12-31-2010

Exploring Evaluation in School Districts: School District Evaluators and Their Practice

Susan Hibbard
University of South Florida

Follow this and additional works at: http://scholarcommons.usf.edu/etd
Part of the American Studies Commons, and the Statistics and Probability Commons

Scholar Commons Citation
http://scholarcommons.usf.edu/etd/3514

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Exploring Evaluation in School Districts: School District Evaluators and Their Practice

by

Susan T. Hibbard

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
Department of Measurement & Research
College of Education
University of South Florida

Major Professor: John Ferron, Ph.D.
Jane Applegate, Ph.D.
Robert Dedrick, Ph.D.
Liliana Rodriguez-Campos, Ph.D.

Date of Approval:
November 5, 2010

Keywords:
internal evaluators, education, research on evaluation, manova, survey research

© Copyright 2010, Susan T. Hibbard
Acknowledgments

I would like to thank the wonderful faculty in the Department of Measurement and Research at USF. Bob Dedrick, John Ferron, Connie Hines, Jeff Kromrey, and Liliana Rodriguez-Campos. Your mentorship has provided me with knowledge and experiences I would never have dreamed of before joining the program. Thank you to my doctoral committee for your support throughout this process. Tony Onwuegbuluzie, thank you for showing me the ways of writing and publishing. I would never been able to write a dissertation if I had not taken Carol Mullen’s class, thank you. Jane Applegate, your mentorship in writing and teacher education was exceptional, thank you.

I need to thank Liliana Rodriguez-Campos for introducing me to the world of evaluation. The lessons learned and experiences I gained from you have already proved to be invaluable. I am so glad I met you. Thank you for allowing me to work with you and learn from you. My fellow graduate students at USF were a great support group and made my experience memorable.

My family and friends have supported me every step of the way. Thank you for your support and encouragement. I want to especially thank my parents who gave me everything I needed to be where I am today. You instilled a strong work ethic and valued a good education. I will always be your professional student. Lastly I thank my husband Jeremy who has encouraged me to work hard every single day even when it was hard to work. Your own perseverance is inspiring and I thank you for your love and patience.
# Table of Contents

List of Tables ............................................................................................................. v
List of Figures ............................................................................................................. vi
Abstract ...................................................................................................................... vii

Chapter One: Introduction ......................................................................................... 1
  Background .............................................................................................................. 1
  Statement of Problem ............................................................................................... 2
  The Role of Theory in Evaluation ............................................................................ 3
  Why Does Research on Evaluation Matter? ............................................................ 5
  Purpose ..................................................................................................................... 7
  Research Questions ................................................................................................. 8
  Procedures ............................................................................................................... 8
  Importance of the Study ........................................................................................... 9
  Assumptions ............................................................................................................ 9
  Limitations ............................................................................................................. 10
  Definitions ............................................................................................................. 10

Chapter Two: Literature Review ............................................................................. 12
  Introduction ............................................................................................................ 12
  History of Evaluation .............................................................................................. 12
  Evaluation Theory ................................................................................................. 15
Classification of Evaluation Theories .......................... 18
Eight Evaluation Theorists ........................................ 24
Methods Dimension .................................................... 27
Values Dimension ....................................................... 28
Use Dimension .......................................................... 31
Empirical Studies on Evaluation .................................... 34
Chapter Three: Methodology ........................................ 43
Introduction ................................................................... 43
Research Design ............................................................ 43
Participant Selection ..................................................... 44
Instrument Selection and Design ....................................... 46
Christie’s (2001) Instrument Pilot Study and Use ................... 50
Reliability and Validity of Original Instrument ..................... 51
Changes to Christie’s (2001) Instrument ............................ 53
Focus Group Instrument .................................................. 54
Focus Group Results ....................................................... 55
Data Collection Procedures ............................................. 55
Pilot Study ..................................................................... 56
Analytic Procedures ....................................................... 58

Research Question 1: What is the breadth and depth of the preparation of school district evaluators? ...................... 58
Research Question 2: What are the reported practices of school
district evaluators? ....................................................... 58
Research Question 3: What is the relationship between evaluator preparation and evaluation practice? ..........................................................59

Conclusion ..................................................................................................................................................61

Chapter Four: Results ..................................................................................................................................62

Overview ......................................................................................................................................................62

Demographic and background information ...............................................................................................62

Research Question One ..............................................................................................................................64

Summary of Research Question One ..........................................................................................................76

Research Question Two ..............................................................................................................................77

Quantitative Analysis .................................................................................................................................77

Item grouping ..............................................................................................................................................81

Qualitative Analysis ....................................................................................................................................87

Open-ended items ......................................................................................................................................88

Summary of open-ended items ....................................................................................................................93

Interviews ....................................................................................................................................................93

Summary of Interviews .............................................................................................................................103

Summary of Research Question Two .........................................................................................................104

Research Question Three ..........................................................................................................................105

First one-way ANOVA ..............................................................................................................................105

Second one-way MANOVA .......................................................................................................................107

Third one-way MANOVA ..........................................................................................................................112

Summary of MANOVAs ............................................................................................................................115

Overall Summary of Results ......................................................................................................................116
List of Tables

Table 1 Mark’s (2008) Taxonomy of Subjects of Inquiry in Research on Evaluation ..........................................................36

Table 2 Mark’s Modes of Inquiry for Research on Evaluation ...............................................................37

Table 3 Amount of Training in Evaluation Areas .........................................................................................67

Table 4 Means, Standard Deviations, Minimum, Maximum Values for Responses from Methods, Values, and Use Items ........................................................................................................78

Table 5 Factors Representing Methods in Evaluation Practice .................................................................83

Table 6 Factors Representing Values in Evaluation Practice ...................................................................85

Table 7 Factors Representing Uses in Evaluation Practice .......................................................................86

Table 8 Descriptive Statistics for Evaluation Practice Dimensions ..........................................................87

Table 9 Means and Standard Deviations for Highest Degree Obtained and Evaluation Practice ..........................................................106

Table 10 Descriptive Statistics for Area of Highest Degree Alignment and Evaluation Practice ..........................................................108

Table 11 Means Differences and Confidence Intervals for Significant Differences among Highest Degree Alignment and Evaluation Practice .................................................................111

Table 12 Descriptive Statistics for Extent of Training in Evaluation Theory and Program Evaluation by Evaluation Practice .....................................................................................................114
List of Figures

Figure 1 Alkin and House (1992) Taxonomy ................................................................. 21
Figure 2 Eight theorists classified in the Alkin and House (1992) Taxonomy ................. 26
Figure 3 Frequency of race/ethnicity of respondents ....................................................... 63
Figure 4 Frequency of highest degree earned by respondents .......................................... 64
Figure 5 Area of highest degree earned by respondents .................................................... 65
Figure 6 Frequency of responses in which respondents completed one or more full courses ................................................................. 68
Figure 7 Amount of training in program evaluation ............................................................ 69
Figure 8 Frequency of respondents’ training in evaluation theory ..................................... 70
Figure 9 Areas of staff development training .................................................................... 71
Figure 10 Respondents’ perception of evaluation knowledge and skills .......................... 73
Figure 11 Programs for which respondents typically conduct program evaluations ...... 74
Figure 12 Type of program selected by respondents .......................................................... 76
Figure 13 Placement on the Continuum for highest degree held .................................... 130
Figure 14 Placement on the Continuum for alignment of highest degree ........................ 131
Figure 15 Placement on the Continuum for extent of evaluation training ........................ 132
Abstract

This study explored the evaluation practices of internal evaluators in public school districts in a large southern state. The individuals who conduct evaluations in school districts as internal evaluators were identified and background information was collected. The education and training in evaluation was investigated and the types of evaluations typically conducted by those individuals. Respondents \( n = 134 \) revealed conducting evaluations was a secondary role and part of their main job responsibilities. The types of evaluations carried out and the way in which evaluation was practiced were revealed. A descriptive framework of the individuals who conduct evaluations in school districts and the ways those evaluations were carried out is presented. Six dimensions were used to summarize evaluation practice: Holistic, Mixed Method Decision Making, Procedures Valued, People Valued, Users Engaged/Embodied, Evaluator as Mediator. Three one-way MANOVAs were conducted to identify differences in evaluation practice. Differences in practice were found among evaluators based on the highest degree obtained, and area of highest degree held by respondents.
Chapter One: Introduction

Background

School districts and education agencies have seen an increased demand for education evaluations since the passage of federal legislation such as the Reading Education Act (1999), the No Child Left Behind Act (2001), and the Education Sciences Reform Act (2002) (Mills, 2008; Thorton, Shepperson, & Canavero, 2007). The most well known federal legislation among educators and public school interest groups is No Child Left Behind Act (NCLB). No Child Left Behind, particularly the Title I component, was intended to provide all children with a fair and equal opportunity to receive a good education and perform proficiently on state academic achievement assessments (Department of Education, 2008). Under this legislation, each state is required to report evaluative results to the federal government regarding the programs and services mandated by federal funds.

School districts and state departments of education must show evidence of the federal dollars at work. The programs and services created under the NCLB undergo internal and/or external evaluations to show the merit or worth to the federal government (U.S. Department of Education, 2008). Evaluation reports are generated for evaluations on Title I, but also Comprehensive School Reform, Class Size Reduction, and Teacher Quality among others. In the state of Florida, each county wide school district is responsible for collecting data and reporting evidence and results of the government funded efforts.
Many school districts in Florida have a department or division focusing solely on evaluation, accountability, research, measurement and/or assessment. This directly reflects the colossal amount of collection, analysis, and evaluation reporting of data by school districts in the state of Florida. The demand is so high that in order to meet the demand to conduct evaluations, some school districts use both internal and external evaluators.

Internal evaluators include personnel hired to work specifically in an area of research, evaluation, assessment, accountability, and some other areas within school districts. Some school districts have an entire department dedicated to program evaluation. Depending on the size of the school district, a variety of different school personnel may be asked to conduct an evaluation. External evaluators include anyone hired to conduct an evaluation who does not work for the school district.

The evaluations conducted and reported to the state and federal government by school districts can be viewed as high-stakes evaluations. The evaluations are important, but the consequences of the results make them high-stakes evaluations. The evaluation results may be used to determine program continuation or elimination. This usually results in either continued financial support for programs and services or elimination of funding and termination of programs and services that children were receiving in school.

**Statement of Problem**

Due to the high-stakes placed on the evaluation results, the people who conduct educational evaluations for school districts play a critical role. The students, schools, teachers, and communities may be significantly affected by the work of education.
evaluators. The evaluators are imperative to the improvement and life of schools, attending students, teachers, parents, and communities.

Little is known about the background, training, and practices of the people conducting education evaluations. Unlike many other disciplines, the evaluation field does not have a required license or certification. If a school district or business is willing to hire someone to work as an evaluator, then a person may receive the title of “Evaluator.” Exploring the people who play the role of “Evaluator” and what the background, training and practices of these people entails, is important for the advancement of education evaluation and evaluation as a profession.

The Role of Theory in Evaluation

Early evaluations were conducted with methods and concepts from other disciplines (Shadish et al., 1991). Over the last thirty years, researchers who conducted evaluations adapted their techniques and created new ways of conducting evaluations with its own unique body of knowledge, evaluation theory. Carol Weiss (1997a) has been a contributor to the literature setting evaluation apart from research. Evaluation and research share some characteristics, but there are key differences in the use, judgment, roles, and motivation. A major difference between evaluators and researchers is the role they play. Evaluators hold a role of power. Unlike researchers, evaluators make a judgment of merit or worth at the conclusion of the study. Evaluators may feel pressure from stakeholders, funding agencies, and program directors if the evaluation is high-stakes (Weiss, 1997b). For this reason proper training in evaluation, specifically different ways evaluations can be approached, is crucial for evaluators to ensure they are conducting evaluations in an ethical and legal manner.
Evaluation theory informs evaluators about ways in which evaluations can be conducted (Alkin, 2011; Alkin, 2007; Shadish et al., 1991). The term theory is used differently in evaluation than some other fields. Many professionals use the term theory to explain phenomena, but in evaluation, the term theory is used in reference to a guide for practice (Christie, 2003). The evaluation community often uses the term theory interchangeably with the terms “models” or “approaches”. An evaluation theory provides a prescriptive guide consisting of statements explaining how an evaluation should be conducted or approached (Alkin, 2003).

As evaluation practice increased throughout U.S. history, the development of evaluation theory was promoted and continues to grow today. Evaluation scholars and theorists hope and intend for evaluation theory to guide evaluators in their practice. Theory provides suggestions for the use of different methods, sequencing and combining different methods, the types of evaluation questions to ask, and strategies for setting up evaluations (Shadish et al., 1991). A good golf player needs to know how each golf club works, what outcomes to expect from each club, and under which conditions the club works best. A golfer may gain this information from practice and experience playing golf, but if a golfer learns how each club works prior to having to use trial and error, s/he may not make mistakes that could have been avoided if information about the clubs was provided and learned. Similarly, an evaluator can benefit from knowing comparable things about different evaluation approaches. Theory is essential to the practice of evaluation, similar to the way it is in other professions.
Why Does Research on Evaluation Matter?

Evaluation has a gap in the body of literature for empirical studies on evaluation. Distinguished evaluation theorists and scholars have made repeated calls for empirical research exploring the practices of people who conduct evaluations (Christie, 2007; Christie, 2003b; Fitzpatrick, 2007; Cousins & Earl, 1999; Mark, 2008; Miller, 2010; Shadish, Cook, & Leviton, 1991; Smith, 1993; Stufflebeam & Shrinkfield, 2007; Stufflebeam & Shrinkfield, 1985). Over the last 40 years, few systematic studies have been conducted on the practice of evaluation. Scholars have shared their ideas in the growing literature base and developed evaluations by sharing personal experience (Stufflebeam & Shinkfield, 2007; Weiss, 1997). The contributions commonly found in the literature are valuable to the knowledge-base and advancement of the field; however, they provide a different type of contribution than systematic studies on evaluation practice.

Empirical studies on evaluation may be the most important resource to advance theory, but they are not found in the literature frequently as other types of contributions (Christie, 2003; Mark, 2008; Miller, 2010; Shadish et al., 1991; Smith, 1993). Few researchers have attempted to fill the gap in the evaluation literature with systematic studies of empirical data on evaluation practice. Shadish and Epstein (1987) and Christie (2003b) are among the few researchers to empirically study evaluation practice. Shadish and Epstein surveyed a sample of members from the Evaluation Research Society and the Evaluation Network, the two organizations that merged to form the American Evaluation Association in 1986. The researchers created a questionnaire to gather information related to evaluation practices including methods, timing, evaluator role, and questions. Four
discriminate patterns of evaluation practice emerged in their study: academic, stakeholder service, decision-driven, and outcome focused. The four patterns were used to describe evaluators’ perception of the purpose of the evaluation.

Christie (2003) derived a comparative framework of the similarities and differences of eight evaluation theorist (Richard Berk, Huey-tsyh Chen, J. Bradley Cousins, Elliot Eisner, David Fetterman, Ernest House, Michael Patton, and Daniel Stufflebeam) and evaluators from California State’s Healthy Start program. She created a Theory to Practice Instrument with input from eight evaluation theorists exploring three main areas of evaluation: methods, values, and use. Christie compared the practice of the internal and external evaluators in relation to the eight evaluation theorists.

Empirical research on evaluation practice is needed in two main areas. The first main area consists of the people conducting evaluations and their practices, and the second main area aims to link the practices of evaluators to the typical outcomes of those practices (Christie, 2003b; Smith, 1993; Stufflebeam & Shrinkfield, 2007). The first area focuses on identifying the population of evaluators and carving out a picture of who conducts evaluations and how they conduct evaluations. Pinpointing the background and education of the evaluators and the extent of their training in evaluation needs further research. The second area concentrates on evaluation practice. The results and outcomes, such as the way the evaluation results are used or interpreted, that typically occur with the use of specific approaches or models need further exploration. Both areas of study are important, but in order to get a sense of what is happening in the real world of educational evaluation practice, the first area should be investigated before the later.
A broad view study of evaluation practice is needed to advance the current understanding practice which can provide information for studies on the relationship between evaluation practice and evaluation theories proposed in the literature. A study investigating who conducts evaluations and the way evaluations are carried out is needed. A study of this kind would inform the evaluation community about evaluation practice and training. The extent to which methods, values, and use summarize the practice of evaluators can be uncovered. The study findings will guide poignant steps for future research. For example, once patterns of evaluation practice are identified in various settings, such as public schools, future studies can investigate the extent to which theories are present or followed (Stufflebeam & Shinkfield, 2007).

**Purpose**

There were three primary objectives overarching this study, (1) to identify the training of K-12 public education evaluators, (2) to examine the practices of education evaluators, and (3) to examine the link between training and practice. First, the background, training, and education of the K-12 public education evaluators were explored. The extent to which their job requires evaluation activities, years of experience, and type of evaluations conducted was investigated. Second, the reported practice of education evaluators was examined via the Theory to Practice instrument (Christie, 2001). Based on the self reported practices of education evaluators, a profile of a population of education evaluators was developed. Current information known about evaluation theory was used as a guide for identifying patterns of evaluation practice. This study used the Theory to Practice instrument developed by Christie (2001) in her quest to better understand the practices of California State’s Healthy Start evaluators and the way
the evaluators’ practice of evaluation compared to eight evaluation theorists. Identifying
the practice of K-12 evaluators and variability among the evaluators’ backgrounds in the
current study provides information to motivate future studies to explore the ways in
which training and background may impact the effectiveness of evaluations. The study
sought to answer three main questions:

**Research Questions**

1. What is the breadth and depth of the preparation of school district evaluators?
2. What are the reported practices of school district evaluators?
3. What is the relationship between evaluator preparation and evaluation practice?

**Procedures**

This research surveyed people who were working in school district offices of
research, evaluation, grants, accountability, or other individuals who typically conduct
program evaluations in the state of Florida. The participants received an online survey
questionnaire via their school district email. The survey questionnaire was a variation of
the *Theory to Practice* instrument designed by Dr. Christina Christie (2001) and
consisted of two sections. The first section contained items regarding education,
academic, and profession background, and the second section asked about their practice
of evaluation. The Theory to Practice instrument collected information about
methodologies used, manner in which value judgments are made, and use of the
evaluation efforts (Christie, 2003b). The background information combined with
information on evaluation practices and interviews provided data to create a descriptive
profile of the people who conduct education evaluations and how those people carried out
the evaluations.
Importance of the Study

As an initial step in a comprehensive investigation of K-12 program evaluators, the findings of this study highlight the patterns and characteristics of current practicing education evaluators, as well as documenting variability across evaluators. Evidence of evaluation training has not been linked to empirical data, or the practical use of established theories. Practices were explored for trends related to evaluation approaches found in the literature.

In addition, this research will suggest potentially fruitful avenues for subsequent inquiry. Evaluation theories, such as participatory or empowerment, have been developed and presented in the literature for evaluators to use, but the extent to which people use them or the way people use them in practice still needs further empirical study. In addition, evaluation approaches presented in the literature may not provide sufficient evidence supporting their use (Christie & Fleischer, 2010, Miller, 2010). All credible outcomes of this study should be further studied for the refinement of evaluation theory and practice (Stufflebeam & Shinkfield, 2007).

Assumptions

The researcher made several assumptions in the design of the study. It is assumed people employed by school districts in Florida will have access to their school email. The survey questionnaire asked respondents to answer the questions in response to the way they actually carry out an evaluation. It was assumed respondents who complete the questionnaire actually conducted an evaluation for their school or district and answered the items honestly.
**Limitations**

There are several limitations to consider when interpreting the results of this study. First, the amount of background information collected from participants regarding their personal and professional experiences does not encompass all possible background information on each person. Gathering additional information regarding all courses, seminars, and other training resources would contribute to the study. Second, the population in this study included evaluators who work for a school district in the state of Florida. Including all people who conduct evaluations in an education setting would enrich the information gathered and add to generalizability.

**Definitions**

**Evaluation** The systematic process of determining an object’s merit or worth based on defensible criteria (Fitzpatrick et al., 2004; Stufflebeam & Shrinkfield, 2007).

**Evaluation Approach** A term describing the ways of thinking, designing, and carrying out an evaluation, to include evaluation models (Stufflebeam & Shrinkfield, 2007).

**Evaluation Stakeholders** Individuals or groups of people who may be directly affected by the evaluation results and who have a direct interest in the program (Fitzpatrick et al., 2004).

**Evaluation theory** A framework to guide evaluation comprised of “conceptual, hypothetic, pragmatic, and ethical principles” (Stufflebeam & Schrinkfield, 2007, p. 716).

**K-12 Evaluators** Individuals who conduct program evaluation in public school districts, primarily consisting of students who are enrolled in kindergarten through twelfth grade.
Prescriptive Theory. A guide proposed by an evaluator to help other evaluators, based on reflections and experiences (Stufflebeam & Shrinkfield, 2007).
Chapter Two: Literature Review

Introduction

This study investigated the practice of evaluators in regards to evaluation theory. Thus, this literature review focuses on three main areas; history of evaluation, evaluation theory, and empirical studies of evaluation practice. The history of evaluation sections provides a brief history on the development of evaluation overtime and evaluation today. The evaluation theory section details the development of evaluation theory, classification of different evaluation theories, and specifically describes eight evaluation theories that represent the wide range of theoretical positions within evaluation. The taxonomy used to classify the eight evaluation theories is discussed highlighting the significance of methods, values, and use in evaluation practice. Lastly, empirical studies on evaluation, highlighting theories and practice of evaluation will be discussed.

History of Evaluation

The practices and roles of evaluation in education and social programs have evolved throughout history. The practice of evaluation can be linked back as far as 2200 B.C. (Shadish & Luellen, 2005; Shadish et al., 1991; Worthen & Sanders, 1987). In Ancient China, the emperor administered examinations to people wishing to hold political office, similar to today’s civil service exams (Bowman, 1989; Fitzpatrick, Sanders, & Worthen, 2004). In the early days of craft making, one of the most prestigious occupations was the signature of approval on completed swords in Japanese sword-making (Scriven, 1991). Personnel, product, and program evaluation date back to the
earliest empires and dynasties of the world. Evaluation practice has been going on for centuries. In the United States, evaluation can be traced back to the work of three men, Ralph Tyler (1935) in education, Kurt Lewin (1948) in social psychology and Paul Lazarsfield (1955) in sociology (Shadish & Luellen, 2005; Shadish et al., 1991).

The term “educational evaluation” was originated by Ralph Tyler in the 1930’s, when he described his evaluations as comparing intended outcomes with actual outcomes (Madaus & Stufflebeam, 1989). The early use of educational evaluation includes evaluation of achievement, behavior, and habits, while using instruments such as tests, scales, interviews, and observations. Ralph Tyler developed his views on evaluation focusing solely on objectives and determining whether objectives were met. He is well known for his work directing the Eight-Year Study (Smith & Tyler, 1942), where he measured achievement directly, without considering other inputs. Tyler’s technique for evaluating learning brought about change in evaluation beyond the measurement of student ability and launched new approaches to evaluation.

As the United States underwent major social and economic changes, evaluation began to undergo changes. The Great Depression came to an end, and the United States experienced a period of carefree growth. Schools increased offerings, developed new institutions, and expanded programs. Other areas such as food services, mental and health services, and industry also saw great expansion and growth. The government was spending billions of dollars on social programs. As the government budget increasingly grew through wars and oil embargos, many government leaders began to wonder how the government money was spent and the results from the efforts. The term evaluation began to take on a new meaning from Tyler’s achievement measures. A professional judgment
on the merit or worth of programs and services was greatly needed (Stufflebeam & Shrinkfield, 2007). This was the beginning of the field of evaluation as we know it today.

Up to this time (1960’s), evaluation was an optional endeavor for organizations receiving government funds. In education, some school districts conducted evaluations contingent on financial support from foundations or professional organizations, but evaluations were not yet required for government funding. Government leaders began initiating evaluation requirements for social program evaluations. A major influence for the increase in program evaluation was the 1960’s legislation that mandated, required, and funded program evaluation (Shadish & Luellen, 2005).

In the mid 1960’s, congress passed the Elementary and Secondary Education Act to ensure all children have a fair and equal opportunity for an education. This legislation, of the Elementary and Secondary Education Act, particularly the Title I section, provided large amounts of monies from the government for the public school, specifically in evaluation. Federal, state and local resources contributing to public schools began to use evaluators as a form of management aid (Smith, 1983). Evaluations were used as an administrative guide for decision-makers. The large amounts of money available for evaluations created an incentive for people to conduct them.

Many educational researchers took on roles as evaluators in the mid 1960’s, because they had relevant expertise in social science methods (Shadish & Luellen, 2005; Smith, 1983). After flourishing in the late 1960’s and 70’s, as the 80’s approached, evaluation developed into a new profession. The Evaluation Research Society (ERS) and Evaluation Network (ENet), two professional societies emerged, and several evaluation journals also emerged including the American Journal of Evaluation, Evaluation and
Program Planning, and New Directions for Evaluation (Shadish & Luellen, 2005). People were defining themselves as evaluators, and created communities and venues to share scholarly information pertaining to evaluation.

Evaluation Theory

While evaluation can be traced back centuries, Ralph Tyler’s Eight Year Study in the 1930’s is the most well known systematic evaluation in United States history. Ralph Tyler developed his views on evaluation by focusing solely on objectives. In his Eight Year Study, he measured achievement directly without considering other inputs. In the 1960’s, new legislation mandated evaluation of new programs introduced in schools (House, 1980). Tyler’s technique for evaluating learning needed to evolve beyond the measurement of student ability. The evaluation requirements set by the Great Society programs in 1965 brought about tremendous growth in the field of evaluation. People conducting evaluations began to create a body of knowledge focusing on the process of evaluation, instead of providing personal opinions.

Evaluation has been conceptualized in a variety of different ways. Different approaches used by evaluators have stimulated the creation of a collection of evaluation theories. The evaluation community uses the term theory differently when referring to evaluation theory, than many other common uses for the term theory. An evaluation theory is a prescriptive guide with statements explaining how evaluation should be conducted (Alkin, 2003). When used to describe evaluation theory, the term “theory” is used interchangeably with the terms “models” or “approaches” in evaluation (Alkin, 2004). Many other theories, are designed to explain phenomena, in contrast evaluation theory is not designed to explain phenomena, but to guide practice. This is why
evaluation theories can be thought of as models or approaches to evaluation. Theories contain guidelines for the evaluators concerning the evaluation focus, evaluation questions, implementation procedures, and use of the results (Christie, 2009). Evaluation theory is the knowledge base for the professional field of evaluation.

Evaluation theory is a general term to describe the theoretical writings focusing on experiences conducting evaluation. Two types of evaluation theories are found in the literature: prescriptive and descriptive (Alkin & Ellet, 1990). Prescriptive theory is conceptual and identifies critical components necessary to properly conduct evaluation. Descriptive theory is empirically derived and explains specific evaluation activities that have taken place. Models with empirical composition help build an understanding of when certain approaches should be used, under what conditions specific models work best, and the types of outcomes evaluators can expect (Christie, 2003). Empirically derived theories are fundamental to advance prescriptive theories of evaluation (Cousins & Earl, 1999; Smith, 1983).

A variety of different prescriptive theories have been proposed in the literature. Some of the theories are general theories of evaluation, while some are specific to the field in which the program under evaluation is housed (Shadish, 1998). Evaluation theory provides an expansive methodology. The theories prescribe ways of thinking, designing, and carrying out evaluations including information on “public policy, value theory, and theory of use” (Shadish et al., 1991, p. 31). Theories combine different evaluation procedures and emphasize and prioritize different components of evaluation (Christie, 2009; Christie, 2003). Hence, theories help evaluators by providing a framework to structure their work. For example, evaluators may look to theory to determine whether or
not to include stakeholders, the number of stakeholders to include, or methods for data collection.

Over the past 40 years, people who were conducting evaluations came from a variety of disciplines (Shadish et al., 1991). These evaluators adapted their techniques and methods to meet the needs of each evaluation. As evaluation emerged as its own discipline, the evaluation literature expanded. Many different theoretical approaches emerged, primarily as a result of evaluators’ perception of the role of evaluation and their role as evaluators. An agreed upon goal for evaluators is to help inform the program decision makers (Christie, 2003). The approach used when conducting evaluations describes the theorists’, who wrote the theory, thoughts on the primary role of evaluation.

Some theories suggest evaluation should empower individuals who have a role in the program undergoing evaluation. Under this point of view, the success of an evaluation would then be determined by the extent to which those individuals were indeed empowered. The work of David Fetterman (1996) exhibits these ideas. Fetterman’s work focuses on use, specifically an empowerment evaluation approach.

Another view to evaluation is to provide evaluation results that will support action (Patton, 2008). Patton strongly supports utilization-focused evaluations. In both of the examples provided, Fetterman and Patton focus on the use but distinguish two different ideas within the use of evaluation results. It should not be implied that Patton opposes empowerment evaluation, nor does it imply that Fetterman opposes utilization of evaluation findings. Fetterman and Patton are examples of the many theories existing in evaluation today.
Many people have contributed to the evaluation knowledge base in the area of evaluation theory. Fetterman and Patton are just two examples. The plethora of evaluation theories proposed in the literature has stimulated researchers of evaluation to create ways to organize different evaluation theories. Evaluation theories are typically traced back to the person or people who are known for writing about the theories. The people who have proposed different evaluation theories and contributed to the evaluation literature can be identified as evaluation theorists. Evaluation theorists are individuals who contributed to the literature with a strong commitment to a specific theoretical orientation (Alkin, 2011; Alkin, 2004). Evaluation theories are primarily made of the ideas and experiences of prominent evaluators. For this reason, evaluation theories are associated with the evaluator or theorist.

Some scholars did not intend to take a theoretical position within evaluation, but from an onlooker’s point of view the scholar appears to have taken a specific theoretical position. Over time, people have made a variety of attempts to organize evaluation theories. Organizational frameworks use different criteria to classify each theory. In any type of taxonomy or organization framework subjective decisions are made by the developer(s).

**Classification of Evaluation Theories**

Over the last 35 years, a variety of attempts have been made to classify evaluation theorists and the evaluation theories they propose. Some of the earliest classification frameworks include Worthen and Sanders (1973) and House (1978). Organizing theories into categories and taxonomies allow for comparisons and differentiation of theories in a systematic manner. Taxonomies may also help theorists understand relationships among
theories and further describe or clarify how their theory is perceived (Alkin, 2004). Today, a variety of existing taxonomies can be found.

Worthen and Sanders’s (1973) taxonomy has been revised to include current advances in evaluation theories (Fitzpatrick et al., 2004). The updated framework outlines five categories for organizing evaluation approaches: objectives-oriented, management-oriented, consumer-oriented, expertise-oriented, and participant-oriented. Objectives-oriented theories focus on meeting set objectives identified at the start of the program. Tyler’s (1935) eight year study is an example of an objective-oriented evaluation approach. The second category is management-oriented theories. The primary purpose of management-oriented theories is to provide information to aid in decision making throughout all stages of program development. Stufflebeam’s CIPP evaluation model is an example of a management-oriented evaluation approach (Fitzpatrick et al., 2004).

Consumer-oriented theories provide information to aid individuals in the purchase of products. Scriven (1991) has made major contributions to consumer-oriented evaluation through his Key Evaluation Checklist. Expertise-oriented theories rely on a subjective professional judgment of quality based on someone who is considered an expert. Historically the expertise approach has been widely used in institution or program accreditation. The expertise approach relies solely on professional judgment (Fitzpatrick et al., 2004). The fifth category in Fitzpatrick and associates’ (2004) model is participant-oriented theories. Participatory approaches involve stakeholders throughout the evaluation process (Cousins & Earl, 1995). The needs of the stakeholders as an audience serve as the primary concern in participatory evaluation approaches.
Fitzpatrick and associates’ classification system was developed based on House’s (1983) utilitarian to intuitionist-pluralist evaluation dimension (Fitzpatrick et al., 2004). In the development of their taxonomy, the authors had difficulty classifying individual evaluation approaches. Many approaches do not fit nicely into only one category, but fit into many different categories. Under this taxonomy, approaches are classified into the five categories based on the main evaluation questions addressed and organization of the evaluation.

House’s (1978) classification of theories also consists of five components comparing theoretical assumptions. The theoretical assumptions include: the primary audience of the evaluation; evaluation theory consensus; methodology (data collection); overall expected outcome; and the questions the theory usually aims to address. Shadish, Cook, and Leviton (1991) created a taxonomy consisting of three stages of theories. Stage one theories included social problem solving and scientific rigor, stage two theories involved alternative approaches with focus on use, and the third stage of theories included approaches with a strong emphasis on integrating earlier parts of the evaluation. Shadish and associates (1991) used their taxonomy to explore changes in assumptions and prescriptions over time.

One of the more notable classification frameworks in evaluation is the Alkin and House (1992) taxonomy. Alkin and House (1992) organized evaluation approaches into three dimensions: methods, values, and uses. Based on reviews of evaluation definitions found in the literature, Alkin and House (1992) found three main emphases: “(a) the collection and analysis of data (methodology); (b) ways in which valuing is done and judgments are made (values); and, (c) the broad or specific purposes for providing
evaluation information (uses)” (Alkin & House, 1992, p. 463). Each dimension, methods, values, and use, has a continuum in which theories fall. The Alkin and House (1992) classification system is illustrated in Figure 1.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Continuum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>Quantitative &lt;––&gt; Qualitative</td>
</tr>
<tr>
<td>Values</td>
<td>Unitary &lt;––&gt; Plural</td>
</tr>
<tr>
<td>Uses</td>
<td>Instrumental &lt;––&gt; Enlightenment</td>
</tr>
</tbody>
</table>

Figure 1. Alkin and House (1992) Taxonomy

The Alkin and House (1992) taxonomy is the taxonomy used in this study to investigate the practice of evaluators. This framework was chosen over others because of the extensive elaboration and acceptance of this taxonomy since Alkin and House published it in 1992. A discussion of the Alkin and House (1992) taxonomy and its use in this study follows.

Alkin and associates have continued to develop this comparative framework using the methods, values, and use dimensions (Alkin & House, 1992; Alkin, 2004; Alkin & Christie, 2004; Christie & Alkin, 2008). Within each dimension lays a continuum. The methods dimension ranges from quantitative to qualitative, the values dimension stretches from unitary to plural, and the uses dimension ranges from instrumental to enlightenment.

The methods dimension continuum represents quantitative and qualitative approaches. In the past, evaluators used quantitative methods to measure academic achievement, conducted randomized experimental studies, and analyzed standardized test
scores (Alkin & House, 1992). During the evaluation boom in the late 1960’s and 1970’s, evaluators faced problems conducting large scale evaluations. Large scale evaluations with multiple program sites per program were not finding consistent results (Alkin & House, 1992). Some programs worked well at one site but did not work well at others. Evaluators needed to collect data beyond test scores to identify components in need of improvement. Qualitative methods yielded useful techniques to acquire the information evaluators needed. As evaluation evolved beyond measuring achievement, qualitative methods became increasingly accepted and appreciated.

Utilizing both quantitative and qualitative methods, known as a mixed-method approach is commonly used in evaluation today (Alkin & House, 1992; Stufflebeam & Schrinfield, 2007). Evaluation has evolved from a single method of data collection to a broader approach in which evaluators use multiple methods for the collection of data.

Evaluations require a value or values to be set in order for a judgment of merit or worth to be made. According to Alkin and House (1992), the values dimension ranges from the evaluator using a single criterion (unitary) to using multiple criteria (pluralist). Traditionally, evaluators would look to program goals for determining criteria and may work with the program manager to use a goal-based or managerial-oriented approach. Some evaluators rejected approaches following program or company goals as the focus of an evaluation. Scriven (1980) investigated the management approach to clarify the establishment of values in evaluation. Scriven concluded evaluators must assess the program goals before using the goals as criteria for determining merit or worth.

The type of criteria must be derived by evaluators and should be appropriate for the program. For example, if an educational program was designed to limit students with
exceptional needs from achieving success, then regardless of the criteria set, the program would not be a good program. Evaluators must establish nonarbitrary criteria to be described and defended.

Many programs are evaluated on multiple criteria rather than one single value. Multiple criteria allow more than one interest to be addressed, but the evaluators must decide how to make an overall judgment of merit and worth based on the different measures. Some evaluators make final judgments, while others present only the data and allow the stakeholders and audience to decide. Beyond the final judgment, evaluation has evolved from unitary to pluralist.

The third dimension found in evaluation definitions pertains to the use of evaluation results. Alkin and House (1992) describe evaluation uses along a continuum ranging from enlightenment to instrumental. Traditionally evaluation results were used to understand something and provide information regarding the specific program. This would fall on the enlightenment side of the continuum. On the other end of the continuum, conducting evaluation to gather information for decision making or policy development is considered instrumental use.

Evaluations conducted for enlightenment purposes are similar to those of research studies and are intended for similar academic audiences. Enlightenment evaluations produce knowledge on the program under evaluation (Alkin & House, 1992). Instrumental evaluations directly inform stakeholders and provide the information needed to make decisions.

Evaluators consider methods, values, and use when designing and implementing evaluation. Quantitative methods, qualitative methods, or mixed-method, a combination
of the two are used to provide data in evaluation. Evaluators may choose one or multiple criteria to judge merit or worth of a program. The criteria may depend on the methods used. Evaluations may be conducted for enlightenment or instrumental uses. The way the evaluator addresses the audience reveals some information about the use of the evaluation results.

Alkin and House (1992) first published their three dimensions in the 1992 edition of the Encyclopedia of Educational Research. Since then, Alkin (2004) developed an evaluation theory tree using a branch of the tree for each of the three dimensions. The tree was designed to organize and visually display evaluation theoretical stances and provide a better understanding of evaluation theory. Alkin placed evaluation theorists on the branch (methods, values, use) reflecting their primary theoretical orientation. Christie and Alkin (2008) re-examined the tree in 2006 and 2008 to include newer contributors to evaluation theory and move any theorists who changed their theoretical stance since the original tree was created or were misplaced the first time around. It is important to note that all of the theorists from the original tree stayed on their original branch (dimension), but may have moved along the continuum. Alkin and House’s (1992) taxonomy of the methods, values, and use dimensions continues to be an accepted way to organize evaluation theories and theorists.

**Eight Evaluation Theorists**

The current study uses eight theories proposed by eight different evaluation theorists. The eight theorists and their theories presented here were used by Christie (2001) to develop the Theory to Practice Instrument. Additional information regarding the development of the instrument is presented in the chapter three. The eight evaluation
theorists were chosen by Christie (2001) because of the wide range of their theoretical positions within evaluation. Each of the theorists is well known and published in the evaluation community. Although the theorists were intentionally selected to represent a broad perspective of evaluation theory and practice, it is important to note there are many other theorists who have contributed to the literature and knowledge base in evaluation.

To highlight the differences of the eight theorists the Alkin and House (1992) taxonomy was used. The eight theorists included: Richard Berk for Robert Boruch, Huey-tskyh Chen, J. Bradley Cousins, Elliot Eisner, David Fetterman, Ernest House, Michael Patton, and Daniel Stufflebeam. Robert Boruch was unavailable to provide feedback and information to contribute to the development of the Theory to Practice (Christie, 2001) instrument, therefore based on the close working relationship between Boruch and Berk, Boruch authorized Berk to provide information on his behalf (Christie, 2001). Picturing the eight theorists in Alkin and Houes’s (1992) taxonomy, both Boruch and Eisner are found in the methods continuum. Within the methods dimension, the quantitative and qualitative continuum contains Boruch on the quantitative left and Eisner on the qualitative right (Christie, 2001). In the values continuum, lays House, Fetterman, and Chen. The values continuum arrays from plural with Fetterman and House on the left, to unitary including Chen on the right. The third continuum of Alkin and House’s (1992) taxonomy represents use. This dimension spans from instrumental to enlightenment.

The final three theorists belong on the instrumental end of the continuum. Cousins, Patton, and Stufflebeam all engage stakeholders during their evaluations, and the three can be sequenced at the end of the continuum with Cousins engaging the most
users in his evaluations, Patton is next, and Stufflebeam typically engaging the least amount of users (Christie, 2001).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Continuum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>Boruch</td>
</tr>
<tr>
<td></td>
<td>Quantitative</td>
</tr>
<tr>
<td>Values</td>
<td>Chen</td>
</tr>
<tr>
<td></td>
<td>Unitary</td>
</tr>
<tr>
<td>Use</td>
<td>Cousins, Patton, Stufflebeam</td>
</tr>
</tbody>
</table>

Figure 2. Eight theorists classified in the Alkin and House (1992) Taxonomy.

Alkin and House (1992) classified each theorist into the one dimension that best captures their primary theoretical orientation, however it is important to note that each theorist does not think the other dimensions are unimportant or not part of their theory. The perceived purpose of evaluation, the general approach to conducting evaluations, and the methods used to implement a specific approach were criteria used to place each theorist in the Alkin and House (1992) taxonomy (Christie, 2001).

Placement in the taxonomy was not based on empirically derived data, but categorized based on direct positions stated in writings or inferences made by Christie (2001) from the theorists’ work and writings. The purpose of placing theorists in a dimension was to demonstrate the diverse perspectives among the eight theorists. Many of the theorists have taken a theoretical perspective holding more than one dimension.
from the Alkin and House (1992) taxonomy. Placing the theorists into just one dimension was not to limit the contributions or theories proposed by the theorists. The idea of the theorists taking on multiple perspectives implies the theorists provided a rich contribution to the theory of evaluation. The following section of the literature review provides the rationale for each theorist’s placement within the taxonomy for this study.

**Methods Dimension**

Robert Boruch and Elliot Eisner are the two theorists representing the *methods* dimension in this study. Boruch is found on the quantitative side focusing on randomized field experiments. Eisner is placed on the qualitative end focusing on the process of the evaluation.

Boruch was known for conducting evaluations with quantitative methods, primarily randomized field experiments. He viewed evaluation as conventional scientific research (Alkin, 2004). Boruch proposed the randomized field experiment as the best way to allow for the least ambiguous approximation of program effectiveness (Boruch, Synder, & DeMoya, 2000). In a randomized field experiment, individuals or groups of individuals are randomly assigned to one or two treatment groups (Boruch, McSweeney, & Soderstrom, 1978). The method allows for a fair comparison of groups and allows legitimate conclusions to be made regarding the role of chance in the interpretation of treatment results (Boruch et al., 2000).

Boruch found randomized research designs to be the best way to objectively measure program effectiveness and objectively evaluate programs. His strong commitment to the randomized field experimental approach sets him in the *methods* dimension on the quantitative end of the continuum.
On the other side of the *methods* dimension, is Elliott Eisner. Eisner takes more of a holistic approach to evaluation. When making judgments of quality, he believes evaluators must be educational connoisseurs and critics (Eisner, 1985). According to Eisner (1985) evaluation can occur without measuring or testing because the job of the evaluator is to describe. Eisner’s approach to evaluation includes the use of qualitative methods. Qualitative methods allow evaluators to gather the information necessary to make judgments.

Acting as a connoisseur and critic places the evaluators as the principal methodological tool (Eisner, 1985; Miles & Huberman, 1994). This requires the evaluator to be an expert in the program’s topical area. Eisner’s work focuses on both methods and values in his work. However, he is placed under the *methods* dimension in the Alkin and House taxonomy because of the necessity for the connoisseurs and critics to use qualitative methods in order to make judgments (Christie, 2001). Under Eisner’s theory, the connoisseurs and critics making value judgments are using qualitative methods to gather their information.

**Values Dimension**

The work of Huey-tsyh Chen, David Fetterman, and Ernest House represent the *values* dimension in this study. Chen is known for his *Theory-Driven* approaches to evaluation which utilize program theory to achieve the desired goal(s) for a program. Fetterman and House both use an advocacy oriented model, but still have notable differences in their approaches. Fetterman’s methods are referred to as *Empowerment Evaluation* and House’s methods are known as *Deliberative Democratic Evaluation*, both found on the plural end of the *values* continuum.
Chen’s *Theory-Driven Evaluation* uses program theory to guide the evaluation. Chen defines program theory as “a specification of what must be done to achieve the program’s desired goals, the important aspects that may be anticipated, and how these goals and impacts would be generated” (1990, p. 43). The program theory for the program under evaluation is used by the evaluator to identify program goals and values to state criteria for program effectiveness.

Theory-driven evaluation is sometimes confused with method-driven evaluation. Method-driven evaluation uses pre-determined research steps associated with the design’s research method. In contrast, values are the highlight of theory-driven evaluation (Chen & Rossi, 1992). This makes program theory the sole information source in theory-driven evaluation and therefore the unitary source for judgments and values. Chen’s approach places him on the unitary end of the *values* continuum. If Chen was hired to conduct an evaluation, he would use the underlying theory for the program under evaluation to determine the criteria for which value judgments should be made.

The plural end of the *values* dimension includes Fetterman and House. Fetterman represents *Empowerment Evaluation*, which is extracted from areas involving self-determination (Christie, 2001). Through the program evaluation process, Fetterman encourages individuals to take part in the evaluation by training the participants to evaluate the program themselves. This approach to evaluation is collaborative and the lead evaluator acts as a teacher or coach throughout the process.

Fetterman was difficult to place into one dimension. He falls strongly to the right on all three dimensions. His placement is on the qualitative end of the methods, plural end of the values, and the instrumental end of the use dimension. In this study he is
categorized in the values dimension because of the emphasis on empowerment evaluation to encourage self-determination among the program participants. Empowerment evaluation requires the collaboration of participants, which means the use of the evaluation results should increase if collaboration is achieved. The main focus of empowerment evaluation is to empower the program participants and instrumental use is a bi-product of the approach.

Similar to Fetterman’s *Empowerment* approach, House’s approach promotes social justice through *Deliberative Democratic Evaluation* (House & Howe, 1999). The deliberative democratic evaluation framework allows all relevant interested parties to participate throughout the evaluation process. This model fosters an equitable and democratic process by approaching evaluation with pluralistic values.

House (2004) described deliberative democratic evaluation as a way to reduce evaluator bias as much as possible by inviting extensive communication with stakeholders and by “promoting extensive deliberation about the study’s conclusions,” (House, 2004, p. 220). House conducts evaluations with input from all relevant interests not just the people who run the program.

Of the eight theorists, House has written about the weaknesses of placing evaluation theorists into three categories. He does not dispute his placement on the values branch of Alkin’s (2004) tree. However, he reinforces the cross-influences one theorist may have within the dimensions and other areas not included in the three dimensions. If House were conducting an evaluation he would use a deliberative democratic approach which gathers information from multiple viewpoints to determine the values in an evaluation.
Use Dimension

Evaluations are conducted to inform decision making. This notion means evaluation results are intended to be used. Alkin and House’s (1992) *use* dimension is represented by J. Bradley Cousins, Michael Patton, and Daniel Stufflebeam. The three theorists believe *use* is a critical purpose of evaluation, and they all fall on the instrumental end of the *use* continuum.

Stufflebeam is one of the most well known scholars in evaluation theory. After trying many different evaluation models, Stufflebeam decided the central purpose of program evaluation is to improve the program (Stufflebeam, 1983). Similar to the The Program Evaluation Standards (Joint Committee on Standards for Educational Evaluation, 1994), Stufflebeam (1983) used the term program generically to refer to the object of evaluation. Scriven (1991) later coined the term evaluand to describe the object under evaluation. Stufflebeam believed “evaluation should help program personnel make and defend decisions keyed to meeting beneficiaries needs” (Stufflebeam, 2001, p. 57). This evaluation approach is also referred to as “decision-oriented” evaluation. Stufflebeam created his CIPP model, to represent the four types of evaluation: Context, Input, Process, and Product. Many people who have minimal training in evaluation know about Stufflebeam’s CIPP model.

Stufflebeam’s work also includes The Program Evaluation Standards (Joint Committee on Standards for Educational Evaluation, 1994) which includes four factors: Utility, Feasibility, Propriety, and Accuracy. The standards were created to ensure evaluations meet the needs of intended users, are conducted realistically, legally and ethically, and reveal adequate and accurate information to determine the merit or worth of
the program under evaluation. As one of the contributing authors to The Program Evaluation Standards (1994), Stufflebeam holds them in high regard.

Under Stufflebeam’s evaluation approach, a representative group of individuals are identified as stakeholders and these stakeholders are included throughout the evaluation process (Stufflebeam, 2007). If Stufflebeam were conducting an evaluation, he would engage the stakeholders and maintain continual communication to ensure the evaluation aids in decision-making. Stufflebeam addresses stakeholders’ questions through the evaluation and provides results for direct use.

Patton’s primary evaluation approach is similar to Stufflebeam’s by including stakeholder involvement. Patton is most well known for his *Utilization Focused Evaluation*, which is intended to provide the information needed to directly impact the program under evaluation (Patton, 2007). The stakeholder inclusions distinguish Patton’s approach from Stufflebeam’s approach. Patton’s approach defines stakeholders as people “who have a stake- a vested interest- in the evaluation” (Patton, 2007, p. 61). From the considerable amount of people who could be identified as stakeholders, Patton narrows down the group to a smaller number called “primary intended users” (Patton, 2007).

The smaller group of the “primary intended users” includes only those that are involved in decision making and have the capacity to utilize the findings in the organization from which the program under evaluation resides. Including people who are in a position to use the evaluation findings greatly increases the likelihood of the findings being used (Patton, 2007). If Patton were conducting an evaluation he would use multiple criteria for value judgments, any methods necessary to answer the evaluation questions,
and he would focus on including primary users to ensure the evaluation results will be useful and used.

Branching off of Patton’s utilization framework is Cousins’s *participatory evaluation* (Cousins & Earl, 1999). Similar to Patton, Cousins is concerned with utilization of evaluation findings and engages primary users, but the way he engages primary users is different from Patton. Cousins’s approach places the evaluator(s) and primary users as collaborators in the evaluation process. The primary users and evaluators work together to conduct the evaluation as a joint partnership. According to Cousins and Earl (1995), participatory evaluation engages users who are committed to program improvement and as a result increases utilization.

Participatory evaluation should not be confused or thought of synonymously with empowerment evaluation. The main goal of empowerment evaluation is to empower the individuals within the program under evaluation. The main goal in participatory evaluation is to engage participants to increase utilization of the evaluation findings.

In general, the eight theorists have been consistently classified into their selected dimensions (e.g., Alkin, 2004; Christie, 2001; Christie & Alkin, 2008). Some changes have occurred over time. Alkin (2004) placed Chen on the methods branch in his evaluation tree, but leaning towards the value branch (Christie & Alkin, 2008). In addition, Alkin (2004) placed Eisner on the values branch. This does not seem to present any problems for their placement in this study. The methods branch on the evaluation tree presented by Alkin (2004) does not contain a theorist who strongly represents qualitative methods. Using Eisner as a contributor for qualitative methods is plausible especially considering his use of qualitative methods to determine values in evaluation.
David Fetterman was placed on the values dimension by Christie (2001) and later placed onto the use branch in Alkin’s (2004) evaluation tree and again in Christie and Alkin’s (2008) evaluation tree revisited. Fetterman does not dispute his place on the use branch, but he does not want to be confined to one area of evaluation (Fetterman, 2004). Fetterman has made contributions to all three areas of evaluation theory and provides a copious perspective of evaluation theory and practice.

Richard Berk, Huey-tsyh Chen, J. Bradley Cousins, Elliot Eisner, David Fetterman, Ernest House, Michael Patton, and Daniel Stufflebeam are eight theorists who represent a broad view of evaluation theory. The Alkin and House (1992) taxonomy provides a nice framework to organize the models, philosophies, and approaches proposed by the eight theorists. There are other views and theoretical orientations present in the evaluation community and the inclusion of eight theorists and theories is an important consideration in this study.

**Empirical Studies on Evaluation**

Over the past 35 years, researchers have conducted studies on evaluation (e.g., Bernstein & Freeman, 1975; Christie, 2003; Lipsey, Crosse, Dunkle, Pollard, & Sobart, 1985; Patton et al., 1977; Shadish & Epstein, 1987; Torres, Preskill, & Piontek, 1997; Weiss, 1977). The existing empirical studies on evaluation found in the literature are infrequent (Mark, 2008). The literature contains information about evaluation shared by scholars in journals such as the American Journal of Evaluation and New Directions for Evaluation, but the majority of the articles do not focused on a systematic study of evaluation itself. A content analysis of articles published in the American Journal of Evaluation (AJE) and New Directions for Evaluation (NDE) from 2004-2006 found 11
articles from AJE and three articles from NDE contained information describing the evaluand, evaluation method used, and results of the evaluation (Christie & Fleischer, 2010). Researchers in evaluation must come together and pursue a steady agenda of research on evaluation. Stimulating a positive outlook for the continued study of evaluation is the recently emerged Research on Evaluation topical interest group (TIG) in the American Evaluation Association.

The calls for research on evaluation are not new. Empirical studies with supporting data are needed to answer the lingering questions about evaluation and evaluation theories (Miller, 2010; Mark, 2008; Shadish et al, 1991; Smith, 1993). Some of the questions research on evaluation may answer include: whether a specific evaluation approach meets its promises; what type of evidence provides clients with more useful information; and, what specific applications look like in practice (Mark, 2008). Empirical evidence can provide information to help answer these questions and others of similar nature.

A variety of areas in evaluation have been identified in the literature for research on evaluation. Mark (2008) has organized the general calls for research on evaluation into a taxonomy to help guide future studies, shown in Table 1. The research on evaluation taxonomy is aimed to identify gaps in the evidence base, aid in the planning of additional research on evaluation, and synthesize the existing studies on evaluation (Mark, 2008). Organizing the research on evaluation may positively stimulate discussions in other areas of evaluation without intention.
Table 1

*Mark’s (2008) Taxonomy of Subjects of Inquiry in Research on Evaluation*

<table>
<thead>
<tr>
<th>Concept</th>
<th>Evaluation Context</th>
<th>Evaluation Activities</th>
<th>Evaluation Consequences</th>
<th>Professional Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The circumstances within which evaluation occurs</td>
<td>The procedures used in planning, carrying out, and disseminating evaluation</td>
<td>Changes that do (or do not) occur as a result of evaluation</td>
<td>Issues involving the structure, norms, and continuation of the field of evaluation</td>
</tr>
</tbody>
</table>

Mark’s (2008) framework identifies four categories in which research on evaluation may fall. The four areas presented by Mark are not intended to limit future studies to these categories, but present researchers with a minimum of four subjects to start. Similar to other taxonomies, a single study does not have to fit into one category. Many studies encompass more than one category, and/or look at relationships between categories.

While thinking about research on evaluation and organizing different types of studies, Mark (2008) paired his Subjects of Research on Evaluation (shown in Table 1 with inquiry modes shown in Table 2). Using the inquiry modes taxonomy, studies are organized into four categories: values inquiry, descriptive, causal analysis, and classification (Mark, 2008). Several different studies on evaluation exist in the current literature reflecting a variety of research investigating the work of evaluators. A review of empirical studies on evaluation using the inquiry modes taxonomy follows.
### Table 2

*Mark’s (2008) Modes of Inquiry for Research on Evaluation*

<table>
<thead>
<tr>
<th>Values Inquiry</th>
<th>Description</th>
<th>Causal Analysis</th>
<th>Classification</th>
</tr>
</thead>
</table>

In the values categories, Segerholm (2003) examined evaluation context by studying evaluations in a national context. This study compared the way evaluations were carried out in different states reflecting the values of stakeholders in different geographic locations. Segerholm (2003) found four phases in the evaluation process: initial, implementation, results, and utilization, that are distinguishable from each other.

Descriptive studies on evaluation include Torres and associates (1997), Petrosino (2003), and Agodini and Dynarsky (2004). Torres and associates (1997) investigated the communication and reporting of evaluation results by surveying members of the American Evaluation Association. Participants indicated political and organizational restrictions in communicating and reporting successfully. This study gathers valuable information about the practice of evaluators after the evaluation has been conducted. The reporting of evaluation results is important, but this type of study yields insufficient amount of information about the implementation of evaluation.

Petrosino (2003) reviewed abstracts from bibliographic databases to estimate the relative frequency of randomized controlled trials (RCT). This study found RCT was used in nearly 70% of interventions studies on children’s healthcare, but RCT was used...
only 6-15% in K-12 education and juvenile justice studies. Similarly, Agodini and Dynarsky (2004) studied the consequences of using a comparison group based on propensity-score methods as part of the evaluation methods. Their study found no consistency in using propensity-score methods compared to experimental methods when evaluating drop-out prevention programs.

Both Petrosino (2003) and Agodini and Dynarsky (2004) have studied evaluation by reviewing reports or articles. This type of study is important but does not paint the complete picture of evaluation practice. Typically reports and articles do not report all of the details involved in the design and implementation of evaluation the way it occurs in practice (Christie, 2003; Stufflebeam, 2007).

Causal analysis studies estimate evaluation effects, or identifies how a program has an effect. Two studies in the literature reflect a causal analysis: Henry (2004) and Campbell and Mark (2006). Both studies use causal analysis to examine the consequences of evaluation. Henry (2004) used statistical controls and quasi-experimental methods, and Campbell and Mark (2006) used experimental methods to assess evaluation outcomes. Campbell and Mark’s (2006) study tested the effects of two factors regarding stakeholder dialogue and negotiations: accountability audience (homogeneous versus heterogeneous) and dialogue structure (instructions for problem-solving versus no instructions) for stakeholder discussions. In their experiment, Campbell and Mark (2006) engaged undergraduate stakeholders in a dialogue about their university’s alcohol policy. After conducting 2 X 2 repeated measures ANOVAs they found a heterogeneous audience and problem-solving dialogue had a positive impact on valued outcomes with more effective dialogue, compared to other combinations.
Conducting a causal analysis is an example of a type of research on evaluation that can stem from the results of the current study. A causal analysis comparing the evaluation consequences of different school district evaluations, and the way the evaluations were carried out would be a good follow up study to the current study.

Classification studies are represented by Cousins, Donohue, and Bloom (1996), Shadish and Epstein (1987), and Christie (2003). Cousins and associates (1996) examined different types of participatory evaluation. Their study aimed to classify different types of participatory evaluation based on self-reported practices from evaluators. Shadish and Epstein (1987) surveyed registered members of the Evaluation Research Society and the Evaluation Network asking about their evaluation practices. These two organizations combined to form the American Evaluation Association. Shadish and Epstein’s (1987) self-authored survey questionnaire explored evaluation practice related to methods, the evaluator’s role, timing of evaluation, evaluation questions, and methods for facilitating use. Their study revealed four discriminating factors of evaluation practice: academic, stakeholder service, decision-driven, and outcome. The academic factor consisted of evaluations meant to fulfill basic interests of the evaluators themselves, similar to a typical research study. Stakeholder service usually involved payment from a client for the evaluation service with the evaluator’s role to provide useful information for the client. The decision-driven factor consisted of the evaluator developing questions to be used for program decisions, and the outcome factor viewed the evaluator as the expert to judge the merit or worth of a program.
Shadish and Epstein’s (1987) and Christie’s (2003) studies both explore the practice of evaluators and have been set under the classification category because they both seek to categorize the practice of evaluators. Christie (2003) surveyed evaluators from California’s Healthy Start program with the *Theory to Practice* instrument. The survey was developed by the author with contributions from eight eminent evaluation theorists. Christie (2003) mapped out the reported practices of evaluation theorist and compared the practice of the Healthy Start evaluators to the practices of the eight theorists. The study found two underlying dimensions distinguishing the practice of the theorists: scope of the stakeholder and method proclivity. Christie did not classify the evaluator’s responses, but this study was included in this category because it comes very close to classification (Mark, 2008).

The subjects of inquiry and inquiry modes taxonomies should both be considered when thinking about research on evaluation (Mark, 2008). Researchers should not limit their studies to those proposed by Mark (2008) in his research on evaluation taxonomies. The different organizational frameworks shown by the subjects of inquiry and inquiry modes taxonomies are intended to bring about future studies which may fit into the proposed categories or stimulate the development of other areas not mentioned.

The current research on evaluation existing in the literature provides a base for further studies. The existing research on evaluation focuses on specific areas of evaluation or specific evaluators. Research on the training of people who conduct evaluation and how people implement evaluation in general is still needed. The research on evaluation conducted and presented in this study will carve a picture of the education
evaluators and their training and practice. The results will provide evidence-based data pertaining to evaluation training and practice.

The current study encompassed two of Mark’s overarching categories for subjects of inquiry, evaluation activities and professional issues. Studying the practices of evaluators falls into the evaluation activities category from Mark’s (2008) taxonomy. To gain a better understanding of evaluation practice, the work of evaluators and their implementation of evaluations must be studied. Studying practice provides insight into frequently used methods, the type of information gathered, and evaluation interpretation (Christie, 2003).

Participants in the current study were asked to provide information regarding their evaluation training and experience. The background information falls into Mark’s (2008) professional issues category. Previous studies have revealed that many people who conduct evaluations lack formal training in evaluation and do not identify evaluation as their primary field of work based on empirical investigation of California Healthy Start evaluators and members of the Evaluation Research Society and the Evaluation Network who were surveyed (Christie, 2003; Shadish & Epstein, 1987). Identifying the background of people conducting evaluation in education settings and how they practice evaluation is important to discover links or missing links between theory and practice.

Under the inquiry modes taxonomy, the current study falls into the description category. Evaluators reported their own practices of evaluation and provided a description of the way evaluation looked in K-12 settings. In addition, the people conducting evaluations provided descriptive information about themselves to describe what evaluators look like in K-12 public school districts.
The current study gathered information on evaluation practice in general, rather than focusing on evaluation practice using a specific approach. Evaluators who practice in the field of education were surveyed. Similar to Christie’s (2001) and Shadish and Epstein’s (1987) studies, evaluators were surveyed regarding their practice and experiences. The participants included evaluators who are affiliated with the American Evaluation Association and those who are not members. Many studies in the past included members of the American Evaluation Association or people who were hired to evaluate a specific program. A missing component in many studies is the inclusion of people who are not members of the American Evaluation Association and those working internally as evaluators.
Chapter Three: Method

Introduction

This chapter presents the methods used to address the research questions of this study: What is the breadth and depth of the preparation of education evaluators? What are the reported practices of education evaluators? What is the relationship between evaluator preparation and evaluation practice?

First, an overview of the research design is presented. Then, participant selection, instrument selection and design, and data collection procedures are described and explained. This information is provided in enough detail so that, if desired, the current study can be replicated. Next, the analytic procedures to be used to address the research questions are presented.

Research Design

To address the research questions and learn about the people who conduct evaluations, a descriptive study design will be used. An internet survey questionnaire containing the Theory to Practice (Christie, 2001) instrument will be used to collect information from K-12 public school evaluators. Survey research methodology gathers data from a population of interest with questions or interviews to collect data (Gall, Gall, & Borg, 2007). The research design will allow for a descriptive account of the current practices of evaluators gathered in a systematic manner.
Participant Selection

The practicing evaluators surveyed in this study were recruited through the public school districts within the state of Florida. Participants were contacted via their work email found on the school district website. The state of Florida had county wide public school districts, each responsible for conducting program evaluations for government programs. Many school districts had an in-house department or unit dedicated to program evaluation activities. Other districts, typically smaller districts, had one or two people responsible for directing evaluation activities.

In addition, district or school administrators may have also played a role in district evaluation activities. For example, a school’s Pre-K coordinator may be in charge of evaluating his or her school’s government funded Voluntary Pre-Kindergarten program (VPK). School district personnel who worked in a department or unit assigned to evaluation and district or school administrators who worked in roles which may require evaluation were invited to participate in this study.

The state of Florida had 67 school districts identified by the Florida Department of Education. Generating the list of participants to invite to participate in the study was conducted similarly to the methodology used in Hines and associates’ (2007) study. The researcher searched the 67 school district’s websites for departments or units in which evaluation was likely to be carried out. The following department names were used to collect participant names for invitation to the study: Research, Accountability, Evaluation, Assessment, Grants, Title I, No Child Left Behind, School Improvement, and Federal Programs. Individuals listed as staff in the department or units related to
evaluation were included in the invitation list with the exception of administrative assistants or secretaries.

For the 67 school district websites, 803 people were identified for invitation to the study. The number of people invited to participate in the study was not expected to be the number of people who are eligible to complete the survey. The size of each school district as well as factors like the number of title I schools, number of government grants obtained by the district, voluntary pre-kindergarten sites, among others could increase or decrease the number of individuals within a school district who conduct evaluation.

Participants must have played a role in a program evaluation for their school district during the past five years. Participants may have been the evaluator or part of an evaluative collaborative effort. Out of the 803 people invited to participate in the study, it was anticipated that approximately half of the people invited may not qualify for inclusion. This estimation was based off of the job titles listed on the school district website. For example, someone with a job title of “Testing Coordinator” may not be involved in the evaluation in which the test scores may be used as an information source. The scope of the jobs listed on school district websites varied, so anyone listed as personnel in an evaluation focused department or unit was invited to participate in the study. Based on response rates of similar studies, a response rate of 40% was expected, however not all respondents would qualify for the study. Based on the estimations of people who qualified for study inclusion, it anticipated that around 200 individuals would respond, resulting in a response rate of 25%.

At times, school districts contract external evaluators to conduct evaluations. External evaluators have not been overlooked in this study. However, the people who
work as school district employees and conduct internal evaluation are the focus of this study. The practice of internal evaluators, particularly in education settings is often missed in studies (Christie, 2009). Therefore, external evaluators who are hired by school districts in Florida are not included in the study sample. It is also important to note that some of the people who conduct internal evaluation for school districts may also conduct external evaluations for other districts or agencies.

**Instrument Selection and Design**

According to prior studies, people conducting evaluation are often unaware of the evaluation theories they follow. Evaluation is more pragmatic than theoretical in nature (Stufflebeam, 2007). To assess the extent to which evaluators use theory to guide their evaluation practice, an instrument assessing theoretical approaches even when the respondent is unaware of the theoretical approach is needed to identify the practice of K-12 school district evaluators.

A search of the literature revealed two existing instruments designed to gather information from practicing evaluators. These instruments were developed by Shadish and Epstein (1987) and Christie (2001). Shadish and Epstein (1987) self-authored a questionnaire to assess patterns in evaluation practice. Their survey questionnaire was 12-pages long and consisted of 74 items. The survey was sent to a random sample of members from the Evaluation Research Society and the Evaluation Network, which combined to form the American Evaluation Association (AEA) in 1986 but the list of members for AEA was not available at the time of the survey.

Shadish and Epstein’s (1987) survey questionnaire asked participants for background information such as demographic data, training in evaluation, job setting,
professional training and identity. Next the survey asked evaluators 74 questions about their evaluation practice by posing questions about recent evaluations conducted, their perceived purpose of evaluation, influences on their decision making in evaluation, and methods adopted to gather data. Although the authors considered the writing and work of theorists in the field, they did not ask for contributions from a panel of experts.

Christie’s (2001) instrument, titled *Theory to Practice*, was created by the contributions from eight selected eminent theorists from the evaluation field. Eight theorists described in the previous chapter served as an expert panel for instrument development. The instrument consisted of two parts. The first part of the instrument gathers background and demographic information. This part of the instrument collects data pertaining to the training, education, and work of the respondent. The second part of the instrument collects information pertaining to evaluation practice with 38 items.

The two survey questionnaires both gather data related to the practice of evaluators. However, Christie’s (2001) *Theory to Practice* instrument was designed for people who may not be familiar with evaluation specific terminology and perspectives. Shadish and Epstein’s (1987) participants were members of an evaluation professional organization. Although membership in an open enrollment organization does not imply expertise and training in the field, it does imply an active interest in evaluation as a field.

Christie’s (2001) instrument was administered to the state of California’s Health Start Program evaluators. The author knew some of the people hired to conduct evaluation for the Healthy Start Program were program directors and traditionally did not take on the role of evaluator. This survey questionnaire is more appealing for studies.
including participants from a diverse background of training and experiences in evaluation.

The survey questionnaire instrument selected for this study is the *Theory to Practice* instrument (Christie, 2001). Permission was obtained from Dr. Christina A. Christie of Claremont Graduate University for the use of her instrument in part or completely. Christie’s (2001) instrument was created to gather information on evaluation practice from people conducting evaluation and see how their practice related to evaluation theory. Christie (2001) compared the practice of evaluators from California’s Healthy Start Program to the practice of the eight theorists who contributed to the instrument development. The current study sought to describe the practice of education evaluators and explore the relationship between evaluation preparation and practice. Although the purpose of the current study and Christie’s study is different, the *Theory to Practice* instrument provided data to fully answer the research questions presented.

The instrument development is described as presented by Christina Christie (2001) and from personal conversations with Christie regarding instrument development. The *Theory to Practice* instrument was developed in two stages. First, eight well-known theorists were selected to participate as an expert panel. The eight theorists were selected because of their individual theories and theoretical orientations of evaluation. The theorists included in the instrument development were: Robert Boruch, Huey-tsyh Chen, J. Bradley Cousins, Elliot Eisner, David Fetterman, Ernest House, Michael Patton, and Daniel Stufflebeam. Robert Boruch was not available to contribute at the time the survey was created, but he authorized Richard Berk, a colleague he worked with very closely, to
answer any questions and contribute on his behalf. The theorists on the expert panel were discussed in detail in the previous chapter.

The eight theorists were sent a letter inviting them to participate in the instrument development and all eight theorists agreed to participate. The next step in the development process required each theorist to submit a minimum of one statement related to each of the three dimensions of evaluation (methods, values, uses) highlighted by Alkin and House’s (1992) taxonomy. The theorists were directed to submit a statement exhibiting a practical application corresponding to his theoretical orientation for each of the three dimensions. Each theorist was invited to provide up to six additional statements related to the three dimensions.

The instrument aimed to “assess the degree to which an activity is or is not carried out when using a particular theory” (Christie, 2001, p. 67) therefore the theorists were asked to provide statements to coincide with an eleven point Likert-like scale for assessment. Theorists were given a response range such as, “This statement is very similar to how I conduct evaluation,” to “This statement is very dissimilar to how I conduct evaluation,” to help in statement formatting. Submitted statements were revised for appropriate language and formatted to fit the Likert-like scale for the survey.

The length of the instrument was kept to a minimum due to the amount of missing data presented in similar studies surveying evaluators about practice (e.g., Shadish & Epstein, 1987). Christie (2001) categorized and aggregated the statements into domains and eliminated duplicate items. Fifteen items were removed during the review process because of duplication. The final instrument contained 38 items related to evaluation practice. A total of 16 items were related to methods (survey items 1-16), 12 items related
to values (survey items 17-28), and 10 items related to use (survey items 29-38). The beginning of the instrument consisted of five demographic items, and six items assessing the participants’ background, education, and evaluation training and experiences. See Appendix B for a paper copy of the electronic instrument.

The use of an expert panel to develop all of the survey items was intended to reduce researcher bias. As with Christie’s (2001) study, this study is one of the few empirical derived comparative frameworks of evaluation practice. The majority of the comparative frameworks have been developed by evaluation theory experts without empirical data (Christie, 2001). The eight evaluation theorists were considered eminent theorists at the time of the instrument development. The eight theorists are still known and recognized today as the field’s top experts.

**Christie’s (2001) Instrument Pilot Study and Use**

The instrument underwent a pilot study and was used in Christie’s (2001) study. The pilot study consisted of five practicing evaluators. Feedback was solicited for clarity and readability during the pilot study. Practicing evaluators from the University of California at Los Angeles assessed the instrument and were interviewed for up to 45 minutes for feedback. As a result of the pilot study, items pertaining to theoretical orientation were moved to the end of the instrument. The participants from the pilot study suggested placing the 38 items pertaining to evaluation practice first in an effort to prevent respondents from getting frustrated if they did not identify with a theoretical orientation or book to guide their practice. Christie followed the suggestions from the pilot study. The *Theory to Practice* instrument was administered and completed by 183 California Healthy Start coordinators in Christie’s (2001) study.
Reliability and Validity of Original Instrument

Gathering measures of validity and reliability is important to support the interpretation of the *Theory to Practice* administration results. Reliability is the extent to which values obtained from an instrument are stable and consistent. Prior to administering the original *Theory to Practice* instrument, the instrument was pilot tested with practicing evaluators from the University of California at Los Angeles. After the pilot test, Christie (2001) interviewed the pilot test participants for up to 45 minutes about the clarity and readability. Christie (2001) did not report any reliability estimates in her study. Pilot testing the instrument prior to use by a small sample of a similar population can increase the reliability of measures obtained from the instrument.

In the current study, reliability was measured using Cronbach’s alpha, a coefficient of reliability. This measured how well each set of items measures a single unidimensional latent variable (Crocker & Algina, 1986). The reliability coefficient expressed the internal consistency of the scores from each scale in the *Theory to Practice* instrument.

Validity is the extent to which the instrument is measuring what it is intended to measure. The instrument was developed by the contributions of eight evaluation theorists. The eight evaluation theorists served as an expert panel during the instrument development. The expert panel supports the content validity for the instrument.

Christie used classical multidimensional unfolding (CMDU) to explore underlying dimensions in the *Theory to Practice* instrument from the results collection from the eight theorists only. In the CMDU Alternating Least-square Scaling Algorithm analysis, two dimensions \( R^2 = .928 \) transpired that cut across the three instrument
domains (three instrument domains: methods, values, and use) (Christie, 2001). The first dimension was “Scope of stakeholder involvement” and the second dimension was “Method proclivity” as referred by Christie (2001). It is not surprising to find two dimensions cutting across the three instrument domains. Stakeholder involvement spans across the three instrument dimensions. Regardless of evaluators’ practice for methods, values, and use, evaluators may or may not involve stakeholders and vary the extent of involvement throughout an evaluation. Nine items were found in the scope of stakeholder involvement dimension, containing items from all three of the instrument dimensions.

The method proclivity dimension from Christie’s (2001) CMDU analysis was based on six items, four of which were part of the methods dimension from the instrument. The two items from outside of the methods dimension were items 19 and 21 from the values instrument dimension. Item 19 stated “I believe that evaluation conclusions are mixtures of facts and values,” and item 21 stated “Using my evaluation approach, stakeholders’ assumptions about a program are integrated into the evaluation process in order to ensure its relevancy and usefulness” (Christie, 2001, p. 115). Item 19 brought in a mixed-method idea by stating mixtures of facts and ideas, which may resemble quantitative and qualitative relation. In item 21, specific methodologies with pre-determined steps did not allow the evaluator to take stakeholder assumptions into account. This could make this statement methods focused rather than values focused.

The two dimensions from the CMDU accounted for 92.8% of the variance. The three instrument dimensions used in the instrument development, methods, values, and use, accounted for 95.7% of the variance. Compared to the 92.8% explanation provided
by the two dimensions, the three dimensions, *methods, values,* and *use,* were supported in Christie’s study ($R^2 = .957$).

Results from the study demonstrated the use of the *Theory to Practice* instrument to “assess a theoretical approach when the individual being questioned is unaware of the theory underlying that approach” (Christie, 2001, p. 181). Christie (2001, 2003) did not identify a need for changes to the *Theory to Practice* instrument for future use. After personal discussions with Dr. Christie, and evaluation of the results from her study (Christie, 2001), the *Theory to Practice* instrument appears appropriate for the current study proposed. However, a factor analysis was included in this study to gather additional information regarding the underlying dimensions defining the practice of evaluators.

**Changes to Christie’s (2001) Instrument**

The background information in the first part of the *Theory to Practice* instrument was adjusted to meet the needs of the current study. The first item on the survey questionnaire asked participants if she or he has played a role in a program evaluation in her or his district in the last five years. The purpose of the first item was to exclude participants who have not conducted an evaluation.

Item seven is another addition to the original instrument. Item seven asked participants to indicate the amount of training completed in 13 topic areas related to evaluation. The topic areas were generated from a review of courses and training workshops from evaluation programs, the Evaluator Institute, and pre and post conference workshops conducted at the American Evaluation Association’s annual meeting. The last change made to the *Theory to Practice* instrument was the removal of
item 19. Theoretical orientation were not the focus of the current study and not appropriate to ask the school district evaluators.

**Focus Group for Instrument**

A focus group was held to ensure the survey questionnaire was appropriate for evaluators working in school districts. The purpose of the focus group was to have people who were familiar with the work of school district evaluators look at the survey questionnaire instrument to review the items’ content and language. Invitation emails were sent to twenty individuals who worked in school districts in the role of the director of a unit in or closely related to evaluation (some units were titled accountability, research, assessment, etc.) to participate in the focus group. Eight people responded to the email but each person expressed concern about anonymity among other participants during the focus group. For this reason, the focus group was held over the phone without the use of names or work locations. In some cases the voice of the respondent may have allowed other participants and the researcher to identify the gender of the participant.

An email was sent to the eight individuals who expressed interest in participating in the focus group. The email provided information regarding the telephone conference call toll free number and date and time of the focus group. Five people called in for the scheduled focus group. Each focus group participants had a minimum of seven years of experience conducting evaluations in K-12 public schools and each of them worked for a Florida school district at some point throughout his/her career. Each participant was given an alphabet letter to use as their identification for record keeping purposes. Letters A-E were used as the identifiers.
Focus Group Results

A focus group was conducted using the current instrument with five people who worked in an evaluation capacity within a school district. Feedback was solicited for relevance and clarity of the survey items. The five focus group participants reviewed and completed the survey, and then participated in a 30 minute focus group. The focus group was intended to ensure the items were pertinent to the tasks and training school districts require from their evaluation related staff. Particularly, the focus group addressed the following three categories: 1) overall impression, 2) item clarity/content, and 3) areas of improvement. The results of the focus grouped informed the revision of any survey items identified as problematic.

Two items were added to the end of the survey to allow respondents to provide additional information. Specifically, one item asks respondents to describe the extent to which her/his practice was influenced by politics. Focus group respondents thought political influences may play a role in the way some people carry out evaluations. The last item on the survey allows respondents to provide any information s/he thought was important for the researcher to know about her/his evaluation practice. The item allows respondents to share information they were not able to communicate from the selected-response items.

Data Collection Procedures

The revised instrument was entered into an electronic survey system operated by Surveymonkey®. The participants recruited for this study work for school districts. Based on this characteristic, each individual will have access to computers and the internet. The electronic survey was sent out to the participants. A 10-day period was
designated for participants to respond. In order to secure a sufficient number of responses, a reminder email was sent out two days prior to the deadline, and an additional five days was provided to complete the electronic survey.

To increase response rates, a pre-notice email was sent out inviting participants to take part in the study (Dillman, 2007) (see Appendix A). Proper settings were imposed on the data gathering functions within the survey in SurveyMonkey® to prevent the researcher from obtaining any identifying information from participants such as the computer network or computer address from participants. Institutional review board approval was sought out and met prior to the collection of data.

In addition to the survey questionnaire, two informal semi-structured interviews were conducted to provide additional information on evaluation context. The interviewers were a result of email communication after completing the survey questionnaire online. Three people communicated via email after completing the survey questionnaire, and only two agreed to participate in an interview. A list of questions for the interview was generated by the researcher prior to speaking with the participants on the phone. The interview protocol can be found in Appendix C. The interviews were recorded with an audio recorder and transcribed by the researcher. The transcripts were sent to the interviewees for verification of responses. Both interviewees were satisfied with the transcript and did not indicate any changes should be made.

**Pilot Study**

A pilot study of the current instrument was conducted to see how the items performed and attempt to identify any problems with the administration of the survey. The instrument was sent via email to 145 people who were in the original estimated
sample on January 11th, 2010. Of the 145 who received the survey link, 21 people responded. The first item on the survey asks if the respondent had conducted a program evaluation in the past five years. People who responded “Yes” were taken to the rest of the instrument, people who responded “No” were sent to the last page of the survey and thanked for responding. Fourteen people responded “Yes” and completed the remainder of the instrument.

The results of the 14 pilot study results indicated the survey items were working satisfactorily. The Theory to Practice part of the instrument contained responses with variability which indicated the start of a rich body of data. Respondents did not provide any feedback on problems with administration. Reliability was estimated for the methods, values, and use dimensions for the pilot data and were found to be satisfactory (method $\alpha = .78$, values $\alpha = .66$, and use $\alpha = .75$). Based on the results obtained from the pilot study, all items were left in the instrument and the survey link was sent out to the remainder of the people identified as the sample on February 24th, 2010. A follow-up email was sent on March 3rd, 2010 as a friendly reminder to complete the survey. Some potential participants responded to the invitation email indicating they were interested in the results of the study, and would like to participate, but were not permitted to complete surveys or interviews regarding their job activities. The invitation email indicated that names and/or districts would not be identified and only aggregated data would be reported, however those who responded by email indicated they were advised to reframe form provided any information regarding the nature of their job.
Analytic Procedures

This section describes the basic techniques and procedures used in analyzing the data gathered by the *Theory to Practice* instrument. A description of the technique used and inferences that can be made from each analysis are provided. The relationship between the data, procedures, and research questions are addressed.

A variety of analyses were conducted to answer the research questions posed. Descriptive statistics were computed to examine the breadth and depth of the preparation of education evaluators and the reported practices of education evaluators (research questions 1 & 2). For the third research question, the relationship between evaluator preparation and evaluator practice was answered with multivariate analysis of the variance (MANOVA).

**Research Question 1: What is the breadth and depth of the preparation of school district evaluators?** Descriptive statistics including frequencies are presented on the background information. Survey items five, six, seven, and eight describe the training of evaluators. Frequencies and charts are presented for the degree obtained and the subject area aligning with the highest degree, for items five and six, respectively. Item seven describes the training and extent of training related to evaluation with frequencies.

**Research Question 2: What are the reported practices of school district evaluators?** Descriptive statistics including frequencies, measures of central tendency and variability are presented on the 39 items from the *Theory to Practice Instrument*. Patterns of practice among evaluators, means, standard deviations, and ranges for each of the 39 items from the instrument is presented.
A principal component analysis was conducted in SPSS 17.0 to examine the multiple dimensions of the *Theory to Practice* instrument. The principal component analysis allowed the researcher to summarize the practice of evaluators with different factors. The instrument was developed with the three dimensions, methods, values, and use. However, Christie’s (2001) study found two underlying dimensions, “scope of the stakeholder” and “method proclivity” using Multiple Scaling Analysis (MDS). Due to the discrepancies from the instrument development and Christie’s (2001) analysis, a principal component analysis with a varimax rotation was conducted within the method, values, and use dimensions.

The principal component method was used, and a scree plot of eigenvalues and interpretation of solution were used to identify the number of factors to retain (Loehlin, 1998). Due to the small sample size, the principal component analysis was used over other methods for stability of the solution. Factor patterns were determined by factor loadings >.3, using the highest factor loading per item, and the interpretation of solution. The interpretation of solution sought to make sense of the groups indentified. Descriptive statistics are provided to describe the resulting factors from the factor analysis.

**Research Question 3: What is the relationship between evaluator preparation and evaluation practice?** Multivariate analysis of the variance (MANOVA) was used to explore differences between groups of evaluators and the dimensions underlying their practices. Based on the results of the primary reliability estimates methods, values, and use, were planned to be used for the MANOVA analyses. MANOVA is a statistical technique allowing researchers to determine if groups differ on more than one dependent variable (Stevens, 2002). MANOVA differs from the t test and analysis of the variance
(ANOVA) by allowing the use of multiple dependent variables. This also removes the risk of drastically increasing the possibility of making a Type I error by conducting too many ANOVA’s independently.

To answer the research question, three one-way MANOVAs were conducted. The first MANOVA used item five “highest degree obtained” as the independent variable, and the dimensions of evaluation practice were the dependent variables.

The second MANOVA used item six “subject area or discipline of the highest degree obtained” as the independent variable, and the dimensions of evaluation practice were the dependent variables. After data were collected, the researcher collapsed and removed some categories of highest degree into three categories. The first category, evaluation/research methods, included: Evaluation/Research Methods, Psychology, and Advanced Quantitative Methods. The second category was education, and the third category was school administration. Collapsing occurred because the number of responses was too small to conduct a MANOVA. Additional rationale for the categories is discussed in the results section.

A third MANOVA was conducted by combining the first and second topics from item seven, “the amount of training in Evaluation Theory” and “the amount of training in Program Evaluation” for the independent variable, and the methods, values, and use dimensions as the dependent variables. Total scores were calculated for extent of training in evaluation theory and program evaluation to conduct the analyses. A description of the procedures used to calculate total scores and grouping for the one-way MANOVA are provided in the results section.
Conclusion

The statistical methods employed in this study allowed the three research questions and sub research question to be answered. The main goals of this study were to: (1) to identify the training of education evaluators, (2) to examine the practice of education evaluators, and (3) to examine the relationship between training and practice. The analyses allowed the researcher to present a descriptive framework depicting those who conduct education evaluations and a picture of their perceived evaluation practice.
Chapter Four: Results

Overview

This chapter presents the results of the current study. The results from the analyses describe the demographic, academic, and professional background of respondents. Reliability estimates and a principle component analysis were used to describe the portion of the instrument related to evaluation practice. The research questions were answered with descriptive statistics, reliability estimates, three one-way MANOVAs, and interviews. The interviews were conducted to provide additional information on evaluation context within school districts.

Demographic and background information

Individuals who conduct evaluation in the public schools of Florida were selected for this study because little was known about people who conduct evaluation internally for school districts. The results provide noteworthy information on the people and practice of evaluation from a large state in the United States. The survey was emailed to 803 people and 154 people responded resulting in a response rate of 19%. A reminder email was sent out a week later to try to increase the response rate. The reminder email resulted in very few additional responses, therefore additional reminder emails were not sent after the first one. The amount of missing data for the survey was minimal (2.99%). Respondents typically answered all of the items used in the analysis or stopped taking the survey midway through. Only respondents with complete data were included in the analyses.
The first item qualifies the respondent for the study by asking whether or not s/he has conducted a program evaluation in an education setting in the last five years. Out of the 154 respondents, 130 answered “yes” and were eligible to complete the rest of the survey, while 24 respondents answered “no” and were brought to the last page of the survey. The last page of the survey thanked participants for their participation. The majority of respondents were female with 63% and males comprised 37% of respondents.

The race/ethnicity of respondents consisted of 88% White, 8% Asian/Pacific Islander, 6% Hispanic, and 2% Black. No respondents selected Other or Mixed Race. Figure 3 below displays the race/ethnicity breakdown.

![Race/ethnicity of Respondents](image)

**Figure 3. Frequency of race/ethnicity of respondents**

The majority of respondents (57%) were 50 years of age or older. The 45 to 49 years of age range and 35-39 age range each contained 16% of respondents. Four percent of respondents reported their age was between 40-44 years, 6% reported their age was between 30-34 years, and 2% reported their age was between 25-29 years. No respondents selected less than 25 years old.
Research Question One

The breadth and depth of preparation of school district evaluators is described via highest degree obtained and alignment of highest degree into a subject area. The most frequent degree obtained with the greatest number of respondents was a master’s degree with 60, the second greatest number of respondents held a doctorate with 48. The remainder of respondents indicated 6 held a specialist degree (Ed.S.), 13 a bachelors, and 3 a high school diploma. Figure 4 provides a visual representation of highest degree obtained.

![Figure 4. Frequency of highest degree earned by respondents](image)

Respondents were asked which area their highest degree most closely aligns with from a list of subject areas. The greatest number of respondents identified Education as their aligned field for their highest earned degree. Evaluation/Research Methods and School Administration both had the same number of respondents align their highest earned degree, and ten respondents selected Other. In the Other category, four of
respondents listed fields within education such as reading education and education technology, two respondents were computer science, and four were social sciences/social work. Figure 5 displays highest degree alignment of respondents.

Figure 5. Area of highest degree earned by respondents

Respondents were asked the amount of training completed in twelve areas related to evaluation: evaluation theory, program evaluation, personnel evaluation, quantitative methods, logic models, cost-benefit/cost-effectiveness, needs assessment, measurement, survey research, qualitative methods, focus groups, and interviewing. Respondents were instructed to select all that apply. Respondents selected more than one category for each topic when applicable. Table 3 provides details on the amount of training completed by respondents in the evaluation topic areas.

Measurement was the most frequent topic area (68 responses) for the one or more courses category. Quantitative methods came in second with 66 responses for one or
more courses, while Evaluation theory, Program evaluation, and Qualitative methods had 44, 40, and 40 responses, respectively. One or more courses in Survey research ranked sixth with 32 responses, and the remainder of the topics ranged from eight to 24 responses, with Focus groups having the least responses. Figure 6 provides a visual display of the one or more courses training category.
Table 3

Amount of Training in Evaluation Areas (n = 130)

<table>
<thead>
<tr>
<th></th>
<th>One or more full courses</th>
<th>Substantial part of a full course</th>
<th>One or more workshops</th>
<th>Substantial part of a workshop</th>
<th>Small part of a course or workshop</th>
<th>Staff Development</th>
<th>Independent Learning</th>
<th>Min. exp or no formal training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval Theory</td>
<td>44</td>
<td>26</td>
<td>22</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td>Prog Eval</td>
<td>40</td>
<td>20</td>
<td>28</td>
<td>14</td>
<td>4</td>
<td>14</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>Personnel Eval</td>
<td>16</td>
<td>26</td>
<td>40</td>
<td>12</td>
<td>6</td>
<td>22</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Quant Methods</td>
<td>66</td>
<td>14</td>
<td>16</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Logic Models</td>
<td>10</td>
<td>14</td>
<td>14</td>
<td>4</td>
<td>14</td>
<td>4</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Cost-Benefit/Effec.</td>
<td>20</td>
<td>14</td>
<td>22</td>
<td>10</td>
<td>14</td>
<td>8</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Needs Assess.</td>
<td>24</td>
<td>20</td>
<td>38</td>
<td>10</td>
<td>4</td>
<td>18</td>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>Measurement</td>
<td>68</td>
<td>14</td>
<td>24</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>Survey Research</td>
<td>32</td>
<td>20</td>
<td>18</td>
<td>2</td>
<td>16</td>
<td>10</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>Qual Methods</td>
<td>40</td>
<td>22</td>
<td>18</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Focus Groups</td>
<td>8</td>
<td>12</td>
<td>28</td>
<td>6</td>
<td>10</td>
<td>24</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td>Interviewing</td>
<td>14</td>
<td>28</td>
<td>34</td>
<td>6</td>
<td>10</td>
<td>16</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>382</strong></td>
<td><strong>230</strong></td>
<td><strong>302</strong></td>
<td><strong>86</strong></td>
<td><strong>98</strong></td>
<td><strong>160</strong></td>
<td><strong>250</strong></td>
<td><strong>142</strong></td>
</tr>
</tbody>
</table>

Note: the values provided in the table are n’s not percentages, individuals could check more than one response.
Figure 6. Frequency of responses in which respondents completed one or more full courses.

Program evaluation had 40 respondents indicating one or more full courses in the topic, 20 respondents choosing substantial part of a course, and 28 choosing one or more workshops. Thirty respondents indicated their training in program evaluation was from Independent Learning. See Figure 7 for a visual display of program evaluation.
Figure 7. Amount of training in Program Evaluation

Training in Evaluation theory consisted of 44 responses for One or more full courses, 25 responses for Substantial part of a full course, and 22 responses for each One or more workshops and Independent learning. Sixteen respondents indicated they had minimal exposure or no formal training in evaluation theory. Figure 8 provides a visual display of all training categories for Evaluation theory.
Figure 8. Frequency of respondents’ training in Evaluation Theory

The One or more workshops training categories had the most responses for Personnel evaluation with a frequency of 40, Needs assessment came in second with a frequency of 38, and Interviewing came in third with 34 responses. The Substantial part of a workshop and Small part of a course or workshop had the least amount of responses compared to the other training categories.

Based on the responses collected, in regards to the evaluation topics, staff development consisted mostly of training in Focus Groups ($n = 24$), Personnel Evaluation ($n = 22$), and Needs Assessment ($n = 18$). Interviews, qualitative methods, and program evaluation had frequencies of 16, 14, and 14, respectively. Logic models had the least responses for professional development training. Figure 9 provides additional details on staff development training.
The majority of respondents did not complete an internship or practicum in evaluation. Two respondents reported completing an evaluation internship or practicum during their undergraduate studies, while 32 reported completing an internship or practicum in evaluation during graduate studies. The majority of respondents, 86% conduct evaluation as part of their job responsibilities. Respondents were asked “What percent of your job do you spend conducting evaluations?” Respondents were provided with a drop down menu consisting of percentages ranging from five percent to 100 percent, in increments increasing by five percent (e.g., 5%, 10%, 15%, etc.). The most frequent responses were 5% and 20%, with 20% of respondents selecting each of those categories. Forty-one percent of respondents indicated they conduct evaluations between 5-25% of their job, 33% reported evaluations consist of 25-50% of their job responsibilities, 22% for 50-75% of their job responsibilities, and 17% of respondents spend 75-100% of their job conducting evaluations.
In addition to conducting evaluations, the majority of respondents, 92%, identified themselves as program administrators, program specialists, or program coordinators. Other work included school administrator (18%), university faculty/staff (8%), teacher (6%), and social worker (1%). The majority of respondents have been conducting evaluations for over seven years, with 42% at 10 or more years, 25% seven to nine years, and 19% four to six years. Newer evaluators included 10% of respondents reporting one to three years of experience and four percent reported less than one year of experience.

Respondents were asked to rate their current evaluation knowledge and skills as Excellent, Good, Average, or Minimal. Figure 10 provides a visual display of respondents’ perception of evaluation knowledge and skills. One quarter of respondents rated their knowledge and skills as Excellent, while 52% rated their knowledge and skills as Good. The remainder of responses included 21% rating as Average, and 2% as minimal. Ninety-four percent of respondents indicated they typically conduct evaluations as an internal evaluator, while 17% reported typically working as an external evaluator. Respondents were able to select both the internal and external evaluator options.
Figure 10. Respondents’ perception of evaluation knowledge and skills

Respondents were provided with a list of ten program types, and were asked to indicate whether or not they typically conduct program evaluation on the programs listed. The ten programs were provided based on the focus group held during the survey development period. Participants were directed to check “yes” or “no” to respond to each type of program. The majority of respondents typically evaluate curricular/curriculum programs, with 72% of respondents selected “yes” for that item. Early childhood, special education, and staff development programs each had 45% indicated they typically evaluate those programs, and 36 people selected “Other”. Responses from the “Other” category included school-wide programs, quality assurance, summative assessment K-12, curricular initiatives, district level administrators, grant requirements for federal grants, magnet programs, career and vocational technical schools, adult education programs, accreditation, reading, Title 1 and after school programs, and AVID a tutoring program. See Figure 11 for additional details.
Supplemental education services (SES), part of Title 1 programs was selected by 32 people, ELL/ESOL had 26 responses, and Gifted programs had 22 responses. Substance abuse prevention had the least number of responses with a frequency of six, dropout prevention, and magnet programs had 16 and 20 responses, respectively.

Item 16 of the survey asked respondents to select a program evaluation s/he conducted within the last five years which represented the way s/he typically carries out evaluation. Participants were instructed to base the remainder of the items on the survey on the way s/he carried out the program evaluation identified in item 16. Figure 12 displays the type of program evaluation selected by respondents.

Curricular/curriculum was the most popular choice for respondents with 23. The second most popular category was Other. Many responses in the other category were coded into the categories provided. For example, one person stated “STEM education program” which was coded as a Curricular/curriculum program. Another person provided
the response “reading program” which was also put into the Curricular/curriculum category. The other category had one respondent identify a homeless education program which did not seem to fit into any of the standing categories. The majority of the respondents who selected “Other” did not provide detailed information on the type of program so the responses remain as “Other”. Dropout Prevention was listed as an option but was not selected by any of the respondents for this item. Supplemental education services was changed to “Federal Programs” because supplemental education services are part of Title 1 and included as a federal program. The category federal programs allowed for more responses listed as “Other” to be included as a group because of the similarities in the programs. Accreditation was also grouped into the same category as federal programs because of similarities in requirements and reporting procedures. Figure 12 provides a visual display of the program selected by each respondent.
Respondents were instructed to base their answers for the rest of the survey on the evaluation selected in this item.

**Summary of research question one.** The background information presented provides a description of the breadth and depth of preparation of school district evaluators. The majority of respondents held a master’s degree as their highest earned degree and education was overwhelmingly the most common area of highest degrees held for all degree types. Measurement was identified as the areas in which the greatest number of respondents received one or more full courses of training and quantitative methods had the second greatest frequency in the one or more full courses category. Twenty respondents reported minimal or no training in evaluation theory and program evaluation, and 52 respondents indicated their training in evaluation theory and program evaluation was from independent learning. Staff development activities related to
evaluation consisted mostly of training in focus groups, personnel evaluation, and needs assessments.

The vast majority of respondents, 92%, identified their primary role in their school district was a program administrator, program specialist, or program coordinator. Overall respondents perceived their evaluation knowledge to be “good” (52%), and most have been conducting evaluations for seven or more years (68%). Typical evaluations consisted mostly of curricular or curriculum programs as well as early childhood, magnet, special education, and staff development programs.

**Research Question Two**

Research question two was answered using quantitative and qualitative data. The main part of the survey questionnaire including both selected-response and open-ended items were used as well as two interviews.

**Quantitative Analysis.** The main part of the instrument contained items on evaluation practice in which respondents rated each item on a scale of zero to ten. Zero represented Very Dissimilar or Strongly Disagree, and ten represented Very Similar or Strongly Agree. These items were used to address the second research question: what are the reported practices of school district evaluators? These items were also used in conjunction with the items on background information to answer research question three: What is the relationship between evaluator preparation and evaluation practice? The open-ended items on the survey questionnaire were also used to describe the reported practice of school district evaluators. Table 4 displays a description of each item, as well as the mean, standard deviation, minimum and maximum value.
Table 4

*Means, Standard Deviations, Minimum, and Maximum Values for Responses from the Methods, Values, and Use Item.*

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>S.D.</th>
<th>Min-Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>5.55</td>
<td>3.25</td>
<td>0-10</td>
</tr>
<tr>
<td>m2</td>
<td>6.93</td>
<td>2.83</td>
<td>0-10</td>
</tr>
<tr>
<td>m3</td>
<td>5.08</td>
<td>3.08</td>
<td>0-10</td>
</tr>
<tr>
<td>m4</td>
<td>7.21</td>
<td>2.67</td>
<td>0-10</td>
</tr>
<tr>
<td>m5</td>
<td>6.77</td>
<td>3.28</td>
<td>0-10</td>
</tr>
<tr>
<td>m6</td>
<td>6.92</td>
<td>3.33</td>
<td>0-10</td>
</tr>
<tr>
<td>m7</td>
<td>6.42</td>
<td>2.98</td>
<td>0-10</td>
</tr>
<tr>
<td>m8</td>
<td>8.31</td>
<td>2.32</td>
<td>0-10</td>
</tr>
<tr>
<td>m9</td>
<td>7.11</td>
<td>2.70</td>
<td>0-10</td>
</tr>
<tr>
<td>m10</td>
<td>5.97</td>
<td>3.21</td>
<td>0-10</td>
</tr>
<tr>
<td>m11</td>
<td>6.31</td>
<td>2.97</td>
<td>0-10</td>
</tr>
<tr>
<td>m12</td>
<td>6.94</td>
<td>2.78</td>
<td>0-10</td>
</tr>
<tr>
<td>m13</td>
<td>6.69</td>
<td>2.96</td>
<td>0-10</td>
</tr>
<tr>
<td>m14</td>
<td>7.06</td>
<td>3.13</td>
<td>0-10</td>
</tr>
<tr>
<td>m15</td>
<td>7.63</td>
<td>2.70</td>
<td>0-10</td>
</tr>
</tbody>
</table>
In order for evaluations to be most effective, stakeholders must be included, but don’t necessarily have to participate in the evaluation process.

Included working with primary users to make informed judgments about their program using the evaluation data.

Allowed me to use my own values to select what I believe to be significant.

Objectively assessed the evaluation’s quality against validated technical standards.

Clear contracts of work requirements subjected to review by people both internal and external to the organization were used.

Stakeholders’ assumptions about a program were integrated into the evaluation process.

The primary users judged the appropriateness and quality of the evaluation’s methodology.

It was most important to ensure the evaluation was scientifically sound.

The primary users helped interpret the meaning of the evaluation data.

Including and reflecting the diverse perspectives of relevant stakeholders was important (not just what was required by the funding source).

Evaluation conclusions are mixtures of facts and values.

Evaluator’s interpretation of the findings can be unbiased.

An evaluator is an applied scientist; not an advocate, counselor or policy advisor.

Identified, engaged, and served stakeholders at all levels of the program.

Issued separate reports to serve the different needs of the various audiences.
The following stems were used to begin the items below: “When conducting evaluation, my evaluation approach” and “When conducting my evaluation”. The item with the lowest mean was Method three (m3) – “When conducting evaluation, my evaluation approach focused predominately on the observation of events (qualitative)” with a mean of 5.08 and a standard deviation of 3.08. Item Use six (u6) – “When conducting my evaluation, the primary focus of the evaluation design was to improve program performance.”
program performance” had the highest mean with 8.72 and the lowest standard deviation out of all of the items, 1.88. The item with the highest standard deviation was Use two (u2) – “When conducting evaluation my evaluation approach issued separate reports to serve the different needs of various audiences” with a standard deviation of 3.36. The values for almost all responses on the method, values, and use items ranged from zero to ten. Items m6 and v7 ranged from one to ten, and items u6 and u10 ranged from three to ten. The remainder of the items each had responses ranging from zero to ten.

**Item grouping.** Items were grouped together based on the three domains, method, values, and use. The internal consistencies of the original scales were estimated using Cronbach’s alpha to estimate the reliability of the methods, values, and use scales. Reliability estimates were also calculated for the two dimensions Christie (2001) named, “scope of the stakeholder involvement” and “method proclivity”, when analyzing results from California Healthy Start evaluators using the original instrument.

The *method* items consisted of 16 items (m1-m16) and had an estimated reliability of $\alpha = .79$. This is an acceptable value of Cronbach’s alpha and was considered to be good. The reliability estimate for *values* was $\alpha = .70$, and *use* had an estimated reliability of $\alpha = .80$. The reliability estimates for “Scope of Stakeholder Involvement” and “Method Proclivity” were $\alpha = .66$ and $\alpha = .10$, respectively. The original domains, *method, values, and use* each had acceptable reliability estimates ($\alpha$ of .70 or higher). The reliability estimates for “Method Proclivity” and “Scope of Stakeholder Involvement” were lower values which were below the acceptable value. The “Scope of Stakeholder Involvement” dimension did have an alpha that was near the acceptable .70 level, however the Method dimension had a poor reliability estimate ($\alpha = .10$) making the
original three dimensions more appropriate to use to describe the practice of school district evaluators in regards to method. The Method, Values, and Use dimensions each had reliability estimates at acceptable values.

Each dimension contained a variety of items which represented different aspects of the dimension. Responses to items could not be interpreted as correct or incorrect in the dimensions. Due to these variations in the nature of the items used in the instrument, a principal components analysis (PCA) with varimax rotation was conducted within each dimension (method, values, and use) to reveal underlying factors. The scree plot of eigenvalues and interpretation of solution were used to identify the number of factors to retain in the principal components analysis. The scree plots for each principal component analysis can be found in Appendix C. Based on the sample size, interpretability of the solution was sought to make sense of the item groups indentified in order to describe different practices of evaluation. Factor patterns were determined by factor loadings >.3, using the highest factor loading per item, and the interpretation of solution.

Results of the principal component analysis for the method dimension indicated a possible interpretation of two to five factors (see Appendix C for scree plots). Two, three, four, and five factor solutions were investigated. Based on the interpretation of the solution, a two factor solution was selected to describe the method practice of evaluation. The first factor contained seven items m1, m3, m4, m5, m6, m7, and m12, and was termed “Holistic Approach”. The term holistic was selected because of the encouragement to include people at all levels of the organization (m5) and using the necessary methods to complete the evaluation (m4) particularly the use of qualitative methods (m1 and m3). The items represent the idea of included a variety of people in the
process and changing the course of action to meet the needs of the evaluation. The second factor contained eight items, m2, m8, m10, m11, m13, m14, m15, and m16, and was referred to as “Mixed Method Decision Making.” The term Mixed Method Decision Making was selected based on the use of qualitative and quantitative methods (m15) and designing evaluation questions with primary users for decision making purposes. In this factor, the items indicate evaluators work with primary users to develop the evaluation questions and may include stakeholders in the evaluation at some parts but not necessarily the entire process. The factors and the items included are displayed in Table 5. Item m9 aligned with factor one according to the principal component analysis but the item did not correspond well with the other items. Therefore, item m9 was not included as part of the Holistic factor and was not included in any further analyses. The two factors explained 47.6% of the variance for the set of method items. The Holistic factor had a Cronbach’s alpha of .82 and Mixed Method Decision Making had a Cronbach’s alpha of .72.

Table 5

Structure Coefficients Representing Methods in Evaluation Practice

<table>
<thead>
<tr>
<th>Structure Coefficients</th>
<th>Factor 1: Holistic</th>
<th>Factor 2: Mixed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>m1</strong> Used qualitative (words) methods as my primary approach.</td>
<td>.77*</td>
<td>.01</td>
</tr>
<tr>
<td><strong>m2</strong> Used quantitative (numbers) methods as my primary approach.</td>
<td>.17</td>
<td>-.79*</td>
</tr>
<tr>
<td><strong>m3</strong> Focused predominately on the observation of events (qualitative).</td>
<td>.55*</td>
<td>.25</td>
</tr>
<tr>
<td><strong>m4</strong> Used the methods necessary to conduct a useful, feasible, proper and accurate evaluation.</td>
<td>.55*</td>
<td>-.04</td>
</tr>
<tr>
<td><strong>m5</strong> Encouraged people at all levels of the organization to participate in the evaluation process.</td>
<td>.64*</td>
<td>.52</td>
</tr>
<tr>
<td><strong>m6</strong> Involved helping the program staff and clients establish their goals and document evidence of working words their goals.</td>
<td>.55*</td>
<td>.43</td>
</tr>
</tbody>
</table>
m7 Helped the program staff and clients develop a plan for the future. **.74** * .00
m8 The evaluation questions were designed to yield information for making decisions about the program. **.21** * .60
m9 Research methods were selected based on the program’s conceptual framework, model or theory. **.56** * .43
m10 The main evaluation questions were answered using scientifically tested instruments. **-.03** * -.39
m11 The primary users helped conceptualize and determine the evaluation questions. **.06** * .32
m12 Adjustments and changes were made when parts of the current evaluation plan were not working **.59** * .19
m13 Stakeholders participated in conducting the evaluation. **.51** * .57
m14 I observed what was transpiring in the program and then interpreted and judged its significance. **.7** * .73
m15 I combined qualitative and quantitative methods to address each evaluation question. **.60** * .63
m16 In order for evaluations to be most effective stakeholders must be included, but don’t necessarily have to participate in the evaluation process. **-.46** * .47

Note: Structure coefficients in bold with an * were included in the factor; n = 130

The values dimension indicated a possible interpretation of two to four factors. Three solutions were considered, two, three, and four factors. Based on interpretation of the solution, the two factor solution was selected to describe the values practice of evaluation. The first factor contained six items v3, v4, v6, v7, v8, and v9, and was named “Procedures Valued”. The second factor contained four items, v1, v2, v5, and v10, and was referred to as “People Valued.” The factors and the items included are displayed in Table 6. Items v11 and v12 had loadings > .3 however the loadings were low and did not align well with the other items in the factors. Items v11 and v12 were not included in either of the factors and were not included in any further analyses. The two factors explained 38.9% of the variance in the items. The Procedures Valued factor had a Cronbach’s alpha of .78 and People Valued had a Cronbach’s alpha of .63. Ideally an α > .70 is desired for reliability estimates, however, the .63 for the People Valued factor was considered acceptable for this study.
Table 6

Structure Coefficients Representing Values in Evaluation Practice

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1: Procedures Valued</th>
<th>Factor 2: People Valued</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>.06</td>
<td>.69*</td>
</tr>
<tr>
<td>v2</td>
<td>.23</td>
<td>.69*</td>
</tr>
<tr>
<td>v3</td>
<td>.72*</td>
<td>-.38</td>
</tr>
<tr>
<td>v4</td>
<td>.56*</td>
<td>.28</td>
</tr>
<tr>
<td>v5</td>
<td>.12</td>
<td>.58*</td>
</tr>
<tr>
<td>v6</td>
<td>.64*</td>
<td>.21</td>
</tr>
<tr>
<td>v7</td>
<td>.74*</td>
<td>-.33</td>
</tr>
<tr>
<td>v8</td>
<td>.76*</td>
<td>.05</td>
</tr>
<tr>
<td>v9</td>
<td>.56*</td>
<td>.21</td>
</tr>
<tr>
<td>v10</td>
<td>-.19</td>
<td>.59*</td>
</tr>
<tr>
<td>v11</td>
<td>.23</td>
<td>-.31</td>
</tr>
<tr>
<td>v12</td>
<td>.43</td>
<td>-.10</td>
</tr>
</tbody>
</table>

Note: Structure coefficients in bold with an * were included in the factor; n = 130

The uses dimension indicated a possible interpretation of two to four factors.

Three solutions were considered, two, three, and four factors. Based on interpretation of the solution, the two factor solution was selected to describe the values practice of evaluation. The first factor contained five items u1, u3, u4, u5, and u7, and was named “Users Engaged/Embodied”. The second factor contained four items, u2, u8, u9, and u10, and was referred to as “Evaluator as Mediator.” The factors and the items included are displayed in Table 7. Items u11 did not load on either factor and therefore was not included in either of the factors and were not included in any further analyses. The two
factors explained 57% of the variance in the items. The Users Engaged/Embodied factor had a Cronbach’s alpha of .87 and Evaluator as Mediator factor had a Cronbach’s alpha of .77.

Table 7

Structure Coefficients Representing Uses in Evaluation Practice

<table>
<thead>
<tr>
<th></th>
<th>Structure Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Factor 1: Users Engaged/Embodied</td>
</tr>
<tr>
<td>u1</td>
<td>.80*</td>
</tr>
<tr>
<td>u2</td>
<td>.49</td>
</tr>
<tr>
<td>u3</td>
<td>.75*</td>
</tr>
<tr>
<td>u4</td>
<td>.89*</td>
</tr>
<tr>
<td>u5</td>
<td>.84*</td>
</tr>
<tr>
<td>u6</td>
<td>.03</td>
</tr>
<tr>
<td>u7</td>
<td>.72*</td>
</tr>
<tr>
<td>u8</td>
<td>-.05</td>
</tr>
<tr>
<td>u9</td>
<td>.08</td>
</tr>
<tr>
<td>u10</td>
<td>.13</td>
</tr>
<tr>
<td>u11</td>
<td>.08</td>
</tr>
</tbody>
</table>

Note: Structure coefficients in bold with an * were included in the factor; n = 130

Descriptive statistics for the six dimensions of evaluation practice can be found in Table 8. People values dimension had the lowest mean with 27.61 and mixed method decision making had the highest mean with 52.24. In addition, people valued contained the lowest standard deviation at 7.55 and the holistic dimension had the greatest standard deviation at 15.35.
Table 8

*Descriptive Statistics for Evaluation Practice Dimensions*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>130</td>
<td>44.18</td>
<td>15.35</td>
<td>0</td>
<td>70</td>
<td>-.85</td>
<td>.29</td>
</tr>
<tr>
<td>Mixed Method Decision Making</td>
<td>130</td>
<td>52.24</td>
<td>12.84</td>
<td>22</td>
<td>76</td>
<td>-.33</td>
<td>.30</td>
</tr>
<tr>
<td>Procedures Valued</td>
<td>130</td>
<td>41.25</td>
<td>11.85</td>
<td>9</td>
<td>60</td>
<td>-.68</td>
<td>.30</td>
</tr>
<tr>
<td>People Valued</td>
<td>130</td>
<td>27.61</td>
<td>7.55</td>
<td>0</td>
<td>40</td>
<td>-1.55</td>
<td>.30</td>
</tr>
<tr>
<td>Users Engaged/Embodied</td>
<td>130</td>
<td>35.47</td>
<td>10.54</td>
<td>0</td>
<td>50</td>
<td>-1.41</td>
<td>.29</td>
</tr>
<tr>
<td>Evaluator as Mediator</td>
<td>130</td>
<td>28.30</td>
<td>8.57</td>
<td>10</td>
<td>40</td>
<td>-.63</td>
<td>.30</td>
</tr>
</tbody>
</table>

*Note:* the means represent the mean of the sum for each dimension

**Qualitative analysis.** Results of the quantitative analyses yielded differences in evaluation practice based on background and training. In order to gain a better understanding of the way evaluation in carried out in school districts qualitative data were collected on the survey questionnaire via open ended items and through semi-structured interviews over the phone. The three main research questions were aligned with selected response items on the survey questionnaire, and in addition to the selected response items, the open ended items and interviews provided qualitative data to describe the practice of evaluation, which was the main goal of research question two.

Communication with respondents via phone interviews was initiated by respondents who sent an email to the email address provided at the end of the survey. Three emails were received from respondents who chose to send an email asking if s/he could call the researcher to provide more information. Two people called the researcher and participated in an informal semi-structured interview. It is important to note the state of Florida has a very broad public record law. Anything transmitted via email from a school district or government agency email account becomes public record and can be requested by the public. As a result of the public record law, many workers will not
disclose any sensitive information in email. Asking an individual to meet in person or over the phone may be a popular practice in states with similar laws.

**Open-ended items.** Evaluation context was not captured as part of the selected-response survey items. Context can play an important role in the way an evaluation is carried out. Respondents were asked to describe the context and purpose of the evaluation s/he selected for the survey via an open-ended item on the survey. Thirty-four people provided information on the context and purpose of the evaluation s/he conducted. Responses were reviewed by the researcher and themes were identified based on the responses provided. The goal of the coding was to identify recurrent themes using an open coding approach (Springer, 2010). Using the open coding approach, the responses were analyzed and coded. In addition, an outside person with graduate level experience with qualitative data, was asked to review the 34 responses and identify themes. The researcher and outside person reviewed the data independently and both identified four themes. The researcher and outside person discussed and agreed upon names for the four themes. The four themes identified were:

- Accreditation/Standards,
- Overall program/program effectiveness,
- Information seeking, and
- Curriculum.

Responses were coded into more than one category when applicable. An inter-rater reliability check was conducted and resulted in a 91% agreement among the researcher and outside person. Three responses were not coded the same by the researcher and outside person. An agreement was reached regarding the three responses and the
responses were categorized accordingly. One response indicated “N/A” and therefore was not included in the coding analysis. Twelve (34%) of the responses were coded as Accreditation/Standards. The Accreditation/standards theme encompasses responses indicating district alignment with standards of an accrediting body, compliance with standards set by a governing body, and evaluation of programs specifically for accreditation. One respondent stated, “I evaluate the schools against the AVID Essentials to determine if they will become an AVID certified site or not (on a national scale).”

Overall program/program effectiveness had the most responses with twelve responses (34%). Some responses focused specifically on the effectiveness of a program: “Effectiveness of ESE programs in our school district,” “I evaluated the effectiveness of technology professional development,” and “Examined the effectiveness of Pre-K programs” whereas others discussed evaluating the program overall. One respondent said,

> We were tasked to conduct an evaluation of a STEM education program. The purpose of the program was to provide summer training sessions to teachers and follow-up trainings during the school year to increase teachers' knowledge of STEM areas and increase their efficacy in teaching STEM subjects. – 18

Nine (26%) responses were identified as information seeking meaning the focus of the evaluation was to find out more information about the program and the way the program operation as well as areas in need of improvement. An information seeking response included,

> The evaluation was conducted at the request of the School Board and was designed to provide information about the SES program in the district. While the program is federally mandated under current NCLB legislation, and therefore must be continued, information about the program, the schools involved, the providers, and the students served may help to inform future decisions about exactly how the
program operates and indicate possible program improvements. – 08

Evaluation of curriculum consisted of four (11%) responses. Curriculum evaluations consisted of evaluations to determine the appropriateness of the curriculum, fidelity, rigor, and/or coverage of required information. One respondent reported,

The purpose of the evaluation was to examine the curriculum of one of our AP science programs. Many students were not passing the AP exam with a 3 or higher in this particular subject and the district wanted to know if our curriculum was the problem. We have gone through several different teachers, so the superintendent thought the curriculum may be the problem rather than the instruction. – 16

The qualitative data provided information regarding the context and purpose of evaluations carried out within school districts. Complying with accreditation, federal, and state governing bodies was present in the responses. Responses indicate program evaluations are being carried out to find out information on a variety of programs offered in school districts. Alignment with internal and external standards, program operation, program effectiveness and impact, and the populations served by programs were described by respondents as information sought from the selected program evaluations.

To compliment the context and purpose item, another open-ended item asked respondents the extent to which the evaluator’s (the respondent’s) decisions made during the evaluation were influenced by political issues surrounding the program and/or school district. Eighteen people answered the item. The majority of respondents (56%) clearly indicated they did not think his/her decisions were influenced by political issues. One participant was not clear as to whether or not political issues influenced their decisions with one person stating “Depends on what you are evaluating…” while another person
said “somewhat”. Six responses (33%) indicated some type of influence from political issues. One respondent stated the following about political influences:

At every step. School Districts = politics. As an internal evaluator you have to be prepared to fight for what you think is right and to work within the system to make changes that are politically and systemically palatable. For the most part this goes back to reframing your arguments and working with the people who are trying to drive the decisions with political agendas. – 43

Many respondents recognized the politics surrounding the program and district but reported not being influenced by the politics when making decisions about the evaluation. One person stated,

As an evaluator, the decisions that I offer are not influenced by political issues surrounded by the program or the school district. The information that the stakeholders use from the evaluations are often influenced by the political climate! – 46

The respondent recognized the political aspects surrounding the evaluation but the evaluator reports not being influenced when making decisions.

The final item on the survey allowed respondents to provide any additional information they wanted to share regarding their evaluation. Four participants provided a response. One response detailed information regarding requirements for a funding agency stating,

Like many projects I work on, we had to comply with the information written in the grant proposal. We were able to add and remove some things to meet the needs of the project. Many times budget restraints prevent additional types of data collection, like interviews and focus groups. – 55
The respondent carried the evaluation out the way the evaluation was written in the grant proposal. Evaluation plans submitted in grant proposals can play a role in the resources available to carry out evaluations which may affect the decisions evaluators make related methods, values, and use.

Another response provided information about supplemental education service (SES) providers which is part of a NCLB, Title I school service. The respondent reported practices of an evaluation of SES providers. Authorized providers are approved by the state and by law schools must contract with a state approved provider. Evaluation results could not be implemented at the school or district level because the policy of contracting with specific providers was required by state law. Further, the respondent recommended improving the process of dealing with private providers of supplemental education services at the state level which is out of the district’s control in many aspects.

Two additional respondents provided information regarding their evaluation. One person mentioned political influences from the previous item stating, “it is difficult to eliminate political influences when the people who work above you are pushing for certain things.” The respondent seemed to be following up the previous answer in which the respondent indicated s/he did not find outside political issues to influence decision making but did state “there are always people trying to push for something.” The last response included in the last item on the survey indicated the recommendations from the evaluation were effective in reaching the program’s goals when implemented with fidelity. Fidelity of program implementation as well as fidelity in carrying out recommendations from formative evaluations could both be built in to an evaluation plan.
Summary of open-ended items. The open-ended items provided details on the context and purpose of the evaluations carried out. The responses revealed two trends in practice which coincide with the dimensions found in the principal component analysis: collecting information for decision making, and following procedures and guides set forth by granting agencies. Respondents indicated evaluations were conducted primarily to learn more about a program and when the evaluation was conducted for a grant, the approach was governed by the granting agency’s guidelines. For example one respondent stated the main purpose of the evaluation was to “…inform future decisions about exactly how the program operates, and indicate possible program improvements,” and “Always looking to improve our system,” providing evidence of the need of information for decision making. In terms of conducting an evaluation as part of a grant, one person responded “Like many projects I work on, we had to comply with the information written in the grant proposal,” indicating an emphasis on procedures. Overall the open-ended responses provided additional details on the practice of evaluation and corresponded with some of the dimensions identified in the principle component analysis.

Interviews. Two conversations took place in the form of an informal semi-structured interview to collect more information on the way evaluations were carried out in school districts. One interview was based on an experience conducting an evaluation of Supplemental Education Services (SES) after-school tutoring services, specifically after-school tutoring programs. The second interview was based on an evaluation of a grant funded Pre-Kindergarten (Pre-K) program. A list of questions for the interview was generated by the researcher prior to speaking with the participants on the phone. The interview questions can be found in Appendix C.
To start the conversation the evaluator shared some background information about the evaluation conducted. To maintain confidentiality of the person and school district, the name, gender, and school district of the interviewee are not identified. Each interview is summarized individually to describe each individual experience. The testimonies reported by the school district evaluators represent their perception of the evaluation they carried out and the school district they worked for. Statements related to the size of the school district and structure of personnel were based on the reported perception of the interviewee.

The evaluator reporting on the SES provider evaluation worked at a mid-sized school district and holds the title of Evaluation Specialist. The evaluator was asked by the superintendent to conduct an evaluation of the SES providers who provided afterschool tutoring at elementary school sites in the school district. As an evaluation specialist for the school district the evaluator was the person selected to conduct the evaluation because the task was the main part of the job description. The evaluator was the lead for the project which s/he reported was typical practice for projects of similar size. Larger projects would include additional personnel to carry out the evaluation allowing more than one person to play a large role in the process.

The school district uses internal evaluators for almost all evaluations unless a grant specifically requires an external person or team to be used. In the case of the SES provider evaluation, no outside money or additional money was set aside for the evaluation. Situations arising which the superintendent or school board ask for something, the staff usually has to use whatever resources are already available and do
whatever it takes to get the task completed. The administration assumes projects like this one as part of their job responsibilities.

The primary stakeholder for the evaluation was the superintendent, and building administrators such as principals and assistant principals were secondary stakeholders. Stakeholder involvement consisted of initial conversations with the superintendent and then the evaluator was directed to work with select principals and assistant principals. The principals and assistant principals selected were the ones who raised questions about the SES providers to the superintendent. The evaluator contacted the principal or assistant principal at each school with afterschool tutoring offered by SES providers to include in the evaluation process. Some responded and wanted to be included in the process while others did not show interested in involvement of the process.

Fostering buy-in for the evaluation from the stakeholders wasn’t very difficult because the school administrators wanted to know what was going on in their schools. The evaluator thought most of the administrators were neutral about the evaluation because the program was not something they had control over. The administrators did not appear to be against or for the evaluation. The evaluator thought this was most likely because of the administrators’ lack of control over the contracts with the SES providers. In addition, the SES providers and their services do not relate to the job performance evaluations of the administrators. The evaluator said this was not typical or atypical for the district. Administrator cooperation does not always occur. Some projects have required time to be used to obtain buy-in from principals, assistant principals, and teachers. For this evaluation, the evaluator estimated the administration was about 50%
vested in the evaluation. Similarly the superintendent was the person who commissioned
the evaluation but did not provide a formal deadline for submission of results.

The evaluation questions were developed collaboratively among the evaluator,
superintendent, and two school administrators. Once the questions were developed and
agreed upon, the evaluator moved forward with the evaluation. The evaluator reported
using Stufflebeam’s CIP model as a framework for developing typical evaluation
projects. The evaluator did not refer to the CIP model when developing this project
because s/he thought this project was somewhat straightforward. Looking back on the
project the evaluator thinks the CIP model guided the project without directly referring to
it in the report.

The SES providers were a separate entity from the school district and were
selected by the state instead of the school districts to provide tutoring. The lack of control
over authorized providers removed political tension for the project. The evaluator thought
there was minimal or no political influences because the school district workers were
separate from the people who were involved in the program under evaluation.

Data collection strategies were governed by the data the government required the
district to collect. The data consisted of standardized test scores, attendance records, and
clerical documentation. Additional information was collected via informal interviews
because it could be done by the evaluator and would not cost additional money to the
district. Test scores and demographic information were housed in the district data system
and were pulled by the data people.

Although the data and other necessary information were in the district’s system,
there were some barriers to obtaining all of the information in a timely manner. The data
department was constantly getting requests to provide data to different people and organizations. The data department complained the requests take up a lot of time which means a lot of money. No additional money was set aside for this project so the person power to provide data to us was at the convenience of the data person. The data person had to pull data for sources paying money for the information via grant money. The district does not have an institutional review board (IRB) for projects like this one. Individuals who want to use district data outside of the district must complete a data request form. Internal affairs like this one do not go through a process.

After all of the information was retrieved and the evaluation report was generated, the report was provided to the superintendent who shared the findings with building administrators who have SES providers in their schools. This was a typical practice in the district. The evaluator reported s/he would not change anything if s/he were to conduct the evaluation again. The main outcome of the evaluation was something the schools themselves ultimately could not control and proper action would have to take place at the state level. Although the outcome was not something the schools or district could immediately change, the evaluator said the administrators were satisfied with the results and thought they could do other things in their schools as a result of knowing more about the SES providers for afterschool tutoring.

As a final thought on recalling the evaluation process for the SES providers, the evaluator thought that if the findings did show the individual schools and/or district could make additional improvements to increase the academic achievement of the students, then the superintendent would have provided time and funds to allow for additional development and evaluation.
The evaluator who discussed the grant funded Pre-K program evaluation, reported working in a mid to large sized school district as part of the Accountability Office. The district received a grant for a Pre-K program at multiple school sites within the school district. The evaluator was assigned to conduct the evaluation and write the evaluation report for submission to the grant funding agency. In the past, the same grant required an external person or team to complete the evaluation for the project, however changes were made to the grants due to budget cuts and internal personnel were able to complete the reporting.

The evaluator served as the lead evaluator for the project which was typical practice for the district. Each evaluation project that came through the department was assigned to one main person. Depending on the size of the project, other people were assigned tasks on the project but one person was typically put in charge of each project. A project of this kind was typical in size and scope of other evaluation projects. The evaluator reported many grant funded projects occur throughout each year so they try to plan ahead to provide each person with a steady flow of projects. Scheduling appeared to be the main reason why the evaluator was assigned to be the lead on the evaluation.

No external contracts were used for the evaluation but the evaluator reported the use of external evaluators as a common practice throughout each year. Projects of high political stigma were typically assigned to external evaluators as well as projects which may need to be done on short notice when the staff was tied up with other deadlines. The evaluator estimated the district probably contracts about five external evaluators a year to write evaluation reports. The term “write reports” was used because the evaluator considered most of the evaluation work to be report writing for the funding agencies or
government. The interviewee said it would be difficult to estimate the portion of evaluations conducted by external people because each project carries different weight. Some projects are five years long whereas others are two years long. External people are also not typically used for a whole project, but smaller parts of a larger project.

The stakeholders for the project could consist of educators, parents, and the community at large but the evaluator stated the funding agency was the real stakeholder. The reason the evaluation was being conducted was for the awarding grant agency. The reporting was for the funding agency and things were carried out a certain way because of the way the evaluation plan was written in the grant proposal. The grant proposal clearly stated the expectations, goals, and outcomes for the program as well as the way each would be measured. The data sources and minimum requirements were also set in the proposal. In essence, the evaluation used the grant proposal as a framework for the evaluation. The evaluator was not part of the grant writing team and did not set any of the criteria.

In general, the teachers of the Pre-K program were vested in the program. The staff knew the program was grant funded and knew the program was supported at large by the grant funds. Each staff member cooperated with any questions or information needed for the evaluation so buy-in was not difficult for this evaluation. For the most part, school administrators accepted the Pre-K program but approximately half of them appeared to be doing it because they were required to do so and probably thought it looked good for them and their school. The evaluator was not sure if the administrators actually thought the program was valuable.
The evaluation was very accountability focused. The main idea of the program was to get the children to be at the appropriate level when they entered Kindergarten. Many assessments were administered to the children to evaluate their entry level abilities, progress monitoring throughout the program, and readiness upon leaving the program to enter Kindergarten. The grant proposal set specific requirements for gain scores and percent increase as a result of attending the program. The proposal was very numbers driven, essentially no qualitative data were needed to answer the evaluation questions. The evaluator thought qualitative data would have been helpful to figure out why the program was helping children or not helping children.

The evaluator did not perceive any political forces due to the project. In areas in which the program was not meeting the goals, the evaluator was asked to provide evidence of the ways the program was attempting to make progress. The evaluator stated this was typical of the district politics. Although the evaluator said no one told them to lie, s/he was asked to provide additional information not asked for by the grant agency to justify or make a case for the program. The evaluator did not view this as being unethical but the district’s attempt to maintain funding for the program. In situations where there is not progress towards a goal, the evaluator says s/he stated what needed to take place in order for progress to be made.

Pre- and post assessment were administered to the program participants by the teachers. Some of the data and pre- and post assessments were from teacher observations. Observation scales were completed by the teachers. This was typical practice for programs targeted for young children since the children cannot read and take a multiple choice test. The evaluator thought the teachers collected rich data but questioned the
accuracy and fidelity of some of the teachers’ assessments. Teachers knew forms and data were due by a certain day and some teachers may have filled out the forms quickly at the last minute without proper administration. The teachers were trained to administer each assessment and the administrations and reporting were simple but required time. Teachers who completed the forms at the last minute would not be able to provide an accurate evaluation of each child’s abilities. There was also no inter-rater reliability which the evaluator thought would greatly increase the credibility of the student reports for observational data. Teachers knew the students should show growth so they could easily make the scores increase over time when recording data.

Data were collected throughout the program by the teachers and program directors. The evaluator was not responsible for collecting the information, but did have problems getting the data electronically. Teachers completed the assessments on paper and submitted the paper files to the program director and the evaluator. The evaluator created a system for the teachers to enter the data electronically for each student in order to save time in the reporting process. The project was considered to fall in the exempt category in terms of an Institutional Review Board (IRB) review. The district had to provide justification of the protection of human rights for the grant agency prior to the start of the program.

The findings were reported back to the funding agency and shared with the superintendent, school administrators, and program leaders. The program leaders were required to provide a plan for improvement in any areas identified as weaknesses or any components not making progress or meeting goals. The evaluator was available to work with program leaders in developing the plan but stated that no one utilized the
opportunity. This was considered typical practice for use of evaluation findings and communication after the report was generated.

The study was a formative evaluation which meant if the program leaders did not use the information provided by the evaluator, and progress was not made the following year, funding could be pulled from the project. The program leader could be fired as well as Assistant Principals or Principals could be in job trouble as well as a result of a failed grant program under his/her jurisdiction. In this case, none of the findings were alarming so the evaluator was not surprised when none of the administrators or program leaders asked for additional feedback. According to the evaluator’s past experience conducting evaluations for the district, stakeholders typically only contact the evaluator for help making improvements if the results were high-stakes and funding was in jeopardy of being pulled.

The evaluator thought the final evaluation report was useful for the district and stakeholders. S/he thought there may be problems in the program related to infidelity of the instruction, assessment, and reporting by the staff. From the information requested by the granting agency, these problems may not have been apparent. This issue was identified as an area worth further study. As previously described, the evaluation was set up to include specific data and criteria. The evaluator felt as though programs could meet all of the criteria stated in the proposal but still have problems meeting the goals of the program, preparing children for kindergarten readiness. If an external evaluator was used program employees may have revealed information about not doing certain things they were supposed to do them. On the contrary, employees may have given the same
impression to the external evaluator(s) or there may have been no problems with program fidelity.

If the evaluator were to conduct the evaluation again, s/he would have requested to be involved with the evaluation plan during the writing process for the grant proposal. The grant team told the evaluator the proposal was worded a certain way so the district would be awarded funds. The district wanted to make sure money was available to offer needed services but the evaluator thought sometimes the policies in grants may compromise the way programs are evaluated. After working in the school district over the last eight years, the evaluator indicated that most of the evaluations conducted in his/her district were not really evaluations but data analyses and reporting of test scores. If policy only asks for this information and no additional information then it is unlikely grant funds will be awarded to fully evaluate programs. The evaluator stated s/he doesn’t see or hear about a lot of full evaluations being carried out in the public schools, “everything is about the numbers”.

**Summary of interviews.** The SES provider evaluation and grant funded Pre-K program evaluation interviews both provided insight into the ways evaluations were carried out in two school districts. Both of the evaluations were conducted by internal evaluators, however the SES provider evaluation was commissioned internally and the grant funded Pre-K evaluation was required for an outside funding source. The Pre-K evaluation was restricted to criteria written in the grant proposal which the evaluator found to be shortcoming for demonstrating the true performance of the program. In relation to the six dimensions of evaluation practice, following the grant criteria in the Pre-K evaluation related to an emphasis on procedures. The interviewee stated, “I used
the grant proposal and information from the program funders to guide the evaluation.” demonstrating a procedural emphasis.

The internally commissioned SES provided evaluation included an emphasis on people over procedures. No prior procedures were in place for the evaluation and the evaluator contacted various stakeholders to develop the evaluation questions. Although some stakeholders were involved in the evaluation process, the SES provider evaluation did not engage/embody the users but focused more on the evaluator as a mediator. The interviewee stated, “the district did not have much control over the providers, so the principals just wanted to know what was going on in their schools in an info briefing kind of way.” Each evaluator reported s/he would not change very much if given the chance to conduct the evaluation under the same circumstances. Overall the interviews provided information on the practice of evaluation which was related back to three of dimensions of evaluation practice.

**Summary of research question two.** Evaluation practice can be described using the six factors found in this study: Holistic, Mixed Method Decision Making, Procedures Valued, People Valued, Users Engaged/Embodied, Evaluator as Mediator. Cronbach’s alphas were calculated to obtain estimates of reliability for the items.

The reliability estimates obtained for Holistic (α = .82), Mixed Method Decision Making (α = .72), Procedures Valued (α = .78), People Valued (α = .64), Users Engaged/Embodied (α = .87), Evaluator as Mediator (α = .77) dimensions were acceptable and appropriate to use to describe evaluation practice. Qualitative data obtained from the survey questionnaire and the interviews provided additional
information on the actual practice of evaluators in school districts. The qualitative data provided context for the quantitative data obtained.

**Research Question Three**

To investigate the relationship between evaluator preparation and evaluation practice, and to determine if professional and/or personal characteristics distinguish the practice of evaluation three one-way MANOVAs were conducted. The dependent variables used to describe evaluation practice were the six dimensions found to describe evaluation practice: Holistic, Mixed Method Decision Making, Procedures Valued, People Valued, Users Engaged/Embodied, Evaluator as Mediator, in each of the one-way MANOVAs. Univariate post hoc follow-up $F$ tests were used to analyze and significant main effects. Data were screened for normality and outliers prior to conducting the one-way MANOVA. In addition, MANOVA's assumption of homoscedasticity was checked using Box’s M due to the differences in group sizes for some of the analyses. The Box’s M test was not significant which indicated the assumption of homoscedasticity was not violated. Box’s M values ranged from 43.8 to 109.2. The means, standard deviations, minimum, maximum, skewness and kurtosis are provided in Table 8 for each dimension of evaluation practice. The skewness and kurtosis revealed appropriate values for normality with no standard errors greater than 2, and observations were independent of each other. All assumptions were met and the three one-way MANOVAs were conducted.

**First one-way MANOVA.** The first one-way MANOVA used highest degree obtained as the independent variable and evaluation practices as the dependent variables. Table 9 provides descriptive statistics for the variables used in the MANOVA. The one-
way MANOVA for highest degree obtained and evaluation practice was statistically significant for differences between groups using Wilks’ Lambda ($\lambda = .60$, $F (12, 96) = 2.38; p = .01$).

Table 9

| Means and Standard Deviations for Highest Degree Obtained and Evaluation Practice |
|-------------------------------|-------|------|------|------|------|------|------|
| Highest Degree               | N    | Mean | SD  | SE  | Skewness | Kurtosis |
| Holistic                     |      |      |     |     |          |          |
| Bachelors                    | 13   | 41.40 | 17.70 | 0.52 | 0.69 | -0.02 | 1.33 |
| Master’s                     | 60   | 50.90 | 7.36  | -1.15 | 0.43 | -1.26 | 0.86 |
| Doctorate                    | 48   | 41.23 | 17.56 | -0.75 | 0.44 | 0.05  | 0.86 |
| Total                        | 121  | 44.71 | 15.22 | -0.85 | 0.29 | 0.50  | 0.57 |
| Mixed Method Decision Making |      |      |     |     |          |          |
| Bachelors                    | 13   | 59.80 | 17.20 | -1.31 | 0.69 | 0.53  | 1.35 |
| Master’s                     | 60   | 55.10 | 10.85 | 0.32  | 0.44 | -1.26 | 0.86 |
| Doctorate                    | 48   | 47.15 | 13.13 | -0.35 | 0.44 | -0.56 | 0.86 |
| Total                        | 121  | 52.25 | 13.89 | -0.33 | 0.30 | -0.40 | 0.58 |
| Procedures Valued            |      |      |     |     |          |          |
| Bachelors                    | 13   | 43.90 | 7.85  | -1.32 | 0.47 | 2.72  | 0.92 |
| Master’s                     | 60   | 47.80 | 13.16 | -0.79 | 0.69 | -1.10 | 1.35 |
| Doctorate                    | 48   | 38.15 | 10.88 | -0.51 | 0.43 | -1.08 | 0.83 |
| Total                        | 121  | 41.93 | 10.85 | -0.68 | 0.30 | 0.29  | 0.59 |
| People Valued                |      |      |     |     |          |          |
| Bachelors                    | 13   | 30.80 | 8.09  | -0.93 | 0.69 | 0.11  | 1.35 |
| Master’s                     | 60   | 27.00 | 6.68  | -0.46 | 0.44 | 0.52  | 0.86 |
| Doctorate                    | 48   | 26.00 | 8.73  | -2.31 | 0.44 | 5.19  | 0.86 |
| Total                        | 121  | 27.21 | 7.99  | -1.55 | 0.30 | 3.98  | 0.58 |
| Users Engaged/Embodied       |      |      |     |     |          |          |
| Bachelors                    | 13   | 37.00 | 12.84 | -0.19 | 0.69 | -2.03 | 1.35 |
| Master’s                     | 60   | 38.70 | 5.33  | -1.28 | 0.44 | 1.11  | 0.86 |
| Doctorate                    | 48   | 35.00 | 11.72 | -2.13 | 0.43 | 4.98  | 0.83 |
| Total                        | 121  | 36.68 | 10.10 | -1.41 | 0.29 | 2.45  | 0.57 |
| Evaluator as Mediator        |      |      |     |     |          |          |
| Bachelors                    | 13   | 31.40 | 6.88  | -0.20 | 0.69 | -1.86 | 1.35 |
| Master’s                     | 60   | 26.40 | 7.01  | -0.29 | 0.44 | -0.82 | 0.86 |
| Doctorate                    | 48   | 28.31 | 9.50  | -1.01 | 0.44 | -0.31 | 0.86 |
| Total                        | 121  | 28.18 | 8.29  | -0.63 | 0.30 | -0.57 | 0.58 |
Follow-up univariate comparisons showed that there was a significant difference in the “Mixed Method Decisions Making” variable of evaluation practice \( [F(2, 53) = 4.05, p < .05] \) and “Procedures Valued” variable \( [F(2, 53) = 3.70, p < .05] \). Further, the data showed the remainder of the variables were not statistically significant, Holistic \( [F(2, 53) = 2.73, p = .07] \), People Valued \( [F(2, 53) = 1.33, p = .27] \), Users Engaged/Embodied \( [F(2, 53) = 0.76, p = .47] \), and Evaluator as Mediator \( [F(2, 53) = 1.23, p = .30] \). Multiple comparisons were made within the values variable for evaluation practice, and a statistically significantly higher mean was found between respondents whose highest degree obtained was a bachelor’s degree than those whose highest degree obtained was a doctorate with a mean difference of 12.65 \( [SE = 4.90, p = .03, Cl_{.95} = 0.82, 24.47, d = .87] \) in the Mixed Method Decision Making dimension, and 9.66 \( [SE = 3.85, p = .04, Cl_{.95} = 0.35, 18.94, d = .11] \) in the Procedures Valued dimension. No other groups within highest degree obtained had a statistically significant mean difference.

**Second one-way MANOVA.** The second one-way MANOVA was conducted using highest degree earned alignment as the independent variable and evaluation practice as the dependent. Results from the original categories listed for highest degree alignment contained some small cell sizes. The categories were reviewed for similarities and combined to form categories of similar degrees. Evaluation/Research Methods and Advanced Quantitative Methods are closely related with quantitative methods falling under the research methods umbrella. Psychology does not necessarily include advanced quantitative methods but advanced statistics such as factor analysis and other psychometric analyses are considered quantitative methods which made psychology an appropriate degree to group with Evaluation/Research Methods and Advanced
Quantitative Methods. Due to the similar background training of the three degrees, it made sense to place the respondents into one category. The category is referred to as Evaluation/Research Methods.

Teaching/Learning (Education) and School Administration were listed as separate categories on the survey and were kept separate for the analysis even though school administration could be categorized as a broad area within education such as teaching and learning. School administration programs require a variety of courses which are different than other education programs due to the administration component. Teaching/Learning, Evaluation/Research Methods, and School Administration were the three categories used to conduct the second MANOVA. Respondents who identified their highest degree obtained as public health, computer science, and social work each had very small cell sizes and it did not seem to make sense to categorize into one of the three categories and therefore were not included in the analysis for this MANOVA. Table 10 provides descriptive statistics for area of highest degree alignment by evaluation practice.

Table 10

<table>
<thead>
<tr>
<th>Area of Highest Degree</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching/Learning</td>
<td>65</td>
<td>49.17</td>
<td>12.78</td>
<td>.41</td>
<td>.49</td>
<td>-.81</td>
</tr>
<tr>
<td>Evaluation/Research Methods</td>
<td>28</td>
<td>49.80</td>
<td>5.71</td>
<td>.24</td>
<td>.69</td>
<td>-1.35</td>
</tr>
<tr>
<td>School Administration</td>
<td>18</td>
<td>25.50</td>
<td>18.92</td>
<td>-.45</td>
<td>.75</td>
<td>-1.81</td>
</tr>
<tr>
<td>Total</td>
<td>111</td>
<td>44.81</td>
<td>15.81</td>
<td>-.85</td>
<td>.30</td>
<td>.50</td>
</tr>
<tr>
<td>Mixed Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching/Learning</td>
<td>65</td>
<td>56.33</td>
<td>12.01</td>
<td>.07</td>
<td>.49</td>
<td>-1.23</td>
</tr>
<tr>
<td>Evaluation/Research Methods</td>
<td>28</td>
<td>56.00</td>
<td>12.18</td>
<td>-.19</td>
<td>.69</td>
<td>-1.86</td>
</tr>
<tr>
<td>School Administration</td>
<td>18</td>
<td>46.25</td>
<td>18.16</td>
<td>.10</td>
<td>.75</td>
<td>-2.57</td>
</tr>
<tr>
<td>Total</td>
<td>111</td>
<td>54.33</td>
<td>13.62</td>
<td>-.34</td>
<td>.30</td>
<td>-.41</td>
</tr>
</tbody>
</table>
The one-way MANOVA for highest degree alignment and evaluation practice was statistically significant for differences between groups using Wilks’ Lambda ($\lambda = .31, F (12, 68) = 4.56; p < .01$). Follow-up univariate comparisons showed that there was a significant difference in the Holistic [$F(2, 39) = 10.96, p < .01$], Procedures Valued [$F(2, 39) = 9.54, p < .01$], People Valued [$F(2, 39) = 4.87, p = .01$], Users Engaged/Embodied [$F(2, 39) = 3.77, p = .03$], and Evaluator as Mediator [$F(2, 39) = 6.13, p = .01$]. The Mixed Method Decisions Making variable did not have a significant difference [$F(2, 39) = 1.81, p = .18$]. Multiple comparisons were made within the Holistic, Procedures Valued, People Valued, Users Engaged/Embodied, and Evaluator as Mediator dimensions. Table 11 provides mean differences, standard errors, significance, confidence intervals, and effect size ($d$) for significant findings. In the Holistic

<table>
<thead>
<tr>
<th></th>
<th>Teaching/Learning</th>
<th>Evaluation/Research Methods</th>
<th>School Administration</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Procedures Valued</strong></td>
<td>65</td>
<td>47.58</td>
<td>7.78</td>
<td>-1.79</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>32.60</td>
<td>8.60</td>
<td>-.70</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>44.25</td>
<td>13.06</td>
<td>-1.19</td>
</tr>
<tr>
<td></td>
<td>111</td>
<td>43.38</td>
<td>10.87</td>
<td>-.69</td>
</tr>
<tr>
<td><strong>People Valued</strong></td>
<td>65</td>
<td>30.33</td>
<td>5.80</td>
<td>-.14</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>27.80</td>
<td>3.08</td>
<td>-.11</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>20.50</td>
<td>14.47</td>
<td>-.60</td>
</tr>
<tr>
<td></td>
<td>111</td>
<td>27.86</td>
<td>8.42</td>
<td>-.56</td>
</tr>
<tr>
<td><strong>Users Engaged/Embodied</strong></td>
<td>65</td>
<td>40.00</td>
<td>7.62</td>
<td>-1.37</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>35.00</td>
<td>4.06</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>28.50</td>
<td>19.98</td>
<td>-.58</td>
</tr>
<tr>
<td></td>
<td>111</td>
<td>36.62</td>
<td>11.16</td>
<td>-1.42</td>
</tr>
<tr>
<td><strong>Evaluator as Mediator</strong></td>
<td>65</td>
<td>31.50</td>
<td>7.24</td>
<td>-1.14</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>22.20</td>
<td>7.47</td>
<td>-.18</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>30.50</td>
<td>6.52</td>
<td>.59</td>
</tr>
<tr>
<td></td>
<td>111</td>
<td>29.10</td>
<td>8.02</td>
<td>-.64</td>
</tr>
</tbody>
</table>
dimension, those with their highest degree in teaching/learning and those with their highest degree in evaluation/research methods had a significantly higher mean than those with their highest degree in school administration. No significant difference was present between teaching/learning and evaluation/research methods in the Holistic dimension.

In the Procedures Valued dimension there was a significant difference between teaching/learning and evaluation/research methods, as well as school administration and evaluation/research methods. Those with their highest degree in evaluation/research methods had a significantly lower mean. No significant differences were found between teaching/learning and school administration.

In the People Valued dimension, respondents with their highest degree in teaching/learning had a significantly greater mean than those with their highest degree in school administration. Similarly, in the Users Engaged/Embodied dimension, those with their highest degree in teaching/learning had a significantly greater mean than those with their highest degree in school administration education. There were no significant differences between school administration and evaluation/research methods, or teaching/learning and evaluation/research methods in either dimension.

The Evaluator as Mediator dimension had two significant differences. There was a significant difference among respondents with their highest degree in teaching/learning and evaluation/research methods, and school administration and evaluation/research methods. Teaching/learning and school administration each had a significantly greater mean than evaluation/research methods. There was not a significant difference between teaching/learning and school administration.
In comparing evaluation practice and highest degree alignment, those with their highest degree aligning with teaching/learning and those aligning with school administration differed significantly in the Holistic, People Valued, and User Engaged/Embodied dimensions. Teaching/learning and evaluation/research methods differed significantly on the Procedures Valued, and Evaluator as Mediator dimensions. School administrators and evaluation/research methods differed significantly in the Holistic, Procedures Valued, and Evaluator as Mediator dimensions. Each of the significant differences found were large in magnitude with each of the effect sizes described as large.

Table 11

*Mean Differences and Confidence Intervals for Significant Differences among Highest Degree Alignment and Evaluation Practice*

<table>
<thead>
<tr>
<th>Holistic</th>
<th>Teaching/Learning</th>
<th>School Administration</th>
<th>Lower Group Mean Diff</th>
<th>SE</th>
<th>p</th>
<th>95% Confidence Interval</th>
<th>Higher Group Mean Diff</th>
<th>SE</th>
<th>p</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval/Research Methods</td>
<td>Higher Group</td>
<td>Lower Group</td>
<td>23.67</td>
<td>5.29</td>
<td>.00</td>
<td>10.77 - 36.57</td>
<td>24.30</td>
<td>6.15</td>
<td>.00</td>
<td>9.31 - 39.29</td>
</tr>
<tr>
<td>School Administration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedures Valued</td>
<td>Teaching/Learning</td>
<td>Higher Group</td>
<td>14.98</td>
<td>3.44</td>
<td>.00</td>
<td>6.61 - 23.36</td>
<td>11.65</td>
<td>4.33</td>
<td>.03</td>
<td>1.09 - 22.21</td>
</tr>
<tr>
<td>School Administration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>People Valued</td>
<td>Teaching/Learning</td>
<td>Higher Group</td>
<td>9.83</td>
<td>3.15</td>
<td>.01</td>
<td>2.15 - 17.51</td>
<td>11.50</td>
<td>4.28</td>
<td>.03</td>
<td>1.08 - 21.92</td>
</tr>
<tr>
<td>Users Engaged/Embodied</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Third one-way MANOVA. The third one-way MANOVA was conducted using the first and second topic, training in evaluation theory and program evaluation from the list of evaluation topics. Respondents were to indicate the amount of training completed in topics related to evaluation. The first two topics were of interest to answer the third research question regarding the relationship between training in evaluation and evaluation practice. For each topic listed, respondents checked the type of training received based on the eight types of training:

- One or more full courses,
- Substantial part of a full course,
- One or more workshops,
- Substantial part of a workshop,
- Small part of a workshop,
- Staff development,
- Independent learning, and
- Minimal exposure or no formal training.

Respondents were instructed to select all that apply for the list of twelve evaluation related topics. A total score was calculated based on the total amount of training reported by each participant. In order to calculate the extent of training for each respondent, weights were assigned to each of the eight types of training. Weights were strategically selected to allow for the sum of training to provide a representative value of
the total amount of training from the categories listed. The category “One or more full courses” is the single highest training a respondent could have completed. Substantial part of a full course is the second greatest single category of formal training. The “One or more workshops” category was estimated to be the third greatest formal training respondents may have received. The remainder of the training categories: substantial part of a workshop, small part of a course or workshop, staff development, and independent learning provide training but the extent of the training, length of training is unknown. For this reason, the first three categories were considered to be less subjective in terms of content coverage. Based on the rationale provided, the first category, one or more full courses was given a weight of seven, substantial part of a full course was given a weight of five, one or more workshops was weighted two, small part of a course or workshop, staff development, and independent learning were each given a weight of one. Minimal or no experience was entered as zero.

A total score was calculated for each respondent based on the amount of training for evaluation theory and program evaluation. The total scores ranged from zero to 17 (note: respondents selected all applicable trainings in topic area). An individual who completed a substantial part of a course (weight of five) and one or more full workshops (weight of two) would receive a total score of seven which is equivalent to one or more full courses (weight of seven) based on the weights selected. The weights and total scores for training provide only an estimate of the extent of training in evaluation theory and program evaluation based on the self rated perception of respondents’ training in the evaluation topics. Total scores for respondents ranged from zero to 17. The cell sizes for each value were not large enough to conduct analyses using the total score values as
categories. Total scores were combined into three categories to make statistical
comparisons of the different extents of training.

Total scores of zero indicated minimal or no formal training in evaluation theory
or program evaluation. The weight for a minimum of one course in either evaluation
theory or program evaluation would provide a total score of at least seven. Respondents
with total scores ranging from one to six were considered to have some exposure to
evaluation theory and program evaluation. Individuals with total scores ranging from zero
to six were placed into the “Minimal to Novice Training” category. Total scores of seven
or higher made up the third category which were considered “Intermediate to Advanced
Training”. Table 12 provided descriptive statistics for the extent of training in evaluation
by evaluation practice.

Table 12

Descriptive Statistics for Extent of Training in Evaluation Theory and Program Evaluation by
Evaluation Practice

<table>
<thead>
<tr>
<th>Level of Training</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal to Novice</td>
<td>71</td>
<td>41.13</td>
<td>18.68</td>
<td>-.73</td>
<td>.38</td>
</tr>
<tr>
<td>Intermediate to Advanced</td>
<td>59</td>
<td>48.85</td>
<td>8.48</td>
<td>-.40</td>
<td>.43</td>
</tr>
<tr>
<td>Total</td>
<td>130</td>
<td>44.71</td>
<td>15.22</td>
<td>-.85</td>
<td>.50</td>
</tr>
<tr>
<td>Mixed Method Decision Making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal to Novice</td>
<td>71</td>
<td>50.47</td>
<td>16.11</td>
<td>-.33</td>
<td>.38</td>
</tr>
<tr>
<td>Intermediate to Advanced</td>
<td>59</td>
<td>54.31</td>
<td>10.74</td>
<td>.07</td>
<td>.44</td>
</tr>
<tr>
<td>Total</td>
<td>130</td>
<td>52.25</td>
<td>13.89</td>
<td>-.34</td>
<td>.41</td>
</tr>
<tr>
<td>Procedures Valued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal to Novice</td>
<td>71</td>
<td>41.73</td>
<td>11.50</td>
<td>-.94</td>
<td>.39</td>
</tr>
<tr>
<td>Intermediate to Advanced</td>
<td>59</td>
<td>42.15</td>
<td>10.28</td>
<td>.33</td>
<td>.44</td>
</tr>
<tr>
<td>Total</td>
<td>130</td>
<td>41.93</td>
<td>10.85</td>
<td>-.69</td>
<td>.29</td>
</tr>
<tr>
<td>People Valued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal to Novice</td>
<td>71</td>
<td>28.13</td>
<td>9.78</td>
<td>-1.96</td>
<td>.38</td>
</tr>
<tr>
<td>Intermediate to Advanced</td>
<td>59</td>
<td>26.15</td>
<td>5.21</td>
<td>-.68</td>
<td>.44</td>
</tr>
<tr>
<td>Total</td>
<td>130</td>
<td>27.21</td>
<td>7.99</td>
<td>-1.56</td>
<td>.30</td>
</tr>
<tr>
<td>Users</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal to Novice</td>
<td>71</td>
<td>34.53</td>
<td>11.96</td>
<td>-1.52</td>
<td>.38</td>
</tr>
</tbody>
</table>

114
The one-way MANOVA for evaluation training and evaluation practice was statistically significant for differences between groups using Wilks’ Lambda ($\lambda = .65$, $F (6,49) = 4.49; p < .01$). Follow-up univariate comparisons showed no significant difference in any of the dimension: Holistic [$F(1, 54) = 3.76, p = .06$], Mixed Method Decision Making [$F(1, 54) = 1.07, p = .31$], Procedures Valued [$F(1, 54) = 0.02, p = .89$], People Valued [$F(1, 54) = 0.85, p = .36$], Users Engaged/Embodied [$F(1, 54) = 3.02, p = .09$], and Evaluator as Mediator [$F(1, 54) = 3.50, p = .07$]. Due to lack of significance no additional analyses or comparisons were made within the dimensions.

**Summary of MANOVAs.** Overall, statistically significant differences were found in two of the three one-way MANOVAs. In terms of highest degree obtained, a statistically significant difference was found between respondents whose highest degree obtained was a bachelor’s degree and those whose highest degree obtained was a doctorate in the Mixed Method Decision Making and Procedures Valued dimensions. Based on discipline area of highest degree, statistically significant differences were found on five of the six dimensions of evaluation practice. Respondents who aligned their highest degree with school administration differed significantly with individuals who held their highest degree in teaching/learning on the Holistic, People Valued, and Users Engaged/Embodied dimensions. Respondents whose highest degree was in teaching/learning differed than those whose highest degree was in evaluation/research.
methods on the Procedures Valued and Evaluator as Mediator dimensions. Lastly, respondents whose highest degree was in school administration and those whose highest degree was in evaluation/research methods differed on the Holistic, Procedures Valued and Evaluator as Mediator dimensions. The extent of evaluation training and evaluation practice revealed no statistically significant differences among the respondents on the six dimensions.

**Overall Summary of Results**

The results from the survey questionnaire and interviews provided useful data on the types of evaluations conducted in school districts and the way evaluations have been carried out. The quantitative and qualitative data collected answered the three research questions presented. A descriptive picture of the background and training of school district evaluators was presented highlighting the academic and professional experiences of respondents. The majority of respondents held the position of a program administrator, program specialists, or program coordinator within their school district. Evaluations consisted mostly of curricular or curriculum, early childhood, magnet, special education, and staff development.

The qualitative data identified funding as a primary reason for conducting evaluations. In some cases evaluations were conducted in a manner to provide information specifically called for by the granting agency. Some program evaluations were based on set criteria from an outside party such as an accrediting body was indicated by several respondents. Similarly, responding to criteria set by granting agencies was also identified. Other evaluations focused on learning more about a curricular program and searching for areas to make improvements. Considering the
context provided by the open ended items, the reporting requirements set forth by accrediting bodies and funding agencies should be considered when interpreting the results of the theory to practice portion of the instrument.

The second part of the instrument was referred to as the theory to practice portion. The instrument was developed based on three dimensions, methods, values, and use. Based on the data collected in this study, the methods, values, and use dimensions did not describe the practice of the K-12 school district evaluators. The results of the principal component analysis indicated six dimensions: holistic, mixed method decision making, procedures valued, people valued, users engaged/embodied, and evaluator as mediator. These six dimensions were used to describe the evaluation practice of the K-12 school district evaluators. Evaluation practice (holistic, mixed method decision making, procedures valued, people valued, users engaged/embodied, evaluator as mediator) served as the dependent variables for each of the one-way MANOVAs. The first one-way MANOVA investigated highest degree held and evaluation practice. A significant difference was found in the mixed method decision making dimension between respondents who held a bachelors degree and those who held a doctorate \( (d = 0.87) \), and among the same groups in the procedures valued dimension \( (d = 0.86) \). The second one-way MANOVA investigated area of highest degree and evaluation practice. A statistically significant difference was found in five of the six dimensions with a difference among respondents with their highest degree in teaching/learning and those aligning with school administration differed significantly in the Holistic \( (d = 1.69) \), People Valued \( (d = 1.17) \), and User Engaged/Embodied dimensions\( (d = 1.01) \). Teaching/learning and evaluation/research methods differed significantly on the
Procedures Valued ($d = 1.93$), and Evaluator as Mediator dimensions ($d = 1.31$). School administrators and evaluation/research methods differed significantly in Holistic ($d = 1.95$), Procedures Valued ($d = 1.15$), and evaluator as mediator ($d = 1.28$). The third one-way MANOVA investigated the relationship between training in evaluation theory and program evaluation and evaluation practice. No statistically significant differences were found. Respondents who had missing data for any of the theory to practice items were not included in the analyses. Missing data may explain why significant differences were not found in some areas.
Chapter Five: Discussion

Overview

This chapter discusses the results presented for the current study. Conclusions are drawn from the analyses and explanations of the results of the research questions are presented. Specifically, conclusions regarding the 1) breadth and depth of the preparation of school district evaluators, 2) reported evaluation practices of school district evaluators, and 3) relationship between evaluator preparation and evaluation practice. A summary of the study’s contributions to evaluation, limitations, and implications for future research are presented.

People conducting evaluation for public schools in Florida comprise a variety of educational and training backgrounds and job positions. The majority of the respondents held job titles such as program coordinator, program administrator, and program specialists. Educational training in evaluation areas and areas of higher education study ranged from high school to doctoral level, and discipline areas such as social sciences, school administration, teaching and learning, quantitative methods, and public health.

Discussion of Findings

The results of this study inform one of the main areas in need of further research in evaluation: identifying the population of evaluators and carving out a picture of who conducts evaluations and how they conduct evaluations. The background and education of K-12 public school evaluators and the extent of their training in evaluation was captured in this study. Christie’s (2001) Theory to Practice instrument was adapted and
administered to obtain information about people who conduct evaluations in K-12 public schools and the ways in which they conduct those evaluations. The original instrument was created with the input of eight evaluation theorists (Richard Berk, Huey-tsyh Chen, J. Bradley Cousins, Elliot Eisner, David Fetterman, Ernest House, Michael Patton, and Daniel Stufflebeam) and focused on three dimensions used to describe evaluation practice: method, values, and use. Unlike Christie’s (2001) study, the purpose of this study was not to link the practice of K-12 public school evaluators to a specific theory, but to use the three dimensions of methods, values, and use as a framework for describing the practice of the K-12 public school evaluators.

**Descriptive findings.** This study offered an inside view of the people conducting internal evaluation and the way evaluation was carried out in public school districts. The demographic information provided interesting findings which may be related to the geographic location of the study participants. The majority of respondents indicated they were 50 years of age or more.

At the time this study was conducted, the current state of Florida retirement system allowed individuals to retire from employment within the state system and re-apply for hire after a brief period of time. The county-wide public school districts employing respondents of this study were part of the state of Florida retirement system. A system functioning in this manner may promote an older workforce such as the one found in this study. Workforces in many disciplines and entities outside of the state of Florida may have a similar workforce age due to the large number of baby boomers present in the workforce (Hewlett, 2009). In addition to an aged workforce in general, the downturn of
the U.S. economy and the economy in the state of Florida may have caused some individuals to maintain employment rather than retire (U.S. Department of Labor, 2008).

School districts are typically known for employing teachers, who over time, may change roles from working in classrooms as teachers to working in district offices in administrative roles such as program specialists, program administrators, or other non-instructional positions. A small portion of the respondents (6%) indicated teaching as another part of their job. Respondents indicated they had been conducting evaluations for many years, with 67% reporting seven or more years of experience conducting evaluations. Due to the majority of respondents indicating an age of 50 or more, future studies should include ranges beyond 50 to collect detailed information on the aged workforce population.

**Findings from Research Question One.** Results from highest degree held consisted of 37 percent holding a doctorate and 46 percent holding a master’s degree. These results were very similar to Christie’s (2001) study of evaluators from the California Healthy Start program where 31 percent of respondents held a doctorate and 51 percent held a master’s degree. Christie’s (2001) survey did not provide a space for a specialist degree which may mean some respondents selected a master’s degree for the highest degree held if they actually held an education specialist degree. Educational specialist degrees are a newer degree, in some areas of the United States, representing an intermediate graduate degree consisting of work after a master’s degree (U.S. Department of Education, 2008).

Teaching/learning was the most common area of highest degree held by respondents. Considering the respondents were all employed by school districts, holding
degrees in areas of education or an education related field such as school administration was expected. For the most part people who study educational leadership were teachers first (Hancock, Black, & Bird, 2006; Grogan & Andrew, 2002). Some of the respondents held their highest degrees in fields outside of education which was not surprising since schools employ a variety of different non-instructional employees such as school psychologists, computer scientists, and social workers. In order to keep the number of survey items at a minimum, additional items such as all of the degrees held by respondents were not asked. Some individuals may hold more than one master’s degree or have completed all of the course work for a doctorate but have not completed the dissertation (ABD).

Aside from the educational specialist degree, education was identified as the most popular content area for highest degree held. Respondents who indicated an educational specialist degree was their highest degree reported school administration (67%) and evaluation/research methods (33%) as the area of the degree. This is not surprising considering an educational specialist degree focuses on an expertise in an area within education. Other degrees are often in education, even some with education in the title of the degree such as a Bachelors of Education (B.Ed.), Master of Education (M.Ed.), or Doctor of Education (Ed.D.). An Educational Specialist (Ed.S.) degree implies a specialization within an area of education.

The majority of doctorates were aligned by respondents as doctorates in education. Respondents were not asked whether their doctorate was a Ph.D. or Ed.D. due to the variations in program requirements among and within doctorate programs at doctoral granting institutions. Traditionally Ph.D. programs contain more research
preparation than an Ed.D. program, which is considered by some as a practitioner doctorate. However an Ed.D. program at one school may require more research courses and research experiences than a Ph.D. program at another school. Overall, little difference can be found between the two degrees (Carnegie Foundation, 2003). In this study those who did not hold a doctorate in teaching/learning, held doctorates in evaluation/research methods (21%), school administration (11%), and advanced quantitative studies (11%).

Areas of study may be directly related to the graduate programs offered at local colleges and universities. Particularly at the advanced graduate level, individuals working in schools often pursue graduate degrees while working full-time (Eisenhart & DeHann, 2005). According to the National Science Foundation (2009), the median age of doctorate recipients in the field of education was 41.5 years of age. Considering the median age at the time a doctorate was earned, people most likely attended a doctoral program that was accessible to them and fit into their work and person lifestyle. Programs offered at local colleges or universities or online programs may play an important role in the subject area or courses completed in graduate degrees earned. Alternatively, some graduate programs offer degrees in broad areas of education such as curriculum and instruction, and require students to select a concentration. Students who are enrolled in curriculum and instruction programs with a concentration may have selected the area of the concentration as the area in which their highest degree aligns or identified with a broader term like education.

**Extent of Training.** The extent of training in areas related to evaluation produced interesting results. Measurement was identified as the area in which the greatest number
of respondents completed one or more full courses. Due to the majority of respondents holding their highest degree in the field of education, many of the respondents were probably trained as teachers or school counselors. A course in measurement is a common required course in teacher preparation programs to meet required criteria for state certification in teaching. Standards encompassing assessment and accountability are currently required by the state of Florida for initial teacher preparation and educational leadership preparation as outlined in the Florida Educator Accomplished Practices (1999) and the Florida Educational Leadership Practices (2005). Florida’s standards correspond with the national standards set forth by the National Policy Board for Educational Administration (2007).

Interestingly, this finding may relate to findings from Hines and associates’ (2007) study of doctoral preparation in education. Hines and associates (2007) investigated the research preparation of doctoral students in education fields at research universities. The results of the national survey of doctoral programs in the field of education found measurement was the least common area among required courses for doctorate programs. The lack of required measurement courses in doctoral programs may be a result of coverage in undergraduate and graduate teacher preparation and educational leadership programs. Measurement courses centered on classroom assessment commonly completed by teachers tend to focus on different aspects of measurement when compared to more advanced measurement courses which could be offered at the graduate level.

In Hines and associates’ (2007) study, faculty who were coordinators or chairs of doctoral programs reported a lack of qualified faculty members and resources to teach courses in measurement and other advanced research areas at the doctoral level. If
students are bound to a particular program due to geographic location and offerings, or university-based programs have limited course offerings, they may be limited in the type and amount of training available in evaluation and research. In the evaluation field these issues have brought about week long training sessions such as the Summer Evaluation Training Institute sponsored by the American Evaluation Association and the Center for Disease Control, and the Evaluator’s Institute housed at George Washington University. Claremont Graduate University has started to offer fully online professional development workshops in addition to an on-campus series which is located at the Claremont campus in California. Individuals interested in pursuing additional training in evaluations areas have these options if they have the funds to enroll and in most cases a large budget to travel. The online workshops represent the newest training options for those with limited time and money. In the future, additional low cost and widely accessible training options may be available to increase the skills set of those conducting evaluations, particularly in the public sector.

Training in evaluation theory ranked third in the one or more full courses category \( (n = 44) \), and also program evaluation closely followed in fourth \( (n = 40) \). Overall the extent of training reported by participants in evaluation theory and program evaluation represented a wide variety of areas. Interestingly, program evaluation had the greatest number of responses for independent learning \( (n = 30) \). A self directed approach to program evaluation may be common for school district evaluators. Evaluation is known as a pragmatic discipline (Stufflebeam & Shinkfield, 2007) in which logical decisions and resources often guide the study. This type of approach often occurs when a content expert is asked to evaluate a program without prior experience conducting evaluations. In
cases such as these, the expert possesses the knowledge and skills related to the content area but needs to use outside resources to guide him or her through the evaluation process.

The question of whether to hire a person who is trained in the content area, which in this case is education, or a person who has an expertise in evaluation, remains unanswered. An educator who was also trained in evaluation and research methods would be the ideal candidate to conduct evaluations in school districts and vice versa. Based on the findings of this study, training in education was widely covered and training in evaluation was clearly evident, but additional training in evaluation for those who conduct evaluations would be beneficial to all those with stakes in the public schools. According to Engle et al. (2006) many graduate programs with a concentration in evaluation lack course offerings in specialized advanced topics related to evaluation practice. “Given the increase need for evaluation at the federal, state, and local levels for accountability purposes, evaluators’ preparation may be incomplete without these courses” (Engle et al., 2006, p. 359). The lack of offerings in evaluation programs may be problematic even for those with university-based evaluation training.

Unlike states such as Michigan and California, the state of Florida does not have its own evaluation association under the national American Evaluation Association. Evaluators in Florida are limited to participating in the Southeast Evaluation Association or the American Evaluation Association. The lack of a state evaluation association may indicate the state overall is not a leader in evaluation.

**Findings from Research Question Two.** Based on the instrument development and findings from Christie’s (2001) study, method, values, and use were proposed to
describe the practice of evaluators in this study. After conducting a principal component analysis, the original three dimensions did not provide useful information regarding the description of evaluation for the respondents in this study. There appeared to be underlying dimensions within the method, values, and use dimensions. Method, values, and use dimensions were used as a starting point, then broken down into two dimensions each, resulting in six dimensions: Holistic, Mixed Method Decision Making, Procedures Valued, People Valued, Users Engaged/Embodied, Evaluator as Mediator.

In Christie’s (2001) study, two dimensions were identified which cut across the method, values, and use dimensions: scope of stakeholder involvement and method proclivity. Her sample consisted of internal and external evaluators from the California Healthy Start program. When she compared the practices of the evaluators separately, focusing only on the internal evaluators, the method, values, and use dimensions explained the practices of the internal evaluators better than stakeholder involvement and method proclivity.

The results obtained from the principal component analyses raise questions regarding the recommended dimensions to use to describe the practice of evaluators in general. Alkin’s (2011) recently released book refers to “three general prototypes of evaluation orientations” (p. 35): methods-oriented approaches, values-oriented approaches, and use-oriented approaches. The results of the current study used the method, values, and use dimensions and found them to be too general for use to describe the practice of evaluation in K-12 public schools. The method, values, and use dimensions were not disregarded in the current study, but used as a starting point to further define dimensions within them.
Results from the qualitative data implied evaluations were carried out to meet the information needs of funding agencies and governing bodies. Relying solely on the information funding agencies and governing bodies need can cause valuable information to be missed. For example, only collecting data for randomized controlled trials (RCT) eliminates the collection of qualitative data which can provide fruitful information as to why a program was working or not working.

The practices reported in the theory to practice section of the survey may not represent the way respondents would carry out the evaluation if s/he were able to choose. Respondents were asked to select a program evaluation conducted in the last five years and base their answers on the way they carried out the evaluation selected. If the evaluation was bound to administrative and budget constraints, the background and training of the person carrying out the evaluation may not play a role in the way the evaluation was carried out. The results do represent the way evaluations were carried out in school districts, but the actions may not be a result of the background or training of the individuals who conducted the evaluation.

Evaluations often conclude with a final report. The final reports are not always used to make changes or improvements to a program. In the case of school districts, internal evaluations do not appear to be conducted unless a problem is identified or specific information is needed for a critical decision. One of the interview participants indicated a lack of money and resources prevented evaluations from being conducted for the “good of the order”. If evaluations in school districts are conducted when information is needed, then one of the reasons the evaluation results were used was because the evaluations were conducted solely to obtain that information.
Findings from Research Question Three. Reflecting back on the Alkin and House (1992) taxonomy, the responses from the school district evaluators were summarized using their three categories: method, values, and use. Based on the findings from the principal component analysis, the Alkin and House (1992) taxonomy provided the starting point in which additional dimensions were found to summarize evaluation practice and make comparisons among the respondents. Evaluation practice was summarized using the six dimensions found in this study in terms of Holistic, Mixed Method Decision Making, Procedures Valued, People Valued, Users Engaged/Embodied, Evaluator as Mediator. Similar to dimension used to display Alkin and House’s (1992) taxonomy (see Figure 1), each dimension can be placed on the continuum to visually describe the reported evaluation practice by respondents. The dimensions are presented with continuums in the order of the research question posed. Placement on the continuum was based on the mean group within each dimension relative to the total possible score. A percentage was calculated for each group using the group mean divided by the total possible sum for the scale.

Respondents were placed along a continuum shown in Figure 13 to visually display participants’ responses based on highest degree obtained. The holistic and procedures valued dimensions each revealed significantly higher means for respondents with a bachelor’s degree than respondents holding a doctorate. In general, those with a bachelors degree had greater means than those with a master’s or doctorate degree. The mean for respondents with doctorates were typically the lowest mean of the three categories. In general people tend to respond to surveys with a positive bias in self-reports (Jonathan, Kim, & Salleh, 2009). The term positive bias was used to indicate
participants responded in a way to make them looked better or do better. In general, research participants want to provide a socially desired response to questions posed to make themselves look good (Donaldson & Grant-Vallone, 2002). Similarly, in the current study, the three MANOVAs revealed group means aligning with agreement to the statements presented in the majority of the groups investigated. The respondents may not have responded with a socially desirable response, however, in general respondents selected ratings of agreement for all areas.

![Continuum](image)

*Figure 13. Placement on Continuum for Highest Degree Held. B = respondents with bachelors degree, M = respondents with master’s degree, D = respondents with doctorate. The left side of the continuum represents 100% and the right side represents 0%. Placement on the continuum was based on the individual group means relative to the total possible score within each dimension. A percentage was calculated for each group using the group mean divided by the total possible sum for the scale and multiplied by 100.*
Area of highest degree alignment yielded many significant differences. Respondents whose highest degree aligned with teaching/learning rated items on the theory to practice portion of the survey higher than those whose highest degree was in evaluation/research methods and school administration. The only exception was the holistic dimension where the mean for respondents with their highest degree in evaluation/research methods was 0.6 greater than those with their highest degree in teaching/learning.

<table>
<thead>
<tr>
<th>Continuum</th>
<th>Holistic</th>
<th>Mixed Method Decision Making</th>
<th>Emphasis on Procedures</th>
<th>Emphasis on People</th>
<th>Users Engaged/Embodied</th>
<th>Evaluator as Mediator</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>E</td>
<td>A</td>
<td>TE A</td>
<td>TE A</td>
<td>TE A</td>
<td>TE A</td>
</tr>
<tr>
<td>Less Holistic</td>
<td>Less focus on Mixed Method Decision Making</td>
<td>Less Emphasis on Procedures</td>
<td>Less Emphasis on People</td>
<td>Users Separate</td>
<td>Evaluator Inter-woven</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 14. Placement on continuum for alignment of highest degree. T = highest degree is teaching/learning, E = highest degree evaluation/research methods, A = highest degree in school administration. The left side of the continuum represents 100% and the right side represents 0%. Placement on the continuum was based on the individual group means relative to the total possible score within each dimension. A percentage was calculated for each group using the group mean divided by the total possible sum for the scale and multiplied by 100.**
The relationship between training in evaluation and evaluation practice did not yield any significant differences among respondents with minimal or novice training and those with intermediate to advanced training. Although there were no significant differences between the two groups, respondents with intermediate to advanced training had slight greater means in the four of the six dimensions.

Figure 15. Placement on continuum for extent of evaluation training. M = minimal to novice training, A = intermediate to advanced training. The left side of the continuum represents 100% and the right side represents 0%. Placement on the continuum was based on the individual group means relative to the total possible score within each dimension. A percentage was calculated for each group using the group mean divided by the total possible sum for the scale and multiplied by 100.

Respondents indicated there was a strong use of grant proposal guidelines and already established criteria to follow when reporting evaluation results. The policies set
forth by granting agencies may influence different values (procedures versus people) much more than the preferences of individual evaluators. For example, a review of request for proposals (RFP) from the Institute for Education Sciences (IES), U.S. Department of Education, and the National Science Foundation (NSF), examples of three agencies which grant awards in the field of education, each ask for multiple measures to be assessed. An evaluation plan submitted with a proposal that does not state quantitative and qualitative data will be used as part of the evaluation may not be awarded the grant and with no program, no evaluation will be conducted. If school district evaluators are conducting evaluations to answer questions posed by outside agencies, then outside policy strongly influences the choices made by evaluators in regards to values.

Qualitative methods are often more time consuming than quantitative methods. Money has been reported by respondents as an influence in their school district evaluations. One respondent (SES provider interview) stated “We don’t have money to hire external evaluators,” unless the money is coming directly from grant funds. One person reported that a lack of money prevents the district from conducting evaluations on programs which may benefit from exploration. Interviews and focus groups often require transcription of results which is time consuming and expensive to outsourced.

The school district evaluators appeared to demonstrate sensitivity to their allotted budget when making decisions about values. This finding was similar to Barela’s (2005) study of school district evaluators in California. Barela (2005) conducted a case study of individuals working in the evaluation branch of a large urban school district. He created evaluation prototypes to describe the way a junior evaluator and a senior evaluator typically carried out evaluation in the selected school district. Barela’s study (2005)
focused on individuals who are specifically conducting evaluation as their primary job. Similar to his findings, when respondents were considering the type and amount of data to collect, budgetary considerations were made. The amount of data to collect and the resources available may play a large role in the number of indicators used to judge the merit or worth of a program.

Results indicated a very “decision-oriented” approach to evaluation (Stufflebeam, 2001). Information was collected to make decisions about the programs under examination. Although participants were not asked if they followed a specific evaluation model or approach in the survey, the results reflect similarities with the practice of Daniel Stufflebeam. Stufflebeam is one of the most well-known evaluators particularly in the field of evaluation. Stufflebeam’s CIPP model (2001) was mentioned by one of the evaluators interviewed in this study as a resource used to guide his evaluation practice. Under Stufflebeam’s evaluation approach, a group of stakeholders are identified and engaged throughout the evaluation process, and results are provided to the stakeholders for direct use (Stufflebeam, 2007).

Overall, the majority of the respondents conduct evaluation as one of their job responsibilities. Self reported use of the evaluation results and findings found in this study were greater than those found in studies conducted among evaluators who conduct evaluation as a main part of their job (Barela, 2005; Christie, 2001; Fleischer & Christie, 2009). Evaluation results may be more valuable to people when they can play an active role in the evaluation process and sometimes act as the main evaluator. This was evident in the “Users Engaged/Embodied” dimension. Engaging stakeholders in the evaluation process is not a new idea and is highlighted as a main component of the participatory and
collaborative evaluation approaches. Under the participatory and collaborative evaluation approaches, stakeholders are actively engaged and involved in the evaluation process (Cousins & Earl, 1999; Rodriguez-Campos, 2005). Supporters of the collaborative evaluation approach often use the approach to increase the use of the evaluation results. Evaluations reported were conducted to learn more about programs and use the results for decision making.

Interestingly, no significant differences were found between respondents with Minimal or No Training and Intermediate to Advanced Training. Engle and associates (2006) found university-based programs of evaluation to have limited offerings in advanced evaluation areas. Considering the differences found in the second MANOVA in which the highest degree alignment was the independent variable, perhaps the content related methods and knowledge defines the different decisions made when conducting evaluations. In Kundin’s (2008) study, seven out of eleven experienced evaluators reported no formal training in evaluation. Interviewees stated they applied knowledge gained in general research methods courses in their evaluation work (Kundin, 2008). Results from Kundin’s (2008) study suggest evaluators rely on practical reasoning to guide them through evaluations. The differences found in this study may be related to the way practical reasoning is approached as a result of the highest degree held and area of highest degree and should be considered for future research.

Limitations

There were several limitations to the current study. The first limitation of this study is the sample size. A total of 134 people responded to the survey providing a response rate of 17%. While the sample size was large enough to conduct the proposed
analyses, a larger sample size was desired to have more precise estimates. One possible reason for the low response rate could have been related to the broad public record law in the state of Florida. The participants of the study were all employed by public school districts which is a government public entity operating under the “Sunshine Law”. Florida has a very broad public record law commonly referred to as the sunshine law because the law allows for everything to be “out under the sunshine”. The public record law allows for any documentable communication such as emails to be public record and accessible by the public and media upon request. Although participants were assured information collected via the electronic survey would be kept anonymous, many people may have been hesitant to respond to anything asking for information pertaining to any part of their job. Some participants sent emails explicitly stating this. The idea of the sunshine law would lead one to believe a wealth of information is available and easily accessible.

After conducting this study, my perception of the sunshine law has changed. After sending out the invitation survey to potential participants some response emails were received stating information similar to this example, “I will not be able to participate in your study. I am not able to answer any questions related to my job to outside persons.” Some school district personnel appeared to exhibit caution when asked about their job. The law appears to have created a climate of extreme caution among government employees who are often in the public eye. The government’s effort to provide an open-door policy on the happenings of their state government may have done the opposite in some areas. A past study surveyed art teachers in a northern state and was carried out in a similar manner to this one, and was able to obtain a 95% response rate and no
respondents indicated any feeling of threats to their job for participating (Hibbard, 2009). The topic of the survey was on knowledge and attitude towards educational research, and information on whether the teachers stay current in the field.

Due to the various different climates in each Florida school district, the results of this study may not be well distributed among the 67 school districts. The distribution of school districts is unknown to allow participants to remain anonymous. As a result, within and between-group comparisons could not be made because individuals were not asked to provide the name of their school district. Asking participants to provide the name of their school district would have likely resulted in an even smaller sample size because of the political climate surrounding school districts in the state, or a lot of missing data for the items. Based on experience conducting this study, individuals may be uncomfortable documenting information related to their job and work actions.

The use of self-reported measures is the second limitation to the study. Participants in research tend to provide socially acceptable responses when providing self-rating measures (Donaldson & Grant-Vallone, 2002). Although respondents were instructed to answer the survey items based on the way they actually carried out evaluation, the responses may not represent the real actions. The responses also represent respondents’ perceptions of the way they carried out evaluation, or the perceptions of what the respondent thinks is the desirable answer to each survey item.

A third limitation to the study is the length of the survey. The survey contained a total of 66 selected-response items and three open-ended items. Attempts were made to keep the survey items to a minimum amount in order to answer the research questions. Additional items regarding extensive information on the all types of degrees earned,
higher education institutions attended, nature of specific courses completed, and
description of previous work experiences would add a wealth of information on the
preparation of respondents. The items would also drastically increase the length of the
survey.

A fourth limitation to the study was the possibility of unknown evaluation
frameworks followed when conducting evaluations. Some respondents may have been
required to follow a specific framework when conducting their evaluation. In cases where
the evaluation was conducted as part of a grant or meeting requirements for an
accrediting body, the granting or governing body rules for the evaluation may explain
why the evaluation was carried out a specific way. Following a specific evaluation policy
could explain similarities in evaluation practice. Respondents were also able to select the
evaluation s/he thought was representative of the typical way s/he conducts evaluation.
Selecting one type of educational program like voluntary pre-kindergarten (VPK) and
surveying evaluators based on one particular type of evaluation would have provided an
additional context. However each district uses different ways to meet the evaluation
reporting requirements and a sample consisting of primarily internal evaluators would
have been difficult to obtain.

This study focused on the work of internal evaluators. Many large school districts
may contract more external evaluators in one year, than a small district utilizing internal
evaluators. Although external evaluators were not the focus of this study they may play a
large role in evaluative decisions in some school districts.
Study Contributions

This study investigated evaluation context and evaluation activities from Mark’s (2008) taxonomy of subjects of inquiry for research on evaluation. The study investigated the people who conduct evaluations in school districts and provided a descriptive overview of the background, training, and evaluation practice of school district evaluators which falls into the descriptive mode of inquiry (Mark, 2008). Providing information about internal evaluators is one of the main contributions of this study. The study is unique in that the data were collected primarily from people who may not consider themselves “evaluators” but regularly conduct evaluation in school districts as part of their job. In both Shadish and Epstein’s (1987) and Christie’s (2001) studies, the individuals they surveyed were identified as evaluators by their title or self identification by membership in a professional evaluation organization. The majority of individuals surveyed in this study did not hold the title of “evaluator” but conduct evaluation as a part of their job in public schools.

This study represents a broad view of evaluations in school districts and the people who conduct them. Trends in evaluation practice based on evaluator background and preparation were presented. Descriptive studies such as this one help outline the field of evaluation and provide a glimpse of the people who are part of the field. Many of the people conducting evaluation did not have formal training in program evaluation or program theory. Lack of available resources may be a reason for the gap in training. Based on the background information presented in this study, many individuals who work in schools will attend graduate school while working full-time. If training in evaluation is not available locally or online, and at an affordable price, it is likely for
individuals to continue to “commit” evaluations (Datta, 2003) without proper training. Attempts to provide affordable and widely accessible resources in program evaluation and evaluation theory could increase evaluation knowledge among individuals who conduct evaluation as part of their job responsibilities.

Implications for Future Research

The current study provided a look into evaluation practice in public schools in the state of Florida. Four specific areas have been identified for future research. First, an in-depth qualitative study observing the way evaluation is carried out on a day to day basis in school districts. An in-depth study would investigate explanations offered in this chapter and provide additional information on the climate and culture of evaluation in school districts from an emic perspective. In addition, an expansion on the current study to include a large geographic area would add to the findings, and differences among states could be compared.

The second area of research is related to evaluation policy. Participants in this study referred to meeting set criteria set forth by a source outside of the school district such as an accrediting body or grant agency. Further research on the criteria required from governing bodies such as the U.S. Department of Education, Florida Department of Education, National Science Foundation, Institute for Education Sciences (IES), among many others to compare the type of data requested and desired methods is needed.

The third area for future research is to investigate the outcomes and impacts of using certain evaluation methods. Considerations such as the requirements set forth by grant agencies and government bodies should be included as part of the study. In general, more studies are needed to explore the way evaluation is carried out under different
evaluation approaches, as well as the impact of using certain evaluation approaches on
the program and participants.

The fourth area for future study includes external evaluators for school districts. External evaluators for school districts were not the focus of the study but could provide another view of the way evaluation is carried out. A future study comparing evaluator practices when conducting evaluations under the same evaluation framework is needed and/or the same type of program is needed. Similarly a study could investigate the different ways evaluation is carried out at multiple program sites within one district.

**Overall Summary**

School district evaluators in the state of Florida were identified in this study. The information collected in this study can be used as a basis for future studies on evaluation practice in education settings. The description of those conducting program evaluations, and the types of programs under evaluation offer a snapshot of the field, and a platform to build a stronger empirical base for the field of evaluation.
References


Appendices
Appendix A: Letter to Participants

Dear XXXX,

By way of this e-mail I am requesting your assistance with a research study for my doctoral dissertation. I am a doctoral candidate in the department of Educational Measurement & Research at the University of South Florida, Tampa. During the week of February X\textsuperscript{th} 2010, I plan to administer an electronic survey to the individuals who have taken part in a school or district program evaluation. The survey is designed to obtain information from these program coordinators, directors or evaluators on the practice of evaluation in education settings.

My dissertation work intends to examine the influence of evaluation theory on practice. The aim is to develop a critical understanding of which evaluation theories, and particular components of theories, are most prevalent in practice. This will provide the insight necessary for developing and refining prescriptive evaluation theories for practice use. Through my study, I propose to yield a descriptive picture of prescriptive evaluation theories in application. This will provide an insight into the actual practice of evaluation. This study promises to make a significant contribution to the field of evaluation. Your participation is an essential part of this work.

You were identified as a person who may conduct evaluation as part of your job responsibilities. I am asking your kind participation in this study and hope that you will take a few minutes to complete the survey when you receive it. If by chance, you have not played an evaluative role in education programs, please indicate that in the first question of the survey.

I thank you for your consideration and hope you agree to participate. If you have any questions about the survey or project, please feel free to contact me (email: sthibbar@mail.usf.edu, phone: (239) 590-7808).

Thank you in advance for your assistance in this study.

Respectfully,

Susan T. Hibbard
Appendix B: Survey Questionnaire

Evaluation in Education Settings, February

1. Introduction

This is research being conducted as part of my doctoral dissertation and I am asking for your participation in my study. As part of my work, I am interested in how you are conducting evaluation. The information collected in this survey is completely confidential and will not be used to judge your job performance and/or any current program evaluation in any way. Individuals, schools, or districts will not be identified and only aggregated data will be reported in my study.

Participation in this survey is completely voluntary and you may choose not to participate or may stop participation at any time without penalty.

Thank you in advance for assisting me with my efforts to learn more about evaluation practice and theory.

2. Background Information

1. Have you evaluated a program in an education setting in the past five years?
   - [ ] Yes
   - [ ] No

3.

2. Gender:
   - [ ] Female
   - [ ] Male

3. Race Ethnicity:
   - [ ] Black
   - [ ] Hispanic
   - [ ] Asian/Pacific Islander
   - [ ] White
   - [ ] Mixed race/ethnicity
   - [ ] Other
4. Age:
- Less than 25
- 25 to 29
- 30 to 34
- 35 to 39
- 40 to 44
- 45 to 49
- 50 +

5. Your highest degree obtained:
- High School/Some college
- Bachelors
- Masters
- Specialist (Ed.S.)
- Doctorate

6. Your highest degree most closely aligns with:
- Anthropology
- Art/Music
- Advanced Qualitative Methods
- Advanced Quantitative Methods
- Business
- Economics
- Education
- Evaluation/Research Methods
- Public Health
- Psychology
- Public Policy
- Social Welfare
- School Administration
- Sociology

Other (please specify)
Evaluation in Education Settings, February

7. Please indicate the amount of training you completed in the following areas, read the descriptions below before selecting your answers (Check all that apply):

**One or more full courses:** A course on the topic taken through an institution (college, university, or online) equivalent to a full semester course.

**Substantial part of a full course:** Topic was a main part of a full course.

**One or more workshops:** The topic was the focus of the workshop(s).

**Substantial part of a workshop:** The topic was a main part of a workshop.

**Small part of a course or workshop:** Topic was briefly covered in a course or workshop.

**Staff Development (In-service):** District or State offered staff development on the topic.

**Independent Learning:** You learned the topic independently via self exploration of the topic.

<table>
<thead>
<tr>
<th></th>
<th>One or more full courses</th>
<th>Substantial part of a full course</th>
<th>One or more workshops</th>
<th>Substantial part of a workshop</th>
<th>Small part of a course or workshop</th>
<th>Staff Development</th>
<th>Independent Learning</th>
<th>Minimal exposure or no formal training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program Evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personnel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logic Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cash</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benefit/Cost-Effectiveness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Needs Assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey Research</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focus Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation in Education Settings, February

8. Did you complete a practicum/internship in evaluation when you were a college student?

<table>
<thead>
<tr>
<th>Undergraduate</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Evaluation Involvement and Experience

9. Conducting evaluation is:

☐ my full-time job.
☐ part of my job responsibilities.

10. What percent of your job do you spend conducting evaluations?

☐ Percent conducting evaluation

11. What other work are you involved in (Check all that apply)?

☐ Program administrator
☐ Teacher
☐ Social worker
☐ University faculty/staff
☐ School administrator
☐ Other

12. How many years have you been conducting evaluations?

☐ Less than 1
☐ 1 to 2
☐ 4 to 6
☐ 7 to 9
☐ 10 or more
13. How would you rate your current evaluation knowledge and skills?
- [ ] Excellent
- [ ] Good
- [ ] Average
- [ ] Minimal

14. You typically conduct evaluation as an (Check all that apply):
- [ ] Internal evaluator
- [ ] External evaluator

15. Of which of the programs listed below do you typically conduct program evaluations:

<table>
<thead>
<tr>
<th>Program</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curriculm/Curriculum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early Childhood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English Language Learners - ELL/ESOL/LEP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dropout Prevention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gifted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnet Programs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staff Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substance Abuse Prevention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supplemental Education Services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (please specify)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation in Education Settings, February

16. Think about an evaluation you conducted within the last five years that is typical of your evaluation practice. Indicate the type of program you evaluated. The remainder of your responses should be based on the program evaluation you select here and the way you carried out that evaluation. 

☐ Curriculur/Curriculum
☐ Early Childhood
☐ English Language Learners - EL/ESOL/LEP
☐ Dropout Prevention
☐ Gifted
☐ Magnet Programs
☐ Special Education
☐ Staff Development
☐ Substance Abuse Prevention
☐ Supplemental Education Services
☐ Other (please specify) _______________________________________________________________________

17. Briefly describe the context and purpose of the evaluation you conducted (this should be the same evaluation you identified in item 16 above).
**Evaluation in Education Settings, February**

16. Please use the scale provided to respond to each statement.

0 - Very Dissimilar (VD) to my evaluation practice  
3 - Dissimilar (D) to my evaluation practice  
7 - Similar (S) to my evaluation practice  
10 - Very Similar (VS) to my evaluation practice

**When conducting the evaluation, my evaluation approach:**

<table>
<thead>
<tr>
<th>Requirement</th>
<th>VD</th>
<th>D</th>
<th>S</th>
<th>VS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used qualitative (words) methods as my primary approach.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used quantitative (numbers) methods as my primary approach.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focused predominately on the observation of events (qualitative).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used the methods (quantitative, qualitative, experimental, naturalistic,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>etc.) necessary to conduct a useful, feasible, proper and accurate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evaluation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encouraged people at all levels of the organization to participate in the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evaluation process.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identified, engaged, and served stakeholders at all levels of a program.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Issued separate reports to serve the different needs of the various</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>audiences.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intended to create changes in the culture of the organization (district)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>where the evaluation was being conducted.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Involved helping the program staff and clients develop a plan for the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>future.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helped the program staff and clients establish their goal and document</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evidence of working towards their goal.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Included working with primary users (the main people using the evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>results) to make informed judgments about their program using the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evaluation data.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allowed me to use my own values to select what I believe to be</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>significant.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objectively assessed the evaluation's quality against validated technical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>standards.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Evaluation in Education Settings, February**

19. Please use the scale provided to respond to each statement.

0 - Very Dissimilar (VD) to my evaluation practice  
3 - Dissimilar (D) to my evaluation practice  
7 - Similar (S) to my evaluation practice  
10 - Very Similar (VS) to my evaluation practice

**When conducting my evaluation:**

<table>
<thead>
<tr>
<th>Requirement</th>
<th>VD</th>
<th>D</th>
<th>S</th>
<th>VS</th>
</tr>
</thead>
<tbody>
<tr>
<td>The evaluation questions were designed to yield information for making</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decisions about the program.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research method(s) were selected based on the program's conceptual</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>framework, model or theory.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear contracts of work requirements subjected to review by people both</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>internal and external to the organization were used.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stakeholders' assumptions about a program were integrated into the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evaluation process.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The evaluation became institutionalized and a part of future planning and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>operation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stakeholders received evidence of the merit of their program, and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>information about how to improve the program.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The primary focus of the evaluation design was to improve program</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>performance.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The evaluation was designed to foster self-determination,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>enlightenment, and liberation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Evaluation in Education Settings, February

#### 20. Please use the scale provided to respond to each statement.

0 - Very Dissimilar (VD) to my evaluation practice  
3 - Dissimilar (D) to my evaluation practice  
7 - Similar (S) to my evaluation practice  
10 - Very Similar (VS) to my evaluation practice

<table>
<thead>
<tr>
<th>When conducting my evaluation:</th>
<th>VD</th>
<th>D</th>
<th>S</th>
<th>VS</th>
</tr>
</thead>
<tbody>
<tr>
<td>The main evaluations questions were answered using scientifically tested instruments.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>The primary users (the main people using the evaluation results) helped conceptualize and determine the evaluation questions.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Adjustments and changes were be made when parts of the current evaluation plan were not working.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Stakeholders participated in conducting the evaluation.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>I observed what was transpiring in the program and then interpreted and judged its significance.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>I combined qualitative and quantitative methods to address each evaluation question.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>The primary users judged the appropriateness and quality of the evaluation's methodology.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>It was most important to ensure the evaluation was scientifically sound.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>The primary users helped interpret the meaning of the evaluation data.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Including and reflecting the diverse perspectives of relevant stakeholders was important (not just what was required by the funding source).</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
</tbody>
</table>

### Evaluation in Education Settings, February

#### 21. Please use the scale provided to respond to each statement.

0 - Strongly Disagree (SD)  
3 - Disagree (D)  
7 - Agree (A)  
10 - Strongly Agree (SA)

<table>
<thead>
<tr>
<th>In order for evaluations to be most effective stakeholders must be included, but don’t necessarily have to participate, in the evaluation process.</th>
<th>SD</th>
<th>D</th>
<th>A</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation conclusions are mixtures of facts and values.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Evaluator’s interpretation of the findings and conclusions can be unbiased.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>An evaluator is an applied scientist; not an advocate, counselor or policy advisor.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Evaluations are to be used to build upon the current generalized knowledge base of the particular program being studied.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>The ultimate purpose of program evaluation is to enhance knowledge for designing and implementing programs to improve human conditions.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>During the final stages of the evaluation, the evaluator is to work with the primary users to help determine the next steps.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Evaluation is a self-evaluative process for a district.</td>
<td>☐</td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
</tbody>
</table>

#### 22. To what extent do you think decisions you made during the evaluation were influenced by political issues surrounding the program and/or school district?
23. Please use the space below to provide any additional information you would like to share regarding your evaluation.

Thank you for your participation in my study. If you have any questions about the study, please contact me, Susan Hibbard at sthibbard@mail.usf.edu
Appendix C: Scree Plots

Method Dimension

Values Dimension
Use Dimension

Scree Plot

![Scree Plot Image]

Component Number

Eigenvalue
Appendix D: Interview Questions

**Hibbard:** Thank you for contacting me to talk about your program evaluation. I would like to ask you a few questions about study. Can you tell me a little bit about the evaluation you conducted?

**Hibbard:** Who specifically commissioned this particular evaluation? The grant agency requires an evaluation.

**Hibbard:** How and why were you selected to take part in this evaluation study?

**Hibbard:** What was your role in the study?

**Hibbard:** Is this the typical process?

**Hibbard:** Is the size and scope of the study typical of the kinds of studies conducted in your department/unit?

**Hibbard:** How often does the district solicit external evaluators?

**Hibbard:** What proportion of the district’s evaluations would you say are done by your department/unit?

**Hibbard:** Thanks, I think that gives us a nice context for understanding your work. Getting to some of the details about the ________ evaluation, who were the primary stakeholders for this study?

**Hibbard:** How were the ________ stakeholders involved in the study?

**Hibbard:** How often did you foster buy-in for your work?

**Hibbard:** How vested would you say the ________ stakeholders were in the evaluation?

**Hibbard:** How accountability focused was the study?

**Hibbard:** How was your process for deciding upon the evaluation questions?

**Hibbard:** If the questions were established prior to you stepping in, how, then, did you decide upon the approach you used to address these ___ (# of eval questions) questions?

**Hibbard:** Did you use a specific evaluation framework to guide your study?
Hibbard: How would you say the politics of this study differed from other evaluations you have conducted?

Hibbard: I’d like to know more about your methodological and analytic approach. What types of data collection methods were used?

Hibbard: Which data collection strategies, or combination of strategies, do you think yielded the most critical or powerful information about the ______ program and why?

Hibbard: What barriers did you encounter when collecting data and how were they resolved?

Hibbard: Do you have an IRB process for internal evaluations?

Hibbard: Did you engage in any specific activities to help facilitate the use of your findings?

Hibbard: Is this a typical process?

Hibbard: How did they use the report?

Hibbard: In retrospect, what would you have done differently to make the report more useful?

Hibbard: How do you think the study would have been different if it had been conducted by an external evaluation team?

Hibbard: Is there anything else you would have done differently?
About the Author

Susan T. Hibbard was born in Buffalo, NY and earned a B.A. in Chemistry from the University of Buffalo, M.S. in Education from Niagara University, and a Ph.D. from the University of South Florida in Measurement & Evaluation. She has worked at Florida Gulf Coast University since 2008 as a faculty member in Research and Evaluation in the College of Education. She teaches undergraduate and graduate courses in evaluation, research methods, and foundations. Her research focuses on evaluation theory and practice, empirical research on evaluation, and education programs. She has worked on research and evaluated programs for state and federal grants, including the Institute of Education Sciences, National Science Foundation, and the Association of Institutional Research. Her peers have recognized her professional qualifications through invitations to her to make presentations and workshops in the areas of evaluation, action research, and rubric writing. She is an active member of the American Evaluation Association, American Educational Research Association, Association of Teacher Educators, and Kappa Delta Pi an international honor society in the field of education. In addition, Susan served as an editorial assistant for the peer reviewed journal Educational Researcher, one of the top journals for educational research. She has presented at conferences such as the American Educational Research Association, Association of Teacher Educators, and Council for Exception Children, in the areas of evaluation, research, and teacher preparation.