a minimum, the relationship between reactions and other training criteria is not a direct one. Thus, employee reactions cannot be expected to accurately assess the effectiveness of most training

programs whose objectives include trainee learning and behavior changes on the job. Furthermore, reaction data are often insufficient to support subsequent planning decisions. Finally, although employee reactions are used quite frequently, they are not very persuasive during lean financial times. When experiencing financial pressures, most decision makers will be more impressed with an indication that a program taught employees new skills, changed employee's behaviors, and contributed to improved performance than with the knowledge that trainees liked the training.

Ideally, one would want to assess Level 1 (reaction), Level 2 (learning) and Level 3 (behavior) using reaction measures alone. Researchers in the field acknowledge that desire for such an approach. For example, according to Alliger and Tannenbaum, (1997) "Because reaction measures are the easiest to collect, it would be ideal if they could be used as surrogate measures of learning and transfer. Interestingly, the magnitude of the relationship between training satisfaction and job performance is about the same as that typically exhibited between job satisfaction and job performance." However, other research findings suggest that the potential success of such an approach appears less certain.

However, according to Alliger and Tannenbaum, (1997) "Regardless of their relationship with other measures, from a pragmatic perspective, trainee reactions are important. Overall results of this meta-analysis support Alliger and Janak's (1989) findings that reaction measures cannot be used as surrogates of other measures. In particular, affective reactions are unrelated to other indicators – liking does not equate to learning or performing. As previously mentioned, if the purpose for collecting reaction measures is to predict or indicate transfer it would be best to ask utility-oriented

questions. If both utility and affective questions are asked, it appears that these should be treated separately, with the utility questions being used to provide the better estimate of potential transfer.”

Basic evaluation strategy.

These are similar to the reaction-only strategy but are based on different criteria. One advantage of this strategy is that it can yield more compelling results than a reaction-only study. Even without a pre-measure or a control group, a well designed post-training assessment of knowledge or performance may be useful for determining whether the trainees know "enough" about a topic and whether they perform "acceptably" after training. However, this strategy tells us little about the impact or usefulness of training; it is difficult or impossible to determine whether the training enhanced, impaired, or had no effect on the knowledge and performance of trainees. A basic strategy, although appealing in its simplicity, can easily yield misleading results.

Intermediate evaluation strategy.

Intermediate evaluations have greater sophistication and fewer threats to validity than the reaction-only and basic strategies, but they do not incorporate some of the controls seen in an advanced, experimentally based strategy. Consequently, there are many advantages to an intermediate evaluation strategy. This approach yields more useful and valid results than a reaction-only strategy and, hence, provides better information for planning purposes. Although an intermediate evaluation strategy is less rigorous than an advanced, experimental strategy, it typically yields information that is interpretable and defensible. This strategy is also less obtrusive than most advanced evaluations and usually fits more readily with organizational realities.

The disadvantage of an intermediate strategy is that it is more obtrusive than a reaction-only strategy and less rigorous than an advanced strategy. In an organization where evaluation is considered a nuisance, it is unlikely that intermediate evaluations can be substituted for reaction-only evaluations. At the other extreme, hard-core experimentalists may not accept the findings from an intermediate evaluation, suggesting that the limitations of the design interfere with the interpretation of the results.

Intermediate evaluation efforts vary in their complexity and obtrusiveness. Those that use multiple criteria, collect information from many trainees, and use designs that are more sophisticated and analyses will allow for greater confidence in the results but will elicit more resistance. Thus, trade-offs exist within this category as well.

Advanced evaluation strategy .

Experimental research designs have been strongly advocated for their ability to imply causality. When conditions allow for the use of experimental designs they yield the greatest confidence in interpretation. There are, however, several obstacles or disadvantages to using an advanced evaluation strategy in many organizational contexts. It is rare that an organization will be willing to assign trainees to training on a random basis. Employees are usually assigned to training to prepare for a recently established or forthcoming change in responsibilities, to remedy a performance problem, or as a reward for past-accomplishments. Each of these reasons interferes with sending people to training at random. Similarly, most organizations are unwilling to establish a true control group and to collect data from the "untrained" group.

Even if an organization chooses to use random assignment and a true control group, advanced evaluations can have problems. A control group is only useful if the employees receiving the training do not communicate with the employees in the control group. Although this is easy to maintain in an academic study, it is difficult, if not impossible, in many organizational settings.

Theoretical Considerations: Training Evaluation Design

Reliability of training evaluation measures.

The literature makes clear that training organizations must take care to assure that they are employing reliable evaluations. Newstrom (1987) offers seven tips for avoiding evaluation anomalies that have proven beneficial as guidelines for designing evaluation measures for the *FLL* training.

1. Use clear statements that respondents are likely to perceive similarly. Focus on concrete trainer behaviors and specific facets of the training environment.
2. Anchor various points on the rating scale with descriptive phrases to anchor various points on the rating scale. According to Newstrom, "Confusion is likely where you have not used anchor phrases..."because trainees frequently misinterpret the direction of scales."
3. Always ask respondents why they answered as they did. According to Newstrom, "This allows you to better assess whether the trainees simply misunderstood the question or the response scale or whether some had an underlying bias that could account for their reaction."
4. Tell respondents in advance if they are to find similar items and instruct them to address each independently.

5. Use the smallest number of items necessary to tap adequately the respondents' perceptions of each major factor. Newstrom advises, "Remember that trainee desire to complete any reaction instrument is inversely proportional to its length!"
6. Use an adequate sample.
7. Decide up-front which criterion – cost-benefit analysis or statistical significance- will be applied to determine the program's success.

Newstrom (1987) also offers seven means of interpreting reliability when designing and testing evaluation measures and instruments.

1. Interpret degree of agreement on a single criterion within a group of respondents. Ideally, a high proportion of the trainees should respond similarly to the program, the trainer, or the training environment. Example, "a bar chart of reaction survey response distribution should show a dominance of responses in one or two adjacent categories."
2. The degree to which the participants respond similarly across related items comprising a single scale. According to Newstrom, "When two or more questions address related issues, the group's response – whether positive or negative- to one item should closely approximate its response to the related item."
3. The degree to which the participants respond similarly to the same items at two different points in time. Statistically referred to as test-retest reliability, group's evaluation of a program remains relatively constant across two or more assessment periods. According to Newstrom, "The retest period could begin with the inclusion of a comparable question late in the basic questionnaire, as an internal and immediate

check on reliability, or could be part of a follow-up evaluation weeks or months after the initial assessment.

4. The degree to which different groups exposed to the same program and trainer concur in their assessment. According to Newstrom, "Assuming that neither the program nor its presentation actually differed, you can expect to find the reactions of one group similar to those of another comparable group."
5. Close agreement among conclusions about a program's effectiveness across two or more of the traditional levels of evaluation (Kirkpatrick, 1996.) According to Newstrom, "If the trainees have a positive reaction to a program, you can logically conclude that they will also learn and retain the material presented." While this assumption has been called into question by other researchers, based on the literature, one can still argue for the value of testing at multiple levels and looking for agreement across levels.
6. The degree to which those who respond to an end-of-program questionnaire fairly represent the feelings and reactions of all participants. One solution to this concern would be to require 100% compliance to evaluation measures.
7. To what degree does the statistical effectiveness of a program converge with its practical significance? According to Newstrom, "...you can place too much emphasis on an assessment approach under the false assumption that statistical and practical significance mean virtually the same thing."

Evaluation Instruments

According to Tyler (2002) “The format of the questionnaire is crucial in eliciting accurate responses. Experts give the following advice.”

- Place easy questions first. Start with simple, uncontroversial questions, and move to more complex questions.
- Keep the evaluation short. Long evaluations will reduce the response rate.
- Avoid unnecessary numbers. Don't number (for identification) the questionnaire
- If there are summary questions, always put them at the end of a block of related questions.
- Avoid mixing rating and ranking questions
- Use a rating scale instead of words. Have two poles labeled with extremes and just numbers in between.

A review of the literature reveals that relatively few researchers publish the instruments that they used in the course of their study, and consequently the existing literature offers little in the way of effective examples. A few authors have published their instruments and they have proven helpful in designing instruments for evaluating *FLL* training. (Brungardt, 1997; Cantrell, 2003) In addition, the American Society for Training and Development publishes a number of publications demonstrating existing effective evaluation instruments and/or offering valuable advice on instrument design. (Conway, 2004; Fisk, 1991; Kristiansen, 2004; Stadius, 1999)

Chapter III

Methodology

Overview

This internship project developed a method for collecting and analyzing data to evaluate the *FLL* training, and ultimately, the NWCG leadership initiative. Initially, efforts were made to examine and synthesize an existing body of knowledge, determine key or core competencies of the *FLL* training program, identify seemingly appropriate training evaluation methods, make a critical and analytic appraisal of those methods, and build a case for a preferred evaluation method for assessing the *FLL* training.

Once a preferred evaluation method was established, data collection instruments were developed and quantifiable and statistical measures created under the supervision of Fort Hays State University faculty. The author conducted a small-scale test of the evaluation instruments and measures, including an evaluation of results that involved statistical analysis as appropriate. The evaluation method and associated instruments were revised as necessary based on this evaluation. During the final steps of the project, the author developed procedural guidance for the client organization. Ultimately, this paper, resulting evaluation instruments, and procedural guidance were delivered to the University and the client organization with the intent of making the resulting evaluation method operational during the 2005 training cycle. This project was completed using a twelve-step research process as follows.

Step 1: Literature Search: Fireline Leadership

This study reviewed a body of literature to comprehensively examine research into the *FLL* training program that others had done, and literature describing what others had said about the training. Data was located and retrieved by searching on-line electronic websites, databases and journals. Specifically, the search used the descriptors “fireline AND leadership” as well as “fireline AND leadership AND training” to locate primary sources. Knowing that little had been written on the subject, this research step primarily involved a targeted search of the websites of associations, societies and institutes committed to wildland fire and leadership in wildland fire industry to examine research-based papers posted at them. Specifically, primary sources were obtained from the websites of the National Interagency Fire Center, the International Association of Wildland Fire, and Mission-Centered Solutions, Inc.

The same search descriptors were used to search for primary sources within the EBSCOHost, OCLC: FirstSearch, ABI/Inform, WilsonSelectPlus and PsychInfo databases, as well as the research databases of the Library of Congress and the American Society for Training and Development (ASTD.) In addition, a general search using Internet search engines was conducted using the same search term descriptors employed for the searches of the online databases and libraries. In total, this step acquired six primary sources of value. This literature search also involved retrieving select print literature, specifically an assessment of the effectiveness of *FLL* and contracted delivery familiar to the author. (Hillman and McDonald, 2003)

Step 2: Literature Search: Training Evaluation

This study also reviewed a body of literature to comprehensively examine research that others have done in the field of training evaluation. The same general process used to examine the *FLL* literature was used. First, data was located and retrieved by searching on-line, electronic websites, databases and journals. Specifically, the search used the search descriptor “training AND evaluation” to locate primary sources within the EBSCOHost, OCLC: FirstSearch, ABI/Inform, WilsonSelectPlus and PsychInfo databases, as well as the research databases of the Library of Congress and the American Society for Training and Development (ASTD.) In addition, a general search using Internet search engines was conducted using the same search term descriptor employed for the searches of the online databases and libraries. In total, this step acquired 33 primary sources of value.

This step also involved retrieving select print literature pertaining to training evaluation, specifically 11 articles and reference texts familiar to the author or discovered during on-line research. All were selected for their relationship to training evaluation, in some cases specific to leadership development training, as well as for their research basis.

Step 3: Evaluate Data Sources

The third research step involved evaluating the information sources retrieved using five criteria. First, each source was assessed to determine that its purpose and results are adequately oriented to either *FLL* or training evaluation. When assessing training evaluation sources, particular priority was placed on sources pertaining to the evaluation of leadership development training. Next, each retrieved source was

evaluated to determine that it was sufficiently research based. Third, when possible, the credibility of the information source was evaluated based on the credentials or reputation of the author. Once determined that the source of information was sufficiently oriented to training evaluation or *FLL*, adequately research based, and produced by a credible author, the source's methods were evaluated, within the abilities of this study's author, to determine that its sample is adequate in terms of size and "representativeness" as well as for the accuracy and relevance of the methods used. During this step, efforts were made to recognize the theoretical orientation, bias or perspective of the author, setting the stage for systematically organizing the works.

The collected data were evaluated to determine their adequacy to purpose, leading to targeted electronic searches to fill-in research gaps as necessary. Of note, when it appeared that the retrospective pretest or "post-then" methodology might have merit, additional electronic searches were conducted using "retrospective AND pretest", "post-then AND retrospective pretest" and "response-shift AND bias" as descriptors. These additional efforts netted eight primary sources of value.

In all, over 3,000 pages of material contained in 58 sources were located, assessed and retained, all in primary sources.

Step 4: Determine Key or Core Competencies of the *FLL* Training Program

An important accomplishment of this study was to determine the key or core competencies of the *FLL* training program, by applying the findings of the literature search. In this step, the findings of the literature search, particularly the identification and validation of key competencies, were supplemented by validation and consultation with the Leadership Committee.

Step 5: Select/Design Evaluation Strategy

Following the literature search and identification of core competencies, the author identified seemingly appropriate methods, made a critical and analytic appraisal of those methods, and built a case for a preferred evaluation strategy for assessing the *FLL* training. For the purposes of this project, "evaluation strategy" refers to the combination of criteria, design, and magnitude chosen, as described by Tannenbaum and Woods. (1992), and an assessment, employing factors described by them appears in Chapter IV - Presentation of Findings elsewhere in this paper.

Based on this assessment and other findings of the literature search, potential evaluation strategies and methods were evaluated against the following criteria:

- An intermediate, quasi-experimental design
- The ability to match what is measured with established learning objectives
- The ability to evaluate *FLL* Training at Kirkpatrick's Levels 1 – 3
- Simplicity
- A relatively low-cost method
- A viable alternative to the common pretest/posttest evaluation design
- Able to evaluate training using multiple methods, techniques and instruments

Step 6: Develop Data Collection Instruments

The literature makes clear that training organizations must take care to assure that they are employing reliable evaluations, and care was taken to employ design criteria discovered during the literature search. The project considered both paper-pencil and on-line data collection instruments. However, a preference for paper-pencil instruments was established early in the project upon consultation with the client organization. Particular attention was paid to evaluation measures that strike the best balance between effective data collection (encouraging 100 percent response), limited obtrusiveness, and maximum confidentiality for the respondent.

Step 7: Develop Quantifiable/Statistical Measures

The project's intent was to determine a strategy focused on quantifiable, supportable and, where appropriate, statistically defensible measures. Statistical measures used for data analysis were developed under the supervision of Fort Hays State University faculty.

In the interest of producing a highly reliable training evaluation measures, six measures of reliability were employed.

1. The degree of agreement on a single criterion within a group of respondents.

Ideally, a high proportion of the trainees should respond similarly to the program, the trainer, or the training environment.

2. The degree to which the participants respond similarly across related items comprising a single scale. When two or more questions address related issues, the group's response - whether positive or negative - to one item should closely approximate its response to the related item.

3. The degree to which different groups exposed to the same program and trainer concur in their assessment.
4. Close agreement among conclusions about a program's effectiveness across two or more of the traditional levels of evaluation.
5. The degree to which those who respond to the end-of-program questionnaire fairly represent the feelings and reactions of all participants. The project addressed this criteria by obtaining 100% compliance on immediate evaluation measures and a by statistical sampling on others.
6. The degree to which the statistical effectiveness of a program converges with its practical significance.

Step 8: Small-scale Test

Once data collection instruments, quantifiable/statistical measures, and testing procedures were established, a small-scale test was conducted on a sample of *FLL* training participants and supervisors of those participants in late 2004 and early 2005. The purpose of this test was to evaluate the utility and the validity of the instruments. In this stage of the project, three instruments were tested; an attitude survey instrument administered immediately after the training program, a survey for past participants in the *FLL* training, and a survey for the supervisors of past participants in the training. Both delayed measures employed the "post-then" methodology, explained elsewhere in this paper, and were conducted approximately six months post-course. For both training participants and supervisors, survey responses were used only for the purposes of this project and all participant responses were kept confidential.

Sample and Limitations of the Study

One hundred and fifteen course participants from five *FLL* courses completed the attitude survey instrument immediately following training sessions during the fall and winter of 2004/2005. However, surveying past *FLL* course participants and the supervisors of past training participants proved more problematic. The client organization lacks a centralized database of the approximately 6,000 people who have completed the *FLL* training, a fact that presented a challenge to this project. The author developed a form to collect contact data from course participants during the 2004 training season. The form was intended to collect contact data for both the participant and their supervisor. However, the project's timing, in relation to the 2004 *FLL* training cycle, allowed only a limited opportunity to develop a sample population complete with contact data specifically for use in this project. Specifically, the form was only distributed in two courses at the end of the training season, netting only 49 participants and 35 supervisors.

With the assistance of the client organization and the contracted provider of the training, the author developed additional contact data, eventually building a sample of 120 past participants of *FLL* courses, selected at random. To complete the database for the participant sample, contact information for the remaining 71 participants was developed by recreating contact information from course rosters and other sources. The project's timing, in relation to the 2004 *FLL* training cycle, created an even more limited opportunity to develop a sample population of supervisors. The opportunity did not exist to supplement the supervisor sample as was done with the participant sample, because supervisor data is not routinely collected on course rosters. Consequently, the

supervisor sample included only 35 persons. Due to the make-up of one of the courses from which the population samples were developed, three people appeared in both the participant and supervisor samples.

Table 1. Sample Sizes for Treatment Groups

Instrument	Treatment Group		
	Current Participants	Past Participants	Supervisors
Immediate Post-Course Attitude Survey	n = 115		
Delayed Post-Course Survey - Participant		n = 120	
Delayed Post-Course Survey - Supervisor			n = 35

In the test, response rates to the delayed instruments administered to past-participants of the *FLL* training and supervisors of past-participants were inadequate to generate the desired statistical sample. In all, 55 training participants (of 120) and 22 supervisors (of 35) responded with completed surveys.

Table 2. Survey Response by Treatment Group

Instrument	Treatment Group		
	Current Participants	Past Participants	Supervisors
Immediate Post-Course Attitude Survey	n = 115		
Delayed Post-Course Survey - Participant		n = 55	
Delayed Post-Course Survey - Supervisor			n = 22

A typographical error in the supervisor instrument required that one element (of 37) be eliminated for the purposes of statistical evaluation. However, the error does not appear to have had an effect on the overall test of the instrument or to have influenced

adjacent elements. The results of the small-scale test of the instruments are reported in the Presentation of Findings (Chapter IV) elsewhere in this paper.

A fourth element of the evaluation strategy (qualitative focus groups) was not tested as part of the small-scale test, because this element of the strategy employs standard, widely accepted focus group protocols described elsewhere.

Step 9: Revise Strategy/Instruments as Necessary

Once the operational test was completed, including statistical analysis as appropriate, results were evaluated. The evaluation strategy and associated instruments were revised as necessary based on this evaluation.

Step 10: Develop Procedure

The resulting evaluation strategy includes procedural guidance, essentially a user's manual, attached to this paper as Appendix A.

Step 11: Deliver Test Results Procedure and Instruments

The resulting evaluation method is documented in this report to be presented to Dr. Curtis Brungardt of Fort Hays State University and the Leadership Committee of the NWCG Training Working Team.

Step 12: Operational During the 2005 Training Cycle

Chapter IV

Presentation of Findings

Introduction

Assessment findings are presented here in four sections. First, the results of an assessment of the *FLL* training using eight organizational factors described by Tannenbaum and Woods (1992) are presented. The author used this assessment to help determine an appropriate evaluation strategy for the *FLL* training. Second, the immediate post-course attitude survey is examined. Third, the survey administered to past participants in the *FLL* training is analyzed. Finally, the results of the evaluation of the survey administered to the supervisors of past participants in the *FLL* training are reported. The tests of the three data collection instruments are intended to determine their utility and validity.

Evaluation Strategy

Tannenbaum and Woods (1992) offer an effective approach to crafting an evaluation strategy by considering key organizational factors including change potential, importance/criticality, scale, purpose and nature of the training, organizational culture, expertise, cost, and timeframe. An assessment of these factors pertaining to evaluation of the *FLL* training program appears below.

Change Potential

There are clearly opportunities to change the *FLL* training program, as it has been modified annually since its inception, and undergoes nearly constant change. In addition, training does not represent the only organizational strategy for leadership development and it is conceivable that the sponsor organizations may someday

consider alternative methods to supplement or replace this training. Financial pressures result in continual calls for cost justification. Ultimately, change potential exists and for these reasons noted, investing time and effort in evaluation makes sense.

Importance/Criticality

The *FLL* training program represents a key component in an important strategic initiative with significant implications to firefighter safety, effectiveness and productivity. As such, it warrants a thorough evaluation to ensure that its objectives are accomplished and to allow for continuous improvement. As mentioned, *FLL* training represents a vital part of a critical strategic initiative, but faces pressure for justification. The implications of erroneous evaluation results are serious, and argue for the use of at least an intermediate design to evaluate the *FLL* training.

Scale

The *FLL* program is a large, ongoing effort that will ultimately train many employees and is part of an overall strategy to affect cultural change. Consequently, evaluation assumes significant importance. A reaction-only strategy will not provide the information needed to determine if this continual expense is a worthwhile investment for the sponsor agencies. Ongoing training efforts require a more systematic evaluation strategy. In addition, the large population of *FLL* “graduates,” multiple classes, and larger sample sizes increase the feasibility of conducting a large-scale evaluation and using a more sophisticated design and analyses.

Purpose and Nature of the Training

In general, it is more difficult to evaluate leadership, supervisory and human relations-oriented training than it is to assess other types of training with readily

measurable or observable outcomes. Leadership training imparts "softer," less objective outcomes. Nonetheless, the *FLL* training program represents a key part of an important strategic initiative and warrants careful evaluation. Typically, leadership training represents a prime opportunity for quasi-experimental evaluation designs.

Organizational Culture

In some organizations cost justification and follow-up evaluation is a way of life. These organizations typically reject reaction-only evaluations as insufficient. In organizations that have a strong orientation toward evaluation, it is particularly important to use evaluation designs that will generate defensible, quantitative indicators of learning, behaviors, and/or results.

Other organizations give greater weight to intuition and quick action. Evaluation is less important in these organizations and, may even be discouraged. In these organizations, elaborate evaluations often run into significant resistance. The agencies sponsoring the *FLL* training program lie at neither pole.

Neither NWCG agencies nor the NWCG training curriculums are particularly strongly oriented to training evaluation. The NWCG curriculum engages in little evaluation beyond Level 1 (Reaction) evaluation immediately following training. However, the agencies are not particularly evaluation averse. At the bottom line lies the need gather useful, supportable information regarding a strategically important and ongoing training program. The challenge will be to collect this information in the least obtrusive method possible.

Expertise

Sophisticated designs and analyses require greater evaluation expertise. The NWCG Leadership Committee must assess whether they and their sponsor agencies have the capabilities to sustain a complex evaluation method over time. The experience of the small-scale test conducted as part of this study suggests that they do not, given their organizational resources, demands on those resources, and the mobile and temporal nature of much of their workforce. The literature shows that many training organizations report that one reason that reaction-only designs are used so frequently is that other designs are beyond the expertise and resources of their training departments.

Cost

As a large scale, expensive training program, the *FLL* training deserves quality evaluation. The potential exists that the evaluation may provide feedback for program improvements or as the mechanism for a decision to continue or discontinue the program. Evaluation efforts cost money, both in the form of direct and indirect expenses. For example, collecting information from supervisors, subordinates and peers takes their time as well as the evaluator's time. Without perspective on the legitimate needs for evaluation, costs associated with conducting an evaluation could approach the cost of conducting the training.

Finally, the NWCG Leadership Committee must consider when it needs evaluation results to best support decisions about the *FLL* training. A shorter, relatively simple, small sample evaluation effort may represent the best approach. While such a study elicits less confidence than a more comprehensive study, it may provide more timely information and better long-term applicability than a large-scale effort.

Immediate Post-Course Attitude Survey

Training participants' attitudes toward their experience in the FLL training were measured with a post-course attitude survey administered immediately at the end of the training. The self-reporting instrument uses a five point Likert scale anchored by descriptive phrases (5 = Agree, 1= Disagree) on 11 quantifiable questions. The instrument also employs six qualitative questions to be answered with written comments. To effectively measure learning, in addition to reactions to the training, the instrument has been specifically designed to inquire about the transferability/utility of the training by asking "utility" questions that estimate the potential for the trainee to transfer what they learned in training to the workplace. Table 3. reports mean and standard deviation scores for the test of this instrument. Mean scores ranged from 4.66 to 4.99, which represents a uniform reporting range skewed far to the "Agree" end of the measurement scale.

Table 3. Immediate Post-course Attitude Survey (Scoreable Portion)

Section	Question	M	SD
Overall Rating	This training was worth attending	4.90	.33
Training Design	The topics were well organized and understandable	4.74	.52
	The pace of the training was appropriate for the material covered	4.66	.58
	The student guides and course materials contributed to my learning	4.75	.53
Instructors	The instructors performed well overall	4.89	.34
	The instructors are knowledgeable about the subject matter	4.92	.28
Training Exercises	I found the classroom exercises valuable in helping me understand the concepts presented and how to apply them	4.67	.60
	I found the field exercises valuable in helping me understand the concepts presented and how to apply them	4.67	.62
Training Application	The content presented applies to my current needs and job duties	4.69	.63
	I will apply what I learned in this training to my job	4.84	.44
	This training will help me be more effective in my current job now	4.76	.54

5-Agree, 1-Disagree

Participants' Experience With The Instrument

Less than 1% of participants ($n=1$) experienced difficulty interpreting the direction of the scale on the "scoreable" portion of the instrument. Twenty-eight percent of participants ($n=31$) answered the twelfth question ("what topics would you have liked to spend more or less time on?") in such a way as to make it difficult for the evaluator to know the intent of the participant's response. Eighteen percent of participants ($n=20$) experienced similar difficulty with the thirteenth question ("What did the instructors do that worked well and what might you suggest to improve their effectiveness?").

Delayed Post-Course Survey Instruments

The strategy described in this paper evaluates the *FLL Training At Kirkpatrick Level 3 (Behavior)* via a retrospective pretest administered to both trainees and supervisors approximately six-months post-course. The self-reporting instruments use a five point Likert scale anchored by descriptive phrases (5 = Agree, 1= Disagree). The survey administered to participants contains 36 questions and the survey administered to supervisors of participants contains 37. In both cases, the questions address the accepted learning targets of the training.

Testing Reliability

Effective survey instruments must be both valid and reliable, and the delayed post-course survey instruments were analyzed employing Chronbach's Alpha, a measurement of internal reliability. For the purposes of this test of reliability, the two forms were analyzed as separate instruments and Alpha coefficients were determined, not by individual learning target, but by the groupings established by the NWCG Leadership Committee (Communicating Vision and Intent, Application of Leadership Skills, Teambuilding, Detecting and Mitigating Error, Managing Stress and Other Human Factors, Decision Making, and Ethics).

Delayed Post-Course Survey Instrument Administered to Training Participants

When completing the delayed post-course survey, the participant completes the survey in two parts, on two separate forms. On Form 1 - Post-Training Skills Evaluation, the respondent is asked to evaluate the skills and knowledge that they feel they have now, after taking the *FLL Training*. On Form 2 - Pre -Training Skills

Evaluation the participant is asked, using the same 36 questions, to evaluate the level of skill and knowledge they had before participating in the training.

Table 4 displays reliability coefficients (Alpha) scores for the test of the Post-Training Skills Evaluation. With one exception, Alpha scores ranged from .6575 to .8778, indicating high reliability of the instrument, based on this limited sample ($n=51$). With the exception of the Ethics group, variations were not statistically significant ($p<.50$).

Table 4. Participant Post -Training Alpha Coefficient s

Group	Alpha
Communicating Vision and Intent	.8254
Application of Leadership Skills	.8778
Teambuilding	.6901
Detecting and Mitigating Error	.6575
Managing Stress and Other Human Factors	.8227
Decision Making	.8416
Ethics	.2194

The variation in the Ethics group is unique to the participant post-training instrument, and could be caused by any number of explanations. One possible explanation is that the Ethics group only contains two items, which can confound the statistical program used to analyze the results of this test, though the same variation did not arise in the participant pre-training instrument or either of the supervisor instruments.

Table 5 displays reliability coefficients (Alpha) scores for the test of the Pre-Training Skills Evaluation. Alpha scores ranged from .8068 to .9191, indicating high

reliability of the instrument, based on this limited sample ($n=51$). Variations were not statistically significant ($p<.50$).

Table 5. Participant Pre-Training Alpha Coefficients

Group	Alpha
Communicating Vision and Intent	.8068
Application of Leadership Skills	.8943
Teambuilding	.8724
Detecting and Mitigating Error	.8070
Managing Stress and Other Human Factors	.8311
Decision Making	.8505
Ethics	.9191

Delayed Post-Course Survey Instrument Administered to Supervisors

As with the participants' survey, when completing the delayed post-course survey, the responding supervisor completes the survey in two parts, on two separate forms. On Form 1 - Post-Training Skills Evaluation, the respondent is asked to evaluate the skills and knowledge that they feel a specified subordinate has now, after taking the *FLL* Training. On Form 2 - Pre -Training Skills Evaluation the respondent is asked, using the same 37 questions, to evaluate the level of skill and knowledge the specified subordinate had before participating in the training.

Table 6 displays reliability coefficients (Alpha) scores for the test of the Post-Training Skills Evaluation. Alpha scores ranged from .8680 to .9334, indicating high reliability of the instrument, based on this limited sample ($n=51$). Variations were not statistically significant ($p<.50$).

Table 6. Supervisor Post -Training Alpha Coefficients

Group	Alpha
Communicating Vision and Intent	.9334
Application of Leadership Skills	.9132
Teambuilding	.8815
Detecting and Mitigating Error	.8995
Managing Stress and Other Human Factors	.8731
Decision Making	.8680
Ethics	.9447

A typographical error in the supervisor instrument required that element 33 (Understands the relationship between experience, memory and decision-making), which falls within the Decision Making group, be eliminated for the purposes of statistical evaluation. This limitation is discussed in the Methodology section (Chapter III) of this paper.

Table 7 displays reliability coefficients (Alpha) scores for the test of the Pre-Training Skills Evaluation. Alpha scores ranged from .8635 to .9384, indicating high reliability of the instrument, based on this limited sample ($n=51$). Variations were not statistically significant ($p<.50$).

Table 7. Supervisor Pre-Training Alpha Coefficients

Group	Alpha
Communicating Vision and Intent	.9384
Application of Leadership Skills	.9252
Teambuilding	.8892
Detecting and Mitigating Error	.8848
Managing Stress and Other Human Factors	.8635
Decision Making	.9349
Ethics	.9337

Testing Validity

The delayed post-course survey instruments were also analyzed by employing T-Tests for paired samples. The Paired Samples T-Test compares the means of two variables, in this case the participant or supervisor's response to the same elements on the pre-training and post-training surveys. The test computes the difference between the two variables for each case, and tests to see if the average difference is significantly different from zero. Unlike the test for reliability, for the purposes of this test, pairs were established by individual learning target.

Table 8 displays Sig. (2-tailed) results for the test of the survey administered to participants. With only one exception (.004), values for all pairs were 0.00, indicating that, for the people who self-reported their ability via this questionnaire, there had been significant improvement between the pre-training period and the post-training period on each of the learning targets ($p < .05$). This test involved a relatively small sample ($n=52/n=54$), and while means and standard deviations could change with a larger sample, what is reported here should be encouraging to those interested in this training.

Table 8. Paired Samples T-Test for Survey Administered to Participants

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	PREQ1 - POSTQ1	-.7115	.5718	7.929E-02	-.8707	-.5524	-8.974	51	.000
Pair 2	PREQ2 - POSTQ2	-.6667	.7004	9.531E-02	-.8578	-.4755	-6.995	53	.000
Pair 3	PREQ3 - POSTQ3	-.4717	.8683	.1193	-.7110	-.2324	-3.955	52	.000
Pair 4	PREQ4 - POSTQ4	-.8148	.8484	.1155	-1.0464	-.5832	-7.058	53	.000
Pair 5	PREQ5 - POSTQ5	-1.1667	.8633	.1175	-1.4023	-.9310	-9.931	53	.000
Pair 6	PREQ6 - POSTQ6	-.9630	1.0633	.1447	-1.2532	-.6727	-6.655	53	.000
Pair 7	PREQ7 - POSTQ7	-.9444	1.0888	.1482	-1.2416	-.6473	-6.374	53	.000
Pair 8	PREQ8 - POSTQ8	-.8611	.7032	9.569E-02	-1.0530	-.6692	-8.999	53	.000
Pair 9	PREQ9 - POSTQ9	-.7547	.7821	.1074	-.9703	-.5391	-7.025	52	.000
Pair 10	PREQ10 - POSTQ10	-1.4906	1.0119	.1390	-1.7695	-1.2117	-10.724	52	.000
Pair 11	PREQ11 - POSTQ11	-1.1111	.8615	.1172	-1.3462	-.8760	-9.478	53	.000
Pair 12	PREQ12 - POSTQ12	-1.0189	.7719	.1060	-1.2316	-.8061	-9.610	52	.000
Pair 13	PREQ13 - POSTQ13	-.8113	.7353	.1010	-1.0140	-.6087	-8.033	52	.000
Pair 14	PREQ14 - POSTQ14	-.8077	.8174	.1134	-1.0353	-.5801	-7.125	51	.000
Pair 15	PREQ15 - POSTQ15	-.8679	.9813	.1348	-1.1384	-.5974	-6.439	52	.000
Pair 16	PREQ16 - POSTQ16	-.7925	.9064	.1245	-1.0423	-.5426	-6.365	52	.000
Pair 17	PREQ17 - POSTQ17	-1.3208	.9151	.1257	-1.5730	-1.0685	-10.507	52	.000
Pair 18	PREQ18 - POSTQ18	-1.1731	.8794	.1220	-1.4179	-.9282	-9.619	51	.000
Pair 19	PREQ19 - POSTQ19	-.9057	.9044	.1242	-1.1549	-.6564	-7.291	52	.000
Pair 20	PREQ20 - POSTQ20	-.9623	.8312	.1142	-1.1914	-.7332	-8.428	52	.000
Pair 21	PREQ21 - POSTQ21	-.7736	.6691	9.191E-02	-.9580	-.5891	-8.416	52	.000
Pair 22	PREQ22 - POSTQ22	-.7736	.7756	.1065	-.9874	-.5598	-7.261	52	.000
Pair 23	PREQ23 - POSTQ23	-.8491	.7941	.1091	-1.0679	-.6302	-7.784	52	.000
Pair 24	PREQ24 - POSTQ24	-1.1481	1.0712	.1458	-1.4405	-.8558	-7.876	53	.000
Pair 25	PREQ25 - POSTQ25	-1.0741	1.0614	.1444	-1.3638	-.7844	-7.436	53	.000
Pair 26	PREQ26 - POSTQ26	-1.0185	.7395	.1006	-1.2204	-.8167	-10.121	53	.000
Pair 27	PREQ27 - POSTQ27	-1.0189	.8877	.1219	-1.2636	-.7742	-8.355	52	.000
Pair 28	PREQ28 - POSTQ28	-1.0755	.8050	.1106	-1.2974	-.8536	-9.726	52	.000
Pair 29	PREQ29 - POSTQ29	-.9808	.7538	.1045	-1.1906	-.7709	-9.382	51	.000
Pair 30	PREQ30 - POSTQ30	-.8868	.9336	.1282	-1.1441	-.6295	-6.915	52	.000
Pair 31	PREQ31 - POSTQ31	-1.3585	1.0018	.1376	-1.6346	-1.0824	-9.872	52	.000
Pair 32	PREQ32 - POSTQ32	-1.0377	.8077	.1109	-1.2604	-.8151	-9.353	52	.000
Pair 33	PREQ33 - POSTQ33	-1.2353	.8852	.1239	-1.4843	-.9863	-9.966	50	.000
Pair 34	PREQ34 - POSTQ34	-1.1321	1.0198	.1401	-1.4132	-.8510	-8.082	52	.000
Pair 35	PREQ35 - POSTQ35	-.8302	.8712	.1197	-1.0703	-.5900	-6.937	52	.000
Pair 36	PREQ36 - POSTQ36	-1.1509	2.7554	.3785	-1.9104	-.3915	-3.041	52	.004

Table 9 displays Sig. (2-tailed) results for the test of the survey administered to supervisors of training participants. Values ranged from .000 to .055 indicating that, for the people who reported their subordinates skills and knowledge via this questionnaire, there had been significant improvement between the pre-training period and the post-training period on all but one learning target ($p < .05$). Based on this small sample ($n = 22$), statistically significant improvement is not indicated for one element (Can effectively delegate tasks to subordinates), which registered a Sig. (2-tailed) value of .055 ($p = .05$). This test involved a small sample ($n = 22$), and while means and standard deviations could change with a larger sample, what is reported here should be encouraging to those interested in this training.

Table 9. Paired Samples T-Test for Survey Administered to Supervisors

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	PREQ1 - POSTQ1	-.2857	.5606	.1223	-.5409	-3.05E-02	-2.335	20	.030
Pair 2	PREQ2 - POSTQ2	-.5238	.6016	.1313	-.7976	-.2500	-3.990	20	.001
Pair 3	PREQ3 - POSTQ3	-.6190	.5896	.1287	-.8874	-.3507	-4.812	20	.000
Pair 4	PREQ4 - POSTQ4	-.4000	.5026	.1124	-.6352	-.1648	-3.559	19	.002
Pair 5	PREQ5 - POSTQ5	-.4286	.5976	.1304	-.7006	-.1565	-3.286	20	.004
Pair 6	PREQ6 - POSTQ6	-.5263	.6967	.1598	-.8621	-.1905	-3.293	18	.004
Pair 7	PREQ7 - POSTQ7	-.5263	.7723	.1772	-.8986	-.1541	-2.970	18	.008
Pair 8	PREQ8 - POSTQ8	-.5789	.8377	.1922	-.9827	-.1752	-3.012	18	.007
Pair 9	PREQ9 - POSTQ9	-.5000	.6070	.1357	-.7841	-.2159	-3.684	19	.002
Pair 10	PREQ10 - POSTQ1	-.6667	.5774	.1260	-.9295	-.4039	-5.292	20	.000
Pair 11	PREQ11 - POSTQ1	-.5714	.7464	.1629	-.9112	-.2317	-3.508	20	.002
Pair 12	PREQ12 - POSTQ1	-.6667	.7303	.1594	-.9991	-.3342	-4.183	20	.000
Pair 13	PREQ13 - POSTQ1	-.4500	.6048	.1352	-.7331	-.1669	-3.327	19	.004
Pair 14	PREQ14 - POSTQ1	-.5263	.6118	.1404	-.8212	-.2314	-3.750	18	.001
Pair 15	PREQ15 - POSTQ1	-.5500	.7592	.1698	-.9053	-.1947	-3.240	19	.004
Pair 16	PREQ16 - POSTQ1	-.3000	.6569	.1469	-.6075	7.461E-03	-2.042	19	.055
Pair 17	PREQ17 - POSTQ1	-.6667	.7303	.1594	-.9991	-.3342	-4.183	20	.000
Pair 18	PREQ18 - POSTQ1	-.8500	.7452	.1666	-1.1987	-.5013	-5.101	19	.000
Pair 19	PREQ19 - POSTQ1	-.5714	.7464	.1629	-.9112	-.2317	-3.508	20	.002
Pair 20	PREQ20 - POSTQ2	-.6316	.7609	.1746	-.9983	-.2648	-3.618	18	.002
Pair 21	PREQ21 - POSTQ2	-.6000	.5982	.1338	-.8800	-.3200	-4.485	19	.000
Pair 22	PREQ22 - POSTQ2	-.5714	.7464	.1629	-.9112	-.2317	-3.508	20	.002
Pair 23	PREQ23 - POSTQ2	-.4706	.7174	.1740	-.8395	-.1017	-2.704	16	.016
Pair 24	PREQ24 - POSTQ2	-.8889	1.1318	.2668	-1.4517	-.3260	-3.332	17	.004
Pair 25	PREQ25 - POSTQ2	-.9000	.9679	.2164	-1.3530	-.4470	-4.158	19	.001
Pair 26	PREQ26 - POSTQ2	-.4737	.6118	.1404	-.7686	-.1788	-3.375	18	.003
Pair 27	PREQ27 - POSTQ2	-.7368	.8057	.1848	-1.1252	-.3485	-3.986	18	.001
Pair 28	PREQ28 - POSTQ2	-.5500	.7592	.1698	-.9053	-.1947	-3.240	19	.004
Pair 29	PREQ29 - POSTQ2	-.5263	.7723	.1772	-.8986	-.1541	-2.970	18	.008
Pair 30	PREQ30 - POSTQ3	-.5625	.8139	.2035	-.9962	-.1288	-2.764	15	.014
Pair 31	PREQ31 - POSTQ3	-.6842	.6710	.1539	-1.0076	-.3608	-4.444	18	.000
Pair 32	PREQ32 - POSTQ3	-.7778	.8782	.2070	-1.2145	-.3411	-3.757	17	.002
Pair 33	PREQ34 - POSTQ3	-.9500	.9987	.2233	-1.4174	-.4826	-4.254	19	.000
Pair 34	PREQ35 - POSTQ3	-.5714	.8106	.1769	-.9404	-.2024	-3.230	20	.004
Pair 35	PREQ36 - POSTQ3	-.4762	.6016	.1313	-.7500	-.2024	-3.627	20	.002
Pair 36	PREQ37 - POSTQ3	-.6000	.8826	.1974	-1.0131	-.1869	-3.040	19	.007

Chapter V

Conclusions and Recommendations

The purpose of this project was to develop an evaluation strategy for the *FLL* training that would provide a template for an evaluation strategy of the entire NWCG Leadership training curriculum. The following material documents the major conclusions reached regarding the development of such an evaluation strategy, and links them to the various research questions associated with this project. These conclusions result from an extensive review of literature pertaining both to training evaluation and the *FLL* training and from a small-scale, operational test of training evaluation instruments.

Research Questions and Conclusions

Research Question (1): *What are the program goals or objectives of the FLL training, and can they be used as a basis for evaluating the training?*

The literature makes clear that to answer how well a leadership development training program is working, one must establish clear program goals, and that clear program goals enable training organizations to evaluate a leadership development program. Most often, these program goals are expressed in the form of training objectives. Program goals or training objectives are alternatively known as learning targets and student outcomes. The evaluation strategy developed during this project and described in this paper uses 32 “learning targets” in lieu of objectives. These learning targets were identified during the literature search conducted as part of this project and validated by the Leadership Committee in consultation.

Research Question (2): *What evaluation strategy and evaluation design will serve the FLL training curriculum effectively?*

Based on the author's assessment of the organizational factors identified by Tannenbaum and Woods (1992) and reported in Chapter IV (Presentation of Findings), an intermediate evaluation strategy of the *FLL* training program seems appropriate. The evaluation strategy resulting from this project employs evaluation measures of quasi-experimental design oriented to produce Level 1 (Reaction), Level 2 (Learning) and Level 3 (Behavior) evaluation as described by Kirkpatrick (1956 & 1996), as well as Tannenbaum and Woods. (1992)

Intermediate evaluations offer greater sophistication and fewer threats to validity than basic, reaction-only strategies; even though they do not incorporate some of the controls seen in an advanced, experimentally based strategy. Consequently, there are many advantages to an intermediate evaluation strategy. This approach yields more useful and valid results than a reaction-only strategy and, hence, provides better information for planning purposes. Although an intermediate evaluation strategy is less rigorous than an advanced, experimental strategy, it typically yields information that is interpretable and defensible. This strategy is also less obtrusive than most advanced evaluations and usually fits more readily with organizational realities.

By using the evaluation strategy described in this paper, the Leadership Committee will evaluate the *FLL* training at Kirkpatrick's Level 1 (Reaction) using an instrument administered immediately post-course, as is done now.

The current form was reviewed and redesigned to assure that it:

- Allows effective quantification of reactions
- Encourages written comments and suggestions
- Attains an immediate response rate of 100 percent
- Helps achieve the Level 2 (Learning) evaluation described below

By employing the described strategy, the Committee will also evaluate the *FLL* training at Kirkpatrick's Level 2 (Learning), but do so without testing, which is not practical for *FLL*. Since testing does not represent a pragmatic approach for *FLL*, the collection of Level 2 evaluation will be accomplished by using the same self-reporting, "reaction-style" measure employed at the conclusion of the training. To effectively measure learning, the instrument has been specifically designed to inquire about the transferability/utility of the training. Consequently, to effectively evaluate the *FLL* training at Kirkpatrick's Levels 1 and 2, (Reaction and Learning) the instrument uses specific questions that mirror the learning targets, are quantifiable or "scoreable", and utilize "utility" questions that estimate the potential for the trainee to transfer what they learned in training to the workplace.

The nature of the target audience (NWCG agency workforces) presents two significant barriers to comprehensive, systematic training evaluation, both of which must be overcome. First, the client organization lacks a centralized database of the approximately 6,000 people who have completed the *FLL* training, a fact that arose as a challenge to this project, and will continue to challenge efforts to systematically evaluate this training curriculum until corrected.

Second, the mobile and temporal nature of the agency work force also presents a barrier to systematic evaluation. Employees, particularly entry-level employees and emerging leaders, (the target audience of this training), relocate relatively often in response to career opportunities in agencies that encourage, and nearly require, employees to move-around in order to move-up. In addition, NWCG agencies employ part-time, seasonal, and temporary employees; many of who are within the target audience of the *FLL* training. Consequently, at any given time, access to some employees within target population samples can prove problematic, a factor at odds with the longitudinal nature of training evaluation, particularly when employing delayed or repeated measures.

The small-scale test conducted as part of this project demonstrated that it may prove impractical to conduct delayed, post-course evaluation on a project basis, as would be done for a research study. In the test, response rates were inadequate to generate the desired statistical sample and, despite giving participants more than one month to complete their simple survey, completed surveys continued to drift in for over one month beyond the deadline. For this reason, in addition to improving data collection and centralizing a database, it appears that using delayed, post-course instruments to evaluate the *FLL* training may best be accomplished through a continuous, seven-step process as follows:

Step One: Collect contact data for participant and supervisor from every participant in every class (approximately 500-700 participants per year.) The author developed a form to collect contact data from course participants. The form is intended

to collect contact data for both the participant and their supervisor, and is included in the user's manual in Appendix A of this report.

Step Two: Enter the contact data for each participant and each supervisor into a database as they come in (or as close as possible - but continuously)

Step Three: At the time of entry into the database, the subordinate employee (participant) who the supervisor's evaluation will be based on is noted (along with the supervisor) in the database (necessary randomization is completed at this step when multiple participants in the database have the same supervisor)

Step Four: Employ a calendar function that alerts the evaluation administrator. Six months after course completion, the calendar function should produce a list of names to contact with evaluation surveys.

Step Five: Surveys will come in continuously, and they will be filed as they come in (anticipate a 30% - 50% return). Using this system, anticipate that surveys from 100-200 participants and 100-200 supervisors will be received each year.) Separate files will be kept for participant surveys and supervisor surveys.

Step Six: Every time either the participant or supervisor file reaches a sample size of 100, the evaluation administrator will run, or arrange for, analysis. It is anticipated that data entry and general administration will be done "in-house", but that the Leadership Committee will have to outsource the statistical analysis. The Leadership Committee or administrator could obtain sustained, cost-effective statistical service by establishing a relationship with a college or university.

Step Seven: Supplement the quantitative surveys with qualitative focus groups (every two years).

Research Question (3): *Can easily obtainable and inexpensive self-report data, supported by other qualitative data collection, provide a sustainable source of feedback on the FLL training approximating Kirkpatrick's Level 1 (Reaction), Level 2 (Learning), and Level 3 (Behavior)?*

The Leadership Committee should endeavor to keep the evaluation system as simple as possible. This is particularly important because, as the level of evaluation goes up the complexities involved increase, and complexity can erect a barrier to evaluation. Consequently, evaluators need to strike a balance between effective measures and an evaluation strategy that the organization can be reasonably expected to accomplish. A well-planned evaluation minimizes the burden by building in practical and unobtrusive ways of gathering data. Based on the review of the pertinent literature, the author does not believe that the Leadership Committee needs to constrain itself to strict compliance with Kirkpatrick's evaluation hierarchy. Given organizational resources and demands on those resources, easily obtainable and inexpensive self-report data, supported by other qualitative data collection, represents the most sustainable source of ongoing feedback on the *FLL* training beyond Kirkpatrick's Level 2 (Learning.) Like all training evaluators, the Leadership Committee is faced with a choice between the strongest design and one that is practical in terms of time, resources, opportunity, and acceptability. The evaluation strategy presented here strives to strike a balance between an effective design and practicality.

Research Question (4): *Is it practical to measure the efficiency of FLL training at Kirkpatrick's Level 4 (Results) or for return-on-investment (ROI)?*

While it may represent the ideal to evaluate training in terms of results, numerous concerns suggest that measuring the *FLL* training in terms of results represents an unrealistic ideal. In response to concerns over the practicality of measuring at Level 4, the Leadership Committee will not measure the efficiency of *FLL* training at Kirkpatrick's Level 4 (Results.)

First, Level 4 evaluation may not be applicable to leadership development training, and there is some evidence that a results oriented evaluation is not applicable to "soft skills" training. Second, organizational constraints can greatly limit the opportunities for gathering results-oriented data. The data needed to support Level 4 evaluation can require substantial time, effort, and expense; and the training organization needs to weigh the potential costs against the potential value of the results. It seems unlikely that the agencies sponsoring *FLL* training could justify a Level 4 (Results) approach.

In addition, it can be difficult to get meaningful measures at this level. The literature shows that not only is it difficult to obtain sound measures at this level (especially in comparison with a previous period), but identification of a single training activity as the cause of any changes observed is dubious, as there are too many variables that can impact performance, other than the training itself. In addition, the literature clearly indicates that Level 4 evaluation should be conducted only when it is likely that training will have a detectable effect on business results, organizational output, or business results. These concerns are particularly important to this effort, as

the agencies sponsoring *FLL* training are not likely to be able to measure changes or observable outcomes. The *FLL* training teaches principles and concepts that could produce a mix of observable outcomes and unobservable attitude shifts.

Finally, in regard to Level 4 evaluation, the literature also indicates a need to consider organizational constraints when considering Level 4 evaluation. Level 4 evaluations may likely require longitudinal research, and the main difficulty in doing longitudinal research is obtaining access to participants. This concern is of importance to this project, as *FLL* participants come from a highly mobile, and frequently temporary, workforce.

Evaluating return-on-investment (ROI) represents a perennial topic in the training evaluation field, and one involving a great deal of controversy. First, some researchers and practitioners question the very concept of ROI for training. In addition, most organizations are not in a position to carry out such an assessment, due to uncertainties over the benefits of training and because of the difficulty in accounting for its true cost. Finally, according to some training evaluation scholars, research techniques for evaluating training in terms of dollars and cents are not currently well developed or adequate.

Research Question (5): Can one use a design that compares pre-training and post-training results without using a control group to evaluate the FLL training?

According to the cited literature, one can use a design that compares pre-training and post-training results without using a control group. By eliminating the control group requirement, evaluators can collect data more easily because the sample size is smaller, and non-control group design is likely to be less expensive than a control group

design. However, this approach does not eliminate the possibility that pre- and post-training changes are due to variables other than training. The literature suggests that if the training organization cannot use a control group, they should instead establish some baselines with which to compare post-training results. Since the method provided here does not to use an experimental control group to collect baseline data, as an alternative, the method collects baseline data by use of the retrospective pretest, explained elsewhere in these conclusions.

Research Question (6): *What evaluation methods, techniques, measures or instruments will best evaluate the FLL training at, at least, Kirkpatrick's Levels 1, 2 and 3?*

There is sufficient support in the literature to suggest that training evaluators will improve the effectiveness of their evaluation by employing multiple methods, techniques, and instruments. Ultimately, the measures provided here will:

- Measure the learning of each trainee so quantitative results can be determined
- Use a before-and-after approach so that learning can be related to the program
- Attain a response rate of 100 percent on immediate measures
- Utilize a statistically valid sample for delayed measures
- Measure learning on an objective basis as much as possible
- Analyze the evaluation results statistically, where possible, so that learning can be proven in terms of correlation or level of confidence

The literature suggests a strategy that uses a combination of immediate measures and delayed measures. Some scholars argue that measures conducted immediately after training have slightly higher reliability than more delayed measures of

the same criterion. Others suggest that by gathering data several months after training, trainees will have experienced whether the training was useful and should be in a better position to judge the utility of the training. Since the literature is not conclusive on this point, the strategy presented here employs a mix of immediate and delayed measures.

The literature also indicates that it is highly desirable to record both prior and subsequent performance of key behaviors before and after training. The described strategy will evaluate the *FLL* Training At Kirkpatrick Level 3 (Behavior) via a retrospective pretest administered to both trainees and supervisors. A review of the extant literature shows that Level 3 (Behavior) can be effectively measured using these self-reporting instruments.

Finally, the research describes evaluation approaches using qualitative interviewing and other “stakeholder-based” approaches as their principal means for data collection and analysis. This type of interviewing for data collection is well established for field research in the social sciences, and is emerging as an application for learning-related research in technology-based organizations. The strategy described here includes the provision of periodic focus groups for this purpose.

By using a retrospective pretest the strategy provided will:

- Conduct a systematic appraisal of on-the-job performance on a before-and-after basis
- Conduct a post-training appraisal at least three and up to six months after training
- Add to the validity of the evaluation
- Evaluate the training from the perspectives of both the trainees and their supervisors
- Sample at least 100 trainees and at least 100 supervisors per year
- Conduct a statistical analysis to compare before-and-after performance and relate changes to the training.

Evaluation instrument design is critically important, and the format of the questionnaire is crucial in eliciting accurate responses. The strategy described here endeavors to:

- Use an adequate sample for all measures.
- Use statistical significance (not cost-benefit) to determine the program's success
- Use the smallest number of items necessary to adequately tap the respondents' perceptions of each major factor
- When possible/appropriate, ask respondents why they answered as they did
- Keep the evaluation short. Long evaluations will reduce the response rate
- Avoid mixing rating and ranking questions
- Use a rating scale instead of words, anchoring various points on the rating scale with descriptive phrases
- Use clear statements that respondents are likely to perceive similarly, focusing on concrete trainer behaviors and specific facets of the training environment.

Other Conclusions

Immediate Post-Course Attitude Survey Instrument

When the immediate post-course attitude survey was tested, question 12 (“What topics would you have liked to spend more or less time on?”) and question 13 (“What did the instructors do that worked well and what might you suggest to improve their effectiveness?”) were “double-barreled” questions that violated a design principle of this project and negatively impacted the utility of the participants’ response. Consequently, each of these questions has been divided into two, separate questions (making four questions of the original two), which will alleviate this design flaw. As a consequence, the form was necessarily lengthened, pushing it onto a second page. However, the author believes that this consequence is acceptable given the improvement the changes make in the utility of the instrument. This design change could increase the possibility that participants will inadvertently return incomplete evaluation instruments, so that possibility is addressed both on the form and in the operational guidance. An additional, positive, consequence is that space for participants’ comments is improved.

Less than 1% of participants ($n=1$) misinterpreted the scale of the “scoreable” portion of the instrument. Consequently, no modification to the scale was made.

A typographical error in the supervisor instrument required that one element (of 37) be eliminated for the purposes of statistical evaluation during the small-scale test. This error has since been corrected.

Delayed Post-Course Survey Administered to Participants

The variation in the Ethics group is unique to the participant post-training instrument, and could be caused by any number of explanations. One possible explanation is that the Ethics group only contains two items, which can confound the statistical program used to analyze the results of this test. However, the same variation did not arise in the participant pre-training instrument or either of the supervisor instruments. The instrument can and should be used as-is, but statistical analysis of this element should be conducted again once a larger sample is available.

If the variance were to continue, three courses of action may be pursued:

1. Include additional items in the Ethics group to accommodate the statistical program
2. Re-evaluate the questions in this group to assure that they can be answered easily
3. Evaluate this portion of the training and its success at achieving this learning target

Delayed Post-Course Survey Administered to Participants

The typographical error in element 33 of the supervisor instrument (Understands the relationship between experience, memory and decision-making) has been corrected, and no further action is required other than that, for the purposes of evaluation, statistical analysis of this element should be conducted again once a larger sample is available.

Summary Of Conclusions

At the bottom-line, it was the intent of this project to develop a method for collecting and analyzing training related data to support the leadership initiative, validate that the *FLL* training is on track, and provide a model or template for ongoing evaluation of, not only the *FLL* program, but the broader leadership curriculum. The project's objective was to develop a quantifiable and, ideally, a statistically supportable method. Literature was reviewed, learning targets were established, an evaluation strategy was designed, evaluation methods were selected and adapted for use, data collection instruments were designed and developed, quantifiable and statistical measures were established, and a small-scale test of the evaluation method was conducted.

While the sample size was limited and some revisions to the strategy and the evaluation instruments were necessary, the results of the small-scale test indicate that the evaluation strategy described here achieves the desired outcomes and may be implemented with confidence. Procedural guidance was developed and is included as Appendix A of this paper. The resulting evaluation strategy works as follows:

- The Leadership Committee of the NWCG Training Working Team will designate an evaluation administrator (or administrators) and implement a comprehensive and systematic approach to evaluation of the *FLL* training and other courses in the leadership curriculum as the Committee sees fit.
- The Committee/administrator(s) will develop a centralized database of *FLL* training participants and their supervisors to facilitate future evaluation. The centralized database will be used only for this purpose, and all contact information and participant responses will be kept confidential.

- Participants will provide contact information for their supervisors and themselves pre-course. This will enable post-course contact of both participants and their supervisors.
- Participants will immediately complete a post-course evaluation instrument at the conclusion of the course. This evaluation tool is designed to collect both Level 1 (Reaction) and Level 2 (Learning) data. 100% compliance will be sought. The immediate post-course instrument allows for quantification and statistical analysis of 11 questions if desired.
- All course participants and at least 100 supervisors will be contacted each year, and requested to complete a post-course survey employing the “post-then”, retrospective pretest methodology per procedures provided as part of the operational guidance. Surveying will be conducted on a continuous (vs. project) basis, and a return rate of approximately 50% is anticipated. These delayed post-course instruments are designed for statistical analysis.
- Both immediate and delayed measures will be quantified and/or statistically validated as appropriate per procedures provided as part of the operational guidance.
- In addition to surveying course participants and their supervisors to obtain quantitative data about the leadership training, the Leadership Committee will periodically conduct focus groups to obtain qualitative data. Through focus groups, the Committee will ask multiple groups of stakeholders to share their views on the leadership training curriculum to supplement and corroborate the results of course evaluation surveys. It is recommended that the Committee place its highest priority on conducting focus groups comprised of current subordinates of past *FLL* course participants, because

the perspectives of these stakeholders are not otherwise captured in the evaluation system. Protocols are provided as part of the operational guidance.

Limitations of the Study

Because this project was of a quasi-experimental design, the author was unable to make the strongest case possible for the resulting evaluation strategy and method.

Sample size and sample bias represent other significant limitations of this study. Though the project's test to validate evaluation instruments was only intended to be of a small-scale, surveying past *FLL* course participants and the supervisors of past training participants proved problematic even for this small-scale test.

A database had to be established specifically for the purposes of the project using a combination random and convenience sampling. The project's timing allowed only a limited opportunity to develop a sample population complete with contact data. This limitation proved most problematic in relation to efforts to develop a sample population of supervisors, and consequently, the supervisor sample included only 35 persons.

Due to constraints presented by the lack of a centralized database, survey samples were developed by including all participants, of entire classes, with subject courses selected as randomly as possible, but not truly at random. In addition, due to the make-up of one of the courses from which the population samples were developed, three people appeared in both the participant and supervisor samples. Exacerbating the problem of sample size, response rate to the surveys was disappointingly low. In all, 55 of 120 course participants and 22 of 35 supervisors of course participants responded to the surveys. Consequently random selection and sample size have

suffered, and the resulting sample size and selection bias limit the ability to generalize the project's findings.

Recommendations

The purpose of these recommendations is to identify areas where changes and actions are needed within the client organization (Leadership Committee of the NWCG Training Working Team). By acting on these recommendations, the Committee will cause organizational action facilitating the implementation of the evaluation strategy and methods described in this paper. Therefore, the following recommendations are intended to close the gap between the status quo and the desired future condition in regard to evaluation of the *FLL* training. The recommendations are as follows:

- Implement the training evaluation method as described in this paper and in the attached operational guidance until such a time that a sufficient experience and data is obtained to further test the validity of the methodology and the associated instruments.
- Begin using the contact form included as part of this evaluation method immediately to assist in building a centralized database of *FLL* participants and their supervisors.
- Establish a centralized database of *FLL* course participants that employs “calendar” or reminder functions that alert those with responsibility to administer the evaluation strategy to the need to survey participants and supervisors.
- Maintain strict confidentiality of the database and responses. The centralized database should be used only for the purpose of evaluating the *FLL* training, and all contact information and participant responses should be kept confidential.
- Designate an “Evaluation Administrator” (or administrators) to implement, manage and maintain the evaluation strategy.

- The small-scale test conducted as part of this project demonstrated that it will likely prove impractical to conduct delayed, post-course evaluation on a project basis, as would be done for a research study. For this reason; in addition to improving contact data collection, centralizing a database, and appointing Evaluation Administrators; to employ the delayed, post-course instruments, implement the continuous seven-step process described in the conclusions made earlier in this chapter,.
- Begin planning for qualitative data collection, through focus groups as described in the operational guidance, as soon as possible.

Implications for Future Research

Several implications can be drawn from this project that may assist those who evaluate leadership development training. Some implications address weaknesses of this project, while others make recommendations for people interested in evaluating the efficacy of training or suggest future research needs. The implications are as follows:

Instrument validity and reliability, as well as the ability to generalize results matter. First and foremost, the evaluating organization must know that the evaluation instruments and methods they are using measure what it is the organization wants to know. Evaluation validity begins with objectives. The presence of clear training objectives is critical to good training evaluation, and knowing whether evaluation instruments are measuring what one wants to know is nearly impossible if the organization has not first laid the foundation for assessment by establishing clear learning objectives. Second, when planning an evaluation strategy, the sampling design and access to the desired population sample have enormous impact on the ability to generalize the findings of the evaluation study. Future research into the

evaluation of leadership development training should focus on developing evaluation instruments that measure participants' progress against specific objectives and then testing the validity and reliability of those evaluation instruments using adequate population samples that minimize selection bias and represent the training population.

To contribute most to the continuous improvement of training, one should evaluate that training at multiple levels, and by using multiple methods, techniques and instruments to cross-check and verify or disconfirm reaction, learning and behavioral change data; particularly when the evaluation method depends on self-reporting measures. Consequently, future research into evaluating leadership training might focus on developing evaluation strategies that enable evaluation of training by relatively simple self-reporting measures, but survey participants as well as their superiors, subordinates and peers and combine quantitative and qualitative research techniques.

Training organizations may effectively evaluate training at Kirkpatrick's Level 2 (Learning) and Level 3 (Behavior) using what are typically thought of as Level 1 (Reaction) measures. However, the research literature supporting reaction level measurement of Learning (Level 2) and Behavior (Level 3) envisions evaluation instruments specifically designed to do so, and does not support general predictability of learning extrapolated from typical reaction measures. Notably, reaction measures that directly ask trainees about the transferability or utility of the training can correlate with on-the-job performance, ascertain the perceived usefulness of training for subsequent job performance, and serve as a predictor of behavioral change. Future research should continue to explore this arena, with the intent of producing practical and effective

methods of achieving training evaluation at Kirkpatrick's Level 3 using cost effective, non-experimental strategies.

The research literature describes potential problems with internal validity when using traditional pre/post evaluation design, and that, the training organization can counter response-shift bias and get more accurate and realistic results describing the effectiveness of their training by employing retrospective pretests as an alternative to the conventional pretest. The strategy described in this paper is intended to use such an approach, evaluating the *FLL* training at Kirkpatrick Level 3 (Behavior) via a self-reporting, retrospective pretest administered to both trainees and supervisors . During the conduct of this study, questions arose about whether survey respondents would be confused by the unfamiliar method. While those concerns were successfully mitigated, one question remained. Would the method prove more or less effective if the retrospective pretest ("then" measure) were worded differently (in past tense) than the posttest ("post" measure)? This line of inquiry presents an interesting research question for future study, and could serve as the basis of a master's or doctoral student's research into evaluation of leadership training.

Final Discussion

In conclusion, this project provides a strategy for evaluating the *FLL* training, and takes first steps toward establishing a mechanism for evaluating the effectiveness of the entire NWCG leadership training curriculum. The objective should be to accurately assess how the leadership training impacts job performance in NWCG member agencies. While the system for evaluating this training may evolve over time, the outcomes of this project have produced what should prove to be an effective evaluation strategy and method. Given the strategic importance of this training, the scope of participation, the relative costs to the participating agencies, and the vulnerability of agency training funds; the sponsor organizations appear to have strong incentive to evaluate the *FLL* training program to maximize the benefits they receive from their substantial investment in this training.

References

- Abernathy, D. (1999). Thinking Outside the Evaluation Box. *Training and Development* (53) 2: 18-23.
- Alliger, G.M. & Janak, E.A. (1989). Kirkpatrick's Levels of Training Criteria: Thirty Years Later. *Personnel Psychology* (42).
- Alliger, G.M. & Tannenbaum, S.I. (1997). A Meta-Analysis of the Relations Among Training Criteria. *Personnel Psychology*, 50 (2.)
- Ashton, D. & Green, F. (1996). *Education, Training and the Global Economy*. Cheltenham: Edward Elgar.
- ASTD. (2004). Retrieved on June 9, 2004 from <http://www.astd.org/>
- Bass, B. & Stogdill, R. M. (1990). *Bass and Stogdill's handbook of leadership: Theory, research, and managerial applications*. (3rd. Ed.) New York: The Free Press.
- Bee, F. (2000). How to evaluate training. *People Management*, 6 (6), 42-43.
- Brown, S.M. (2004) Changing times and changing methods of evaluating training. Retrieved on March 10, 2004 from http://www.ktic.com/TOPIC7/14_BROWN.HTM
- Brungardt, C.L. (1997). *Evaluation of the Outcomes of an Academic Collegiate Leadership Program*. Unpublished doctoral dissertation, Kansas State University. Manhattan, Kansas.
- Campbell, D.T, & Stanley, J.C. (1963). Experimental and quasi-experimental design for research and teaching. In: Gage, N.L. ed., *Handbook of research on teaching*. Chicago, Ill., Rand McNally.
- Cantrell, P. (2003). Traditional vs. Retrospective Pretests for Measuring Science Teaching Efficacy Beliefs in Preservice Teachers. *School Science & Mathematics*, 103 (4.)
- Conway, M. (Ed.). (2004). *Collecting Data With Electronic Tools*. Alexandria, VA. American Society for Training and Development.
- Cousins, J.B. (1995). Participatory Evaluation: Enhancing Evaluation Use and organizational Learning Capacity. *The Evaluation Exchange* I (3/4), Fall 1995. Retrieved on June 9, 2004 from <http://gseweb.harvard.edu/~hfrp/eval/issue2/theory.html>
- Delahoussaye, M. (2001). Leadership in the 21st century. *Training* 38 (9.)

Eckert, A. (2000). Situational Enhancement of Design Validity: The Case of Training Evaluation at the World Bank Institute. *American Journal of Evaluation* 21 (2): 185-193.

Eseryel, D. (2002). Approaches to Evaluation of Training: Theory and Practice. *Educational Technology & Society* 5 (2.)

Fisk, C.N. (Ed.). (1991). *ASTD Trainer's Toolkit: Evaluation Instruments*. Alexandria, VA. American Society for Training and Development.

Fitzgerald, L. & Hammon, M.C. (1999). Assessment and evaluation: TVA University's experience. *Corporate University Review*, 7 (3), 38-40.

Hillman, V. & McDonald, L.M. (2003). *Training Evaluation and Needs Assessment: An assessment of the effectiveness of Fireline Leadership and contracted delivery*. Utah BLM/Forest Service Region 4 – Interagency Fire Training Program. Ogden, Utah.

Hubbard, A. (2001). Training Evaluation. *Mortgage Banking*, 61 (7), 115-118.

Ingram, M. Staten, L., Cohen, S.J., Stewart, R., deZapien, J.G. (2004) The Use of the Retrospective Pre-Test method to Measure Skill Acquisition Among Community Health Workers. *Internet Journal of Public Health Education*, B6-1-15. Retrieved on June 9, 2004 from <http://www.aspher.org/Articles>

Kirkpatrick, D.L. (1956). How to Start an Objective Evaluation of Your Training Program. *The Journal of the American Society of Training Directors*. Retrieved on June 9, 2004 from <http://www.astd.org/>.

Kirkpatrick, D.L. (1996). Great ideas revisited: Techniques for evaluating training programs. *Training & Development*, 50 (1), 54-59.

Kristiansen, N.S. (2004). *Making Smile Sheets Count*. February 2004. Issue 0402. American Society for Training and Development. Alexandria Virginia.

Long, L.N. (1999). ROI: Capturing the Big Picture. *Technical Training*. November/December 1999.

Mann, S. (1997). Implications of the Response-Shift Bias for Management. *Journal of Management Development* 16 (5/6.)

Manthei, R.J. (1997). The Response-Shift Bias in a Counselor Education Program. *British Journal of Guidance & Counseling* 25 (2.)

McDonald, L.S. (2001). Project Report – Fireline Leadership. In *Proceedings 2001 IAWF International Safety Summit*. International Association of Wildland Fire.

- McDonald, L.S. & Shadow, L. (2003.) Precursor for Error: An Analysis of Wildland Fire Crew Leaders' Attitudes About Organizational Culture and Safety. In *Conference Proceedings 3rd International Wildland Fire Conference and Exhibition*. 3rd International Wildland Fire Conference and Exhibition 2003.
- Mezoff, B. (1981). How to Get Accurate Self-Reports Of Training Outcomes. *Training and Development Journal*, September 1981.
- Michalski, G.V. & Cousins, J.B. (2001). Multiple Perspectives on Training Evaluation in a Global network Development Firm. *American Journal of Evaluation* 22 (1): 37-53.
- Newstrom, J.M. (1987). Confronting Anomalies in Evaluation: Reliability and other statistical mysteries are plain at last. *Training and Development Journal*. July 1987.
- Nickols, F. (2003). A Stakeholder Approach to Evaluating Training. Retrieved on June 9, 2004 from <http://www.nickols.us>.
- Nickols, F. (2000). Evaluating Training: There is no "cookbook" approach. Retrieved on June 9, 2004 from <http://www.nickols.us>.
- Nowack, K.M. (1991.) In C.N. Fisk, (Ed.), *ASTD Trainer's Toolkit: Evaluation Instruments* (pp. 105-110). Alexandria, Virginia: American Society for Training and Development.
- Pernick, R. (2001). Creating a leadership development program: nine essential tasks. *Public Personnel Management*, 30 (4), 429-44.
- Pratt, C.C., Mcguigan, W.M. & Katzev, A.R. (2000). Measuring Program Outcomes: Using Retrospective Pretest Methodology. *American Journal of Evaluation*, 21(3).
- Ripley, D.E. (2002). *The Work Environment and Training Effectiveness: An Overlooked Element in Human Resource Management Instruction*. Retrieved on June 9, 2004 from <http://www.westga.edu/~bquest/2002/hresource.htm>
- Rohs, F.R. (2002) Improving the Evaluation of Leadership Programs: Control Response Shift. *Journal of Leadership Education*, 1 (2.)
- Rohs, F.R., Langone, C.A. & Coleman, R.K. (2001). Response Shift Bias: A Problem in Evaluating Nutrition Training Using Self-Report Measures. *Journal of Nutrition Education*, 33 (3.)
- Ruona, E.A., Leimbach, M., Holton, E.F. & Bates, R. (2002). The relationship between learner utility reactions and predicted learning transfer among trainees. *International Journal of Training and Development*. 6 (4.)

- Saari, Johnson, T.R., McLaughlin, S.D. & Zimmerle, D.M. (1988). A Survey of Management Training and Education Practices in U.S. Companies. *Personnel Psychology* (41).
- Santos, S.A. & Stuart, M. (2003). Employee perceptions and their influence on training effectiveness. *Human Resource Management Journal*, 13 (1), 27-45.
- Shelton, S. & Alliger, G. (1993). Who's Afraid of Level 4 Evaluation?: A Practical Approach. *Training & Development* – June.
- Shulha, L.M. & Cousins, J.B. (1997). Evaluation Use: Theory, Research and Practice Since 1986. *Evaluation Practice* 18 (3.)
- Stadius, R. (Ed.). (1999). *ASTD Trainer's Toolkit: More Evaluation Instruments*. Alexandria, VA. American Society for Training and Development.
- Sugrue, B. (2003). 2003. *State of the Industry*. American Society for Training and Development. Alexandria Virginia. Retrieved on June 11, 2004 from <http://www.astd.org>
- Swierczek, F.W. & Carmichael, L. (1985). The Quantity and Quality of Evaluating Training. *Training and Development Journal* - January.
- Tannenbaum, S.I., Mathieu, J.E. & Cannon-Bowers, J.A. (1991). An examination of the Factors that Influence Training Effectiveness: A Model and Research Agenda. Presented at the Sixth Annual Society of Industrial/Organizational Psychology Meetings, St. Louis, MO, April 1991. Retrieved on October 12, 2003 from
- Tannenbaum, S. I. & S.B. Woods. (1992). Determining a Strategy for Evaluating Training: Operating Within Organizational Constraints. *Human Resource Planning*, 15 (2.)
- Training Working Team¹. (2003). The Program. Retrieved on October 12, 2003 from <http://www.fireleadership.gov/program.html>
- Training Working Team². (2002). *Report of the Leadership Task Group to the Training Working Team*. Retrieved on October 12, 2003 from http://fireleadership.gov/committee/reports/February_2001_Task_Group_Report.pdf
- Training Working Team³. (2002). Leadership Committee Charter. Retrieved on October 12, 2003 from <http://fireleadership.gov/committee/reports/Signedcharter.pdf>
- Training Working Team⁴. (2003). Fireline Leadership Course Description. Retrieved on October 12, 2003 from http://fireleadership.gov/committee/courses/L_380.html
- Tyler, K. (2002). Evaluating Evaluations. *HR Magazine* – June 2002.

Umble, K., Upshaw, V, Orton, S. & Mathews, K. (2000). Using the Post-then Method to Assess Learner Change. Presentation at the AAHE Assessment Conference June 15, 2000. Charlotte, North Carolina.

Warr, P. & Bunce, D. (1995). Trainee Characteristics and the Outcomes of Open Learning. *Personnel Psychology*, 48 (2.)

Warr, P. & Catriona, A. (1999). Predicting three Levels of Training Outcome. *Journal of Occupational & Organizational Psychology*, 72 (3), 351-376.

Appendix A. Fireline Leadership Evaluation User's Manual