12-2016

Three Empirical Investigations into the Logic of Evaluation and Valuing Practices

Satoshi Ozeki
Western Michigan University

HREE EMPIRICAL INVESTIGATIONS INTO THE LOGIC OF EVALUATION AND VALUING PRACTICES

by

Satoshi Ozeki

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation
Western Michigan University
December 2016

Doctoral Committee:

Chris L. S. Coryn, Ph.D., Chair
Daniela C. Schröter, Ph.D.
Tarek Azzam, Ph.D.

THREE EMPIRICAL INVESTIGATIONS INTO THE LOGIC OF EVALUATION AND VALUING PRACTICES

Satoshi Ozeki, Ph.D.

Western Michigan University, 2016

This three-paper dissertation investigates the logic of evaluation proposed by Michael Scriven using two different methods of inquiry: content analysis and a survey method. Scriven's logic of evaluation is one of the central concepts in evaluation, but has not been empirically investigated. Therefore, the purpose of this three-paper dissertation is to accumulate empirical knowledge on Scriven's logic of evaluation to investigate the extent to which evaluators follow his logic and their perceptions of and familiarity with his logic. Furthermore, this dissertation aims to explore valuing practices in terms of Scriven's logic of evaluation. Scriven's logic of evaluation is important because it provides a fundamental reasoning process of conducting evaluation. Whether consciously or not, if one is to conduct evaluation, Scriven's logic needs to be followed. Although there is no unanimous agreement on the extent to which Scriven's logic should be followed, Scriven's logic of evaluation is considered one of the most essential concepts in the field of evaluation.

This dissertation aims to empirically investigate different aspects of Scriven's logic of evaluation using two methods of inquiry. In the first study, content analysis of evaluation reports was conducted to examine whether the utilization of Scriven's logic of evaluation is clearly identifiable or not. Findings suggest that standards were not clearly established and synthesis methodology was not found in those evaluation reports. However, it remained unclear whether

the evaluators in the sample of the evaluation reports did not follow Scriven's logic of evaluation. The second study aimed to examine whether evaluators apply Scriven's logic of evaluation into their evaluation practice as well as evaluators' familiarity with and perceptions of Scriven's logic of evaluation. Findings from the second study suggest that many of the evaluators were not familiar with his logic but evaluators typically, although not always, followed Scriven's logic of evaluation. In addition, it was found that a large number of the respondents considered Scriven's logic of evaluation useful and important in conducting evaluation regardless of their familiarity. The third study sought to explore details of performance standards and evaluative conclusions. The findings from the third study suggest that many of the evaluators were engaged in making evaluative conclusions, but did not always use performance standards to reach evaluative conclusions. Most of them considered the task of making evaluative conclusions difficult yet important in their evaluation practice.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my dissertation committee members, Drs. Chris Coryn, Daniela Schröter, and Tarek Azzam. First, as my dissertation chair, Chris has been always supportive of my academic work and I have learned a lot about evaluation from working with him as a doctoral associate. I am particularly thankful to him for awarding a Doctoral Research Associateship, which enabled me to focus on my academic work. Second, I appreciate Daniela for her continuous inspiration for my dissertation, which has been developed while working for her project. Lastly, I am truly grateful for Dr. Azzam's willingness and commitment to be part of my dissertation. I really appreciate his constructive feedback on my dissertation.

I also would like to say thank you to Ms. Mary Ramlow for her consistent support and encouragement. My life at the Evaluation Center was a lot easier simply because of her presence. In addition to her administrative assistance, she helped me with everything while I was at the Evaluation Center. Furthermore, I would like to thank all the IDPE students and the Evaluation Center staff I met during my doctoral study. IDPE and the Evaluation Center were truly an amazing place to study and work. I sincerely enjoyed being part of them.

Satoshi Ozeki

TABLE OF CONTENTS

ACKNOWLEDGMENTS ii

LIST OF TABLES vii

CHAPTER

I. INTRODUCTION 1

Background of the Problem 1

Scriven's Logic of Evaluation and its Related Issues 3

Dissertation Format and Related Purposes of the Three Studies 5

Study One: The Logic of Evaluation in Professional Development Evaluation Practice 6

Study Two: An Examination of Evaluators' Perceptions and Application of Scriven's Logic of Evaluation 8

Study Three: An Investigation of Evaluators' Valuing Practices 10

Significance of the Research 11

References 13

II. THE LOGIC OF EVALUATION IN PROFESSIONAL DEVELOPMENT EVALUATION PRACTICE: EVIDENT OR NOT SO MUCH? 16

Introduction 16

Purpose 19

Research Questions 20

Method 21

Sample 21

Code Development 22

Table of Contents—Continued

CHAPTER

Coding Procedure....23

Results....24

Sources and Types of Criteria....24

Standards....26

Designs and Methods to Measure Performance....29

Synthesis Methodology....30

Discussion....31

Limitations....34

Future Research....35

References....35

III. AN EXAMINATION OF EVALUATORS' PERCEPTIONS AND APPLICATION OF SCRIVEN'S LOGIC OF EVALUATION....38

Introduction....38

Shared Logical Process....40

Purpose and Research Questions....42

Method....43

Sample....43

Instrumentation....45

Results....45

Frequency and Difficulty of Performing Each Step of Scriven's Logic....45

Reasons for Difficulty....47

Evaluators' Familiarity with and Views on Scriven's Logic of Evaluation....49

Evaluators' Views on Scriven's Logic of Evaluation....50

Table of Contents—Continued

CHAPTER

Familiarity and Evaluators' Practices and Perceptions 52

Discussion 54

Limitations 56

Future Research 57

References 57

IV. AN INVESTIGATION OF EVALUATORS' VALUING PRACTICES 60

Introduction 60

Definition of Evaluation 61

Performance Standard Setting 63

Purpose 65

Research Questions 65

Method 66

Sample 66

Instrumentation 67

Results 69

Nature of Evaluative Conclusions 69

Evaluators' Perceptions about Making Evaluative Conclusions 70

Nature of Performance Standard Setting 72

Reasons for Changes in Performance Standards 73

Discussion 74

Limitations 76

Future Research 77

Table of Contents—Continued

CHAPTER

References....77

V. CONCLUSION....80

Review of Main Findings....80

Conclusions....81

Limitations....82

Future Research....84

References....85

APPENDIX

Human Subjects Institutional Review Board Approval Letters....87

LIST OF TABLES

2.1 Variables and Codes Identified by Content Analysis 23

2.2 Criteria and Major Sub-Criteria 26

2.3 Number and Percentage of Standards Coded.....27

2.4 Major Themes Identified as Performance Descriptors 28

2.5 Methods at Criteria Level 30

3.1 Demographics of the Sample 44

3.2 Frequency and Difficulty of Performing Each Step 46

3.3 Usefulness and Importance of Scriven's logic of evaluation.....50

3.4 Descriptive Statistics, Chi-square Test and Odds Ratio 53

4.1 Demographics of the Sample 68

4.2 Percentages of Respondents Conducting the Followings Tasks in Evaluation 70

4.3 Descriptive Statistics of Characteristics of Performance Standards 73

CHAPTER I

INTRODUCTION

Background of the Problem

Different evaluation theories and models have been developed to deal with differing evaluation demands (Stufflebeam & Coryn, 2015). Those theories and models offer practical guidance of how to conduct evaluation in various evaluation contexts (Alkin, 2004). Most evaluation theories and approaches were, however, not empirically investigated. Thus, to understand whether those evaluation theories inform evaluation practice and evaluation approaches were actually practiced, empirical investigation of evaluation theories and approaches is necessary (Christie, 2011; Miller, 2010; Smith, 1993). Without empirical evidence for the effectiveness of evaluation practice, it is difficult for consumers of evaluation to believe what evaluators do. To accumulate empirical knowledge on evaluation, research on evaluation has increased in the past decade (Vallin, Philippoff, Perce, & Brandon, 2015). Past empirical studies investigated various aspects of evaluation practice. Topics of such studies include, but are not limited to, courses on evaluation at universities (e.g., Davies & MacKay, 2014; LaVelle & Donaldson, 2010), evaluation use (e.g., Cousins & Leithwood, 1986; Fleischer & Christie, 2009; Johnson et al., 2009; Patton et al., 1977), methodological practice (e.g., Azzam, 2011; Azzam & Szanyi, 2011; Christie & Fleischer, 2011), theory-practice relationship (e.g., Christie, 2003; Coryn, Noakes, Westine, & Schröter, 2011; Miller & Campbell, 2006), and stakeholder involvement (e.g., Brandon & Fukunaga, 2013; Cullen, Coryn, & Rugh, 2011). It seems that the field of evaluation had benefited from the increased number of studies on evaluation. However, it

was unknown whether those studies successfully dealt with important issues of evaluation. Coryn et al. (under review) reviewed over 3,000 journal articles published in 14 evaluation-related journals over a 10-year period to examine whether past studies on evaluation can be classified into the two types of taxonomies of research on evaluation proposed by Henry and Mark (2003) and Mark (2008). They identified 257 articles on research on evaluation and found that as little as 3.5% of the articles examined issues of valuing. Valuing is defined as ways evaluators attached values to an evaluand (Shadish, Cook, & Leviton, 1991) or the process of making value judgments (Alkin, Vo, & Christie, 2012). Valuing is important because it is one of the key features that make evaluation unique from other types of inquiry (Fournier, 1995). Because evaluators and social scientists both depend on methodology from social science, valuing becomes an essential characteristic that differentiates evaluators from social scientists (Mathison, 2007). The importance of valuing in evaluation is also discussed by evaluation scholars. For instance, Shadish et al. (1991) includes valuing as one of the criteria for good evaluation theories. Furthermore, Christie and Alkin (2012) suggest the importance of valuing in evaluation in their evaluation tree. To make a value judgment, Scriven (1967) proposed a specific guidance on how to conduct evaluation for valuing. Scriven's logic of evaluation is considered an essential concept in evaluation because it provides a basic logical process for valuing practices (Shadish, 1998). Although his theory of valuing has been considered reasonable, his theory has not been under empirical investigations. Therefore, Scriven's theory needs to be empirically investigated to advance the field of evaluation. This proposed dissertation is intended to make a contribution to the field of evaluation by empirically investigating one of the central components that pertain to valuing, namely Scriven's logic of evaluation.

Scriven's Logic of Evaluation and its Related Issues

Scriven (1980, 1991) has discussed the logic of evaluation in his writings. The logic of evaluation is considered as the fundamental logical process pertaining to any evaluation practice and is described as the four steps of conducting evaluation: (1) establishing criteria, (2) setting standards, (3) measuring performance on the criteria relative to the standards, and (4) synthesizing the results into a value judgment (Fournier, 1995). The first step involves determining the aspects of an evaluand to be evaluated. The second step is concerned with setting up standards against which performance on the identified criteria will be compared. The third step is to measure performance relative to the identified criteria and compare it with the predetermined standards. The fourth step is to synthesize the findings into an overall evaluative conclusion about an evaluand. Shadish (1998) indicates that Scriven's logic of evaluation is very important, yet typically underappreciated. Therefore, evaluators may not intentionally employ Scriven's logic of evaluation. Although evaluation scholars indicate that Scriven's logic of evaluation should be followed (Shadish, 1998; Shadish et al., 1991), evaluators might not have any knowledge of the logic of evaluation, therefore not intentionally using it in their evaluation practice. One reason for the lack of recognition is that Scriven's logic of evaluation provides a reasoning process and evaluators may not be aware of the underlying reasoning process; therefore, evaluators might follow Scriven's logic of evaluation without knowing that they do so. Shadish et al. (1991) indicate that Scriven's logic of evaluation is always implicit in evaluation and rarely appreciated by evaluators. Thus, evaluators are unlikely to be aware of the logic when conducting evaluation. In addition, Fournier (1995) discusses that Scriven's logic of evaluation is a general logic of reasoning that underlies any evaluation practice. She explains that it is another logic, called working logic, that guides valuing practices of a particular evaluation approach.

Different types of working logic emerge based on the evaluation approach and evaluation context. Therefore, evaluators might be aware of the working logic because it is specific to the type of evaluation they are conducting. Evaluators are, however, unlikely to be aware of the general logic because it is the logic behind the working logic they use in their evaluation practice.

In addition, not all scholars agree on the entire concept of Scriven's logic of evaluation. For instance, there is no agreement on when and how standards of performances should be set, or even whether standards should be explicitly identified. Patton (2008) recommends setting standards before conducting evaluation. Weiss (1998) states that standards are explicit or implicit, indicating potential confusion about how to establish standards. Davidson (2005) advocates for setting clear standards by using a rubric. Therefore, evaluators are likely to engage in different approaches to establish standards. In addition, not all scholars agree on the last synthesis operation. For instance, Stake and colleagues (1997) claim that Scriven's logic of evaluation is not extensively practiced because valuing involves a more complex process of both intuition and conscious reasoning than simple use of a rubric for valuing. Shadish et al. (1991) also questioned Scriven's last synthesis operation after a detailed analysis of the logic of evaluation. In addition, there is a discussion on who should be involved in a valuing process in evaluation. Although Scriven (1986) argues that the evaluator should determine the value of an evaluand, other evaluation scholars have different views on the degree of evaluators' involvement in valuing. According to Alkin et al. (2012), valuing can be performed in three ways: evaluators only, stakeholders only, and the combination of evaluators and stakeholders. Therefore, evaluators could conduct valuing practices in different ways. In conclusion, Scriven's logic of evaluation seems to be a conceptually sound logic to follow in conducting evaluation.

However, it is unknown whether and how the logic of evaluation is practiced by evaluators. Therefore, the logic of evaluation, like other topics in evaluation, needs to be investigated further.

Dissertation Format and Related Purposes of the Three Studies

The proposed dissertation consists of three distinct studies, which together seek to investigate Scriven's logic of evaluation using different methods of inquiry. Chapter I introduces the broader context of the field of evaluation and the importance of Scriven's logic of evaluation, which led to this empirical investigation of Scriven's logic of evaluation and its related issues. The three papers appearing as Chapters II, III, and IV are briefly introduced individually in the following section. Chapter V summarizes the findings from the three papers and discusses implications and limitations of this dissertation work. Then, future directions for research in Scriven's logic of evaluation are provided.

Each paper is intended to examine different areas of Scriven's logic of evaluation using different methods of inquiry. The first study systematically content-analyzes evaluation reports to investigate whether the practice of Scriven's logic of evaluation can be identified. Based on the first study, the second study attempts to discover what is not identified in the first study; the second study seeks to investigate evaluators' perceptions and application of the logic of evaluation to their evaluation practice. Expanding on the second study, the third study examines further details of Scriven's logic of evaluation with its focus on performance standards and evaluative conclusions. Findings from this dissertation will provide empirical knowledge on evaluators' practice in terms of performance standards and evaluative conclusions. Three studies together empirically explore Scriven's logic of evaluation, an essential concept in evaluation. In the following, each study is described in greater detail along with the study's purpose, research

questions, methodology that will be used to answer the research questions, and contribution to evaluation.

Study One: The Logic of Evaluation in Professional Development Evaluation Practice

Purpose. This study aims to investigate to what extent the logic of evaluation can be explicated from teacher professional development evaluation reports. By conducting a content analysis of the reports, this study aims to determine whether and how the logic of evaluation was applied in the evaluation of professional development programs. To achieve this goal, this study examined the criteria, the standards, and how the findings and standards were integrated into evaluative conclusions.

Research questions. This study focused on research questions that examined the logic of evaluation documented in evaluation reports. As described in the purpose, this study is centered on the documented evaluation process that links evaluation questions to criteria, standards, findings, and conclusions. The focal question is to what extent Scriven's logic of evaluation is identifiable in evaluation reports. The specific questions examined are as follows:

1. Are evaluation criteria explicated?
    a. If so, what were the sources and the types of criteria?
2. Were standards of performance addressed?
    a. If so, what types of standards?
    b. If not, how was performance evaluated?
3. How was performance measured?
4. When conclusions are described, how are the conclusions reached?

Methodology. In the methodology that follows, the sample, instrumentation, procedure, and analytic approach are presented.

Sample. The sample of the evaluation reports included in this study come from a larger evaluation study sponsored by the National Science Foundation, which was designed to investigate evaluations of science-teaching professional development interventions for elementary school science. In this study, a multi-step sampling procedure was utilized to collect a sample of the evaluation products. First, an initial broad scan identified 734 evaluation documents related to elementary science professional development evaluation. Then, various criteria, including documents dealing with science professional development, having sufficient information for coding, addressing professional development in the United States, and being published since 2002, were applied, resulting in the final sample of 55 documents. The 55 documents include 33 peer-reviewed articles and 22 non peer-reviewed documents. Because the purpose of this study is to examine evaluation reports in terms of how evaluators practice the logic of evaluation, the 33 peer-reviewed articles were excluded from this study. In addition, out of those 22 non-peer-reviewed documents, three documents were excluded from the final sample; one of them was a newsletter and two of them were synthesis studies. Thus, the final sample of 19 evaluation reports was investigated in this study.

Instrumentation. A data abstraction form was created based on the research questions. The form includes various codes pertaining to the criteria, the designs and methods, the standards, and the conclusions.

Procedure. Using the coding structure form, two coders coded all the sample of products. Prior to coding, the two coders engaged in a calibration procedure, in which they coded a few products to become familiar with the codes and the coding procedure. This process helps to identify any issues and clarify any ambiguity (Wilson, 2009). Based on the results of the calibration procedure, the coding form could be modified, and once the coding form was

finalized, the two coders coded all studies independently from each other. After all the coding, interrater reliability was calculated, and then the coders resolved all coding disagreements to finalize the coding procedure.

Analytic approach. To answer the research questions, the coded data were analyzed through a content analysis and descriptive statistics. Frequencies were tabulated and presented.

Contribution to evaluation. Following the past studies that examined evaluation practice using content analysis (i.e., Coryn, Noakes, Westine, & Schröter, 2011; Miller & Campbell, 2006), this study aims to accumulate empirical knowledge on evaluation practice by systematically content-analyzing evaluation documents. Results from the study will shed light on the discussion of whether the logic of evaluation proposed by Scriven is practiced often (Stake et al., 1997).

Study Two: An Examination of Evaluators' Perceptions and Application of Scriven's Logic of Evaluation

Purpose. This study aims to examine evaluators' application of Scriven's logic of evaluation and its related questions. The purpose of this study is to investigate whether and to what extent evaluators use Scriven's logic of evaluation. In addition, this study will examine evaluators' views on the logic of evaluation.

Research questions. As described in the purpose, the focal question centers on evaluators' perceptions and applications of the logic of evaluation. Therefore, the specific questions examined are as follows:

1. To what extent do evaluators apply Scriven's logic of evaluation?
    a. How often do evaluators perform each step of the logic?
    b. How difficult it is to perform each step of the logic?
2. What are evaluators' views on Scriven's logic of evaluation?

a. How familiar are evaluators with his logic of evaluation?

b. How important do evaluators consider the logic?

c. How useful do evaluators consider the logic?

Methodology. In the methodology that follows, the sample, instrumentation, procedure, and analytic approach are presented.

Sample. The sample of this study was drawn from AEA members.

Instrumentation. A web-based survey was created on the Qualtrics survey system to answer the focal questions. The survey consisted mostly of closed-response and partially closed-response items. Open-response items were used to follow up with particular responses to closed-response items.

Procedure. Following an application procedure with and approval from the AEA Research Request Task Force, a list of all AEA members' email addresses was obtained. A random sample was drawn from the AEA member list and an invitation email to the survey was sent. After the initial invitation email, three reminder emails were sent once per week to those members who had not completed the survey.

Analytic approach. The results of the survey from the Qualtrics survey system were imported into SAS 9.3 for processing and analysis. Closed-response items were analyzed in SAS 9.3.

Contribution to evaluation. Like Study One, this study aims to contribute to the field of evaluation by accumulating empirical knowledge on the logic of evaluation. Specifically, this study explored evaluators' perceptions and applications of the logic of evaluation. Findings will provide insights into the extent to which the logic of evaluation is recognized and practiced.

Study Three: An Investigation of Evaluators' Valuing Practices

Purpose. Although it provides a general reasoning process for valuing, Scriven's logic of evaluation does not specify all the details for determining values. For instance, the second step of the logic indicates standard setting, but does not provide any guideline on how and when to establish standards. Thus, the purpose of this study is to examine further details of Scriven's logic of evaluation with its focus on performance standards and evaluative conclusions. Specifically, this study will examine how evaluators conducted the tasks of setting performance standards and reaching evaluative conclusions.

Research questions. As described in the purpose, the focal question centers on valuing practices in terms of the logic of evaluation. The focal question is how evaluators make evaluative conclusions. Therefore, the specific questions examined are as follows:

1. How do evaluators make evaluative conclusions?
    a. How often do they make evaluative conclusions?
    b. Who is responsible for making evaluative conclusions?
    c. What are challenges in making evaluative conclusions?
2. How do evaluators establish standards?
    a. Do they use performance standards for evaluative conclusions?
    b. Who are involved in establishing standards?
    c. When do they establish standards?

Methodology. In the methodology that follows, the sample, instrumentation, procedure, and analytic approach are presented.

Sample. The sample of this study was drawn from AEA members.

Instrumentation. A web-based survey was created on the Qualtrics survey system to answer the focal questions.

Procedure. Following an application procedure with and approval from the AEA Research Request Task Force, a list of all AEA members with the names and email addresses as well as background information was obtained. A random sample was drawn from the AEA member list. The members selected for Study Two were excluded from the population. Along with the initial invitation email, three reminder emails were sent once per week to those members who had not completed the survey.

Analytic approach. The results of the survey from the Qualtrics survey system were converted into tab-delimited files and then imported into SAS 9.3 for processing and analysis. Closed-response items were analyzed in SAS 9.3. Open-ended items were content-analyzed to find themes.

Contribution to evaluation. This study aims to reveal evaluators' valuing process in a simulated scenario. It will contribute to the field of evaluation by adding empirical knowledge on how evaluators would approach the valuing process. Although the valuing is a central component of evaluation (Scriven, 1986), the research in this area is scarce. Thus, this study will help to understand the valuing practice of evaluators through empirical investigation.

Significance of the Research

This dissertation study is significant in three important ways. First, this study responds to calls for more research on evaluation. As numerous evaluation scholars suggest, more empirical evidence for evaluation practice is required to advance the field of evaluation (Smith, 1993). This study can contribute to the field of evaluation by increasing the knowledge base. A stronger

knowledge base in evaluation can help to further establish the field of evaluation and an evaluation profession.

Second, this study investigates evaluation theory, which requires empirical investigation to strengthen the link between evaluation theory and practice. Evaluation theories are typically formulated by evaluation scholars' expertise and experiences without much empirical support for them. More research to examine whether evaluation theory empirically informs evaluation practice is beneficial because evaluation practice can be advanced with empirical evidence. This dissertation focuses on such a relationship between evaluation theory and practice. Using content analysis and survey method, this study attempts to have a better understanding of the relationship between Scriven's logic of evaluation and evaluators' practice.

Lastly, this study is concerned with valuing practices of evaluators. Although valuing is one of the central components of evaluation, valuing has not been extensively studied in the past. Coryn et al. (under review) indicate that only 4% of the studies on research on evaluation in the past decade examined values and/or valuing. The field of evaluation can benefit from more research on valuing because valuing is one essential feature that distinguishes evaluation from similar social science practice. Research on valuing will assist in advancing the field of evaluation as an established field of inquiry. This study, by focusing on Scriven's logic of evaluation and valuing practices, can make an important contribution to the field of evaluation. In the following chapters, each study is described in Chapters II, III, and IV, respectively. Then, overall conclusions and future research will be suggested to advance the field of evaluation with empirical investigations in Chapter V.

References

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Alkin, M. C., Vo, A. T., & Christie, C. A. (2012). The evaluator's role in valuing: Who and with whom. In G. Julnes (Ed.), Promoting valuation in the public interest: Informing policies for judging value in evaluation. *New Directions for Evaluation, 133,* 29-41.

Azzam, T. (2011). Evaluator characteristics and methodological choice. *American Journal of Evaluation, 32,* 376-391.

Azzam, T., & Szanyi, M. (2011). Designing evaluations: A study examining preferred evaluation designs of educational evaluators. *Studies in Educational Evaluation, 37,* 134-143.

Brandon, P. R., & Fukunaga, L. L. (2013). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation, 35,* 26-44.

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), The practice-theory relationship. *New Directions for Evaluation, 97,* 7-36.

Christie, C. A. (2011). Advancing empirical scholarship to further develop evaluation theory and practice. *Canadian Journal of Program Evaluation, 26,* 1-18.

Christie, C. A., & Alkin, M.C. (2012). An evaluation theory tree. In M.C. Alkin (Ed.), *Evaluation roots* (2nd ed). Thousand Oaks, CA: Sage.

Christie, C. A., & Fleischer, D. N. (2011). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation, 31,* 326-346.

Coryn, C. L. S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32,* 199-226.

Coryn, C. L. S., Westine, C. D., Wilson, L. N., Ozeki, S., Fiekowsky, E. L., & Hobson, K. A. (under review). *A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014*. Manuscript submitted for publication.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56,* 331-364.

Cullen, A. E., Coryn, C. L. S., & Rugh, J. (2011). The politics and consequences of including stakeholders in international development evaluations. *American Journal of Evaluation, 32,* 345-361.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Davies, R., & MacKay, K. (2014). Evaluator training: Content and topic validation in university evaluation courses. *American Journal of Evaluation, 35*, 419-429.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use: Results from a survey of U.S. American Evaluation Association members. *American Journal of Evaluation, 30*, 158-175.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation, 68*, 15-32.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. In C. A. Christie (Ed.), The practice-theory relationship in evaluation (pp. 69-80). *New Directions for Evaluation, 97*. San Francisco, CA: Jossey-Bass.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*, 377-410.

LaVelle, J. M., & Donaldson, S. I. (2010). University-based evaluation training programs in the United States 1980–2008: An empirical examination. *American Journal of Evaluation, 31*, 9-23.

Mark, M. M. (2008). Building a better evidence for evaluation theory: Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 111-134). New York, NY: The Guilford Press.

Mathison, S. (2007). What is the difference between evaluation and research and why do we care? In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 183-196). New York: Guilford Press.

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation, 31*, 390-399.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation, 27*, 296-319.

Patton, M. Q. (2008). *Utilization-focused evaluation*. Los Angeles, CA: Sage.

Patton, M. Q., Grimes, P. S., Guthrie, K. M., Brennan N. J., French, B. D., & Blyth, D. A. (1977). In search of impact: An analysis of the utilization of federal health evaluation research. In C. H. Weiss (Ed.), *Using social research in public policy making*. Lexington, MA: Lexington Books.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation, Volume I* (pp. 39-83). Chicago, IL: Rand McNally.

Scriven, M. (1980). *The logic of evaluation.* Inverness, CA: Edgepress.

Scriven, M. (1986). New frontiers of evaluation. *Evaluation Practice, 7,* 7-44.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Thousand Oaks, CA: Sage.

Shadish, W. R. (1998). Evaluation theory is Who we are. *American Journal of Evaluation, 19,* 1-19.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *Evaluation Practice, 14*, 237-242.

Stake, R., Migotsky, C., Davis, R., Cisneros, E. J., DePaul, G., Dunbar, C., Jr., . . . Chaves, I. (1997). The evolving synthesis of program value. *Evaluation Practice, 18*, 89-103.

Stufflebeam, D. L., & Coryn, C. L. S. (2015). *Evaluation theory, models, and applications* (2nd ed.). San Francisco, CA: Jossey-Bass.

Vallin, L. M., Philippoff, J., Pierce, S., & Brandon, P. R. (2015). Research-on-evaluation articles published in the *American Journal of Evaluation*, 1998-2014. In P. R. Brandon (Ed.), Research on evaluation. *New Directions for Evaluation, 148*, 7-15.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159-176). New York, NY: Russell Sage Foundation.

CHAPTER II

THE LOGIC OF EVALUATION IN PROFESSIONAL DEVELOPMENT EVALUATION PRACTICE: EVIDENT OR NOT SO MUCH?

Scriven's (1980) logic of evaluation is considered a fundamental reasoning process in conducting evaluation that consists of four steps: (1) identify criteria, (2) set standards, (3) measure performance on the criteria relative to the standards, and (4) synthesize the results into a value judgment. Although the logic is generally supported by evaluation scholars, there is little evidence that the logic is explicitly or implicitly practiced by evaluators. The purpose of this study is to investigate the use of Scriven's logic of evaluation in evaluation reports. To achieve this goal, a content analysis of 18 evaluation reports of educational professional development was conducted. The content analysis sought to identify criteria, standards, measured performance, and a synthesis procedure to determine whether and to what extent the logic of evaluation was utilized. Results indicate that standards were commonly not explicitly established and any synthesis methodology was not identified in the sample of evaluation reports.

Introduction

Scriven (1980, 1990) proposed a specific logic to follow in conducting evaluation. The logic is called "the logic of evaluation," which provides a reasoning process in evaluation practice. Considered a meta-theory of valuing practice, Scriven's logic of evaluation provides a fundamental logical process in evaluation (Shadish, Cook, & Leviton, 1991). The logic of evaluation can be described as the four steps of conducting evaluation: (1) establishing criteria, (2) setting standards, (3) measuring performance on the criteria relative to the standards, and

(4) synthesizing the results into a value judgment (Fournier, 1995). The first step involves determining the aspects of an evaluand to be evaluated. The second step is concerned with setting standards against which performance on the identified criteria will be compared. The third step is to measure performance relative to the identified criteria and compare it with the predetermined standards. The fourth step is to synthesize the findings into an evaluative conclusion. As long as evaluation involves determining value, this basic logic should be utilized in evaluation practice (Shadish, 1998). Fournier (1995) further developed the logic of evaluation by introducing the concept of working logic, a logic that is tailored and modified to individual evaluation situations. She considers Scriven's logic of evaluation as a general logic of reasoning that underlies any evaluation practice, while explaining that the working logic is a logic that appears in different forms when applying the general logic to different evaluation contexts. Thus, the logic of evaluation is an essential and fundamental logic that should be utilized in all evaluation practice.

However, there is no agreement among evaluation scholars on the entire concept of Scriven's logic of evaluation. One area of disagreement is standard setting. There is no agreement on when and how standards of performances should be set, or even whether standards should be explicitly identified. In her definition of evaluation, Weiss (1998) states that standards are explicit or implicit, implying diversity in standard setting among evaluators. Stake et al. (1997) argued that standards are typically not identified explicitly, while Davidson (2005) discusses how to make standards explicit by using a rubric to judge performance. Thus, different evaluators are likely to approach a procedure to establish standards in different ways. In addition, the last step, synthesis, is controversial among scholars. After a detailed analysis of the logic of evaluation, Shadish et al. (1991) questioned Scriven's last synthesis operation. Cronbach (1982)

recommended creating separate conclusions for each criterion because different conclusions have differing degrees of warrant that support them. Stake et al. (1997) objected to Scriven's argument for valuing by claiming that valuing involves a more complex process requiring both intuition and conscious reasoning than simple use of a rubric for valuing. Therefore, although it provides a basic reasoning process in evaluation practice, there is a certain degree of uncertainty in the whole theory of Scriven's logic of evaluation.

Building on Scriven's logic of evaluation, Davidson (2005) provided specific methods for standard setting and synthesis. These methods, called "evaluation-specific methodology," guide evaluators to draw conclusions by specifying processes to determine values. The evaluation-specific methodology particularly explains how to create a rubric to explicate standards of performance. Such rubrics of performance standards identify what level of performance on a particular dimension falls into which value rating. For instance, a rubric to determine values of a workshop to increase teachers' knowledge indicates that an increase in 10 points on a test means that the workshop was moderately effective, while a 20-point increase indicates a highly effective workshop. By applying such a rule, value assignment becomes clear and transparent. Davidson introduces various rubrics to analyze different types of data, such as qualitative, quantitative, and mixed data. Thus, a rubric can clarify the process of converting data into values statements in a well-defined way. In addition to explicit standard setting, the evaluation-specific methodology provides a clear procedure to reach an overall conclusion of a program. One method is a quantitative synthesis methodology, in which an overall conclusion can be reached by considering the importance of each dimension (criterion), assigning each criterion an importance score, and then incorporating the importance score into the value on the criterion. When a criterion has sub-dimensions (sub-criteria), performances on sub-dimensions can be

synthesized into a conclusion on the criterion level by the same synthesis methodology used for an overall conclusion. This study described here also examined the evaluation-specific methodology because the methodology provided by Davidson is an expansion of Scriven's logic of evaluation. Therefore, this study examined standards of performance, criteria-level conclusions, and program-level conclusions. Detailed discussion of the evaluation-specific methodology can be found in Davidson (2005).

Purpose

Although the logic of evaluation seems to be a central component of evaluation practice, few studies exist that have specifically examined the logic. Therefore, the purpose of this study is to accumulate empirical knowledge on evaluators' use of Scriven's logic of evaluation. One way to investigate evaluation practice is through content analysis. Content analysis, a method to investigate the content of documents in a systematic and empirical way, has been applied to examine various aspects of evaluation. Using content analysis, past studies examined evaluators' practice of methodology (Christie & Fleischer, 2011), empowerment evaluation (Miller & Campbell, 2006), theory-driven evaluation (Coryn, Noakes, Westine, & Schröter, 2011), stakeholder involvement (Brandon & Fukunaga, 2013), and mixed-methods (Greene, Caracelli, & Graham, 1989). One issue of these studies is that they examined articles in peer-reviewed journals, which might not contain sufficient information on evaluation practice due to page limits. Journal articles are also usually modified based on reviewers' feedback. In addition, studies with statistically significant findings might be more likely to be published in peer-review journals than those without significant results. Although evaluation reports do not contain all the evaluation activities and processes that occurred during evaluation practice, evaluation reports are likely to include more information than peer-reviewed journals articles. This study

investigated evaluation reports to explore evaluators' use of the logic of evaluation. By conducting a content analysis of evaluation reports, this study identified whether and to what extent the logic of evaluation was applied in professional development evaluation practice in education. Education is a reasonable area for investigating the logic of evaluation because education is one of the major substantive areas of evaluation. In addition, evaluation scholars, such as Stake and Cronbach, who expressed their disagreement on certain parts of the logic, might have practiced evaluation mainly in education as their academic background indicates. Their disagreement might have arisen from their experiences in educational evaluation because evaluation theories and ideas tend to be influenced by the evaluator's experiences and contexts (Alkin, 2004). Thus, investigating the use of the logic of evaluation in education is a reasonable endeavor to accumulate empirical evidence of the use of the logic. This study investigates evaluation reports of professional development (PD) for elementary science education in the United States.

Research Questions

As discussed previously, the purpose of this study is to investigate whether and the extent to which evaluators utilize Scriven's logic of evaluation. Content analysis of evaluation reports was conducted to accomplish this purpose. Thus, the focal question is: To what extent can the logic of evaluation be identified in evaluation reports? In particular,

1. Were evaluation criteria explicated?
    a. If so, what were the sources and types of the criteria?
2. Were standards of performance set?
    a. If so, what types of standards?
    b. If not, how is performance evaluated?

3. How was performance measured?
4. When conclusions are described, how are the conclusions reached?

Method

Sample

The sample of evaluation reports included in this study stemmed from a larger evaluation study sponsored by the National Science Foundation (Award #1228809), which was designed to investigate evaluations of elementary science-teaching professional development interventions. In this study, a multi-step sampling procedure was utilized to collect a sample of the evaluation products. An initial broad scan identified 734 evaluation documents related to elementary science professional development evaluation. Application of inclusion criteria (i.e., documents dealing with science professional development, having sufficient information for coding, addressing professional development in the United States, and being published since 2002) limited the final sample to 55 documents. Of these, 22 documents were considered "non-peer-reviewed." Because the purpose of this study was to examine evaluation reports in terms of how evaluators practice the logic of evaluation, 4 documents were excluded from the non-peer-reviewed documents as they presented a newsletter, a supplement report to a main report, and two synthesis studies that aggregated results of several evaluation reports. Thus, the final sample for this study consisted of 18 evaluation reports. Professional development (PD) in the sample of evaluation reports ranged from a single workshop for teachers to multiple components to improve various aspects of schools, such as strategic planning, partnership, and leadership. Of all the PD components reported in the sample, a teacher workshop to enhance teachers' knowledge and/or teaching practice was the major PD component, addressed in 17 evaluation reports (94%). The most frequently documented purpose of evaluation was formative (n = 11; 61%), while the

rest of the reports indicated an enlightenment purpose (to describe the activities and the effects of PD).

Code Development

Codes were developed based on the focal research questions and a detailed data abstraction form was created. The data abstraction form describes sources of criteria, types of criteria and sub-criteria, standards, designs and methods, and conclusions. The coding variables for criteria and sub-criteria were developed based on the professional development evaluation literature (e.g., Guskey, 2000; Kirkpatrick & Kirkpatrick, 2006) and modified as coding progressed. The evaluation design has three codes: experimental, quasi-experimental, and non-experimental. The method also has three codes: quantitative, qualitative, and mixed methods. Except for using statistical significance as a standard, no codes were identified in the variable, standard. A further qualitative analysis was conducted to explore the nature of standards, which is discussed in the Results section of this article. The variable, conclusions/synthesis, deals with the identification of conclusions and the existence of any synthesis procedure to reach the conclusions. Table 2.1 lists the codes identified during the coding procedure, except for the sub-criteria, which are shown in the Results section.

Table 2.1

Variables and Codes Identified by Content Analysis

Variables	Codes
Sources of Criteria	Program goal, Program theory, Stakeholder, Requirement
Types of Criteria	Quality of PD; Impact on teacher learning; Impact on classroom teaching; Impact on student; Organizational Change; Teacher Leaders/Coaches
Standards: Quantitative	Statistical Significance, None
Standards: Qualitative	None
Standards: Mixed-methods	None
Designs	Experimental; Quasi-experimental; Non-experimental
Methods	Quantitative; Qualitative; Mixed-method
Conclusions/Synthesis	Existence/Nonexistence of Synthesis Methodology

Coding Procedure

Two coders coded the sample of documents. A two-stage coding procedure was employed because of the diversity in the structures of the evaluation reports. First, criteria and sub-criteria were coded and any coding disagreements regarding the criteria were resolved. This process helped to simplify the second coding process because information on standards and methods are embedded in sub-criteria. If criteria and sub-criteria were not identified first, it would have been difficult to find where to locate information on standards and methods. After coding criteria and sub-criteria, the rest of the coding tasks were completed. In addition, when no codes were identified in the variables, standards and conclusions, a qualitative analysis was conducted to explore how performance was evaluated. Themes of performance evaluation were found and organized based on the type of methods used to measure performance. The nature of the synthesis methodology was identified and described in the Results sections.

Before coding all the documents, a calibration procedure was conducted in each stage, in which each coder coded a few documents independently. This procedure is intended to identify any area of coding problem and familiarize coders with the coding procedure (Wilson, 2009). Interrater agreement between the two coders was calculated for all the variables. The average percentage of exact agreement was 82% across all variables and the average agreement for each type of variables was 80% for the criteria variable, including sub-criteria; 85% for the standards variable; 82% for the design and methods variable; and 95% for the synthesis variable.

Results

Sources and Types of Criteria

Most evaluation criteria were clarified via evaluation questions. A few evaluation reports did not include any questions, but the evaluation areas to investigate were explained via the program goal and/or the logic model. Of all the evaluation reports, the most frequently addressed source for criteria was the program goal (n = 18; 100%), followed by program theory (n = 8; 44%), stakeholder involvement (n = 4; 22%), and requirement (n = 1; 0.6%). Of all the criteria identified, impact on classroom teaching was the most commonly coded criterion (n = 14; 78%), followed by impact on teacher learning (n = 13; 72%), impact on students (n = 12; 67%), organizational change (n = 11; 61%), quality of PD (n = 9; 50%), teacher leaders/coaches (n = 3; 17%), and reach (n = 6; 33%). Impact on classroom teaching refers to changes in teaching practice after the PD. The sub-criteria identified include changes in three areas: use of new learning, use of new materials, and change in classroom cultures. Impact on teacher learning indicates knowledge gains for participating teachers as a result of the PD in these areas (sub-criteria): content knowledge, pedagogical knowledge, preparedness for teaching, and changes in attitudes and beliefs. Impact on students is concerned with changes in students' knowledge,

attitudes, and skills. Organization change indicates various external factors that potentially influence new teaching practice after taking PD. Those factors included support from schools and principals, schools' resources and materials, and schools' visions, priority, and policy regarding their science education. Quality of PD deals with how participating teachers experienced the PD, that is, their reactions to the PD content and implementation. Teacher leader/coach indicates a component of PD that dealt with those who take a leadership role in improving science education. For instance, teacher leaders help other teachers to implement a new teaching practice in science. Lastly, reach indicates the aspect of whether the participation goal was or was not met.

Table 2.2 lists a summary of all coded criteria and sub-criteria. As shown in the table, the top three most coded sub-criteria are (1) use of new learning within impact on classroom teaching (n = 15), (2) content knowledge within impact on teacher learning (n = 12), and (3) student knowledge within impact on student. These components were the most frequently evaluated components because most PD programs described in the evaluation reports target participating teachers' knowledge in science and their teaching practice in classrooms, which is theorized to result in increased student knowledge in science. Pedagogical knowledge was also frequently coded (n = 8), because it is an important factor in high quality teaching, as is content knowledge. The higher pedagogical knowledge teachers have, the better their teaching quality is (Shulman, 1986).

Table 2.2

Criteria and Major Sub-Criteria

Criteria	Sub-criteria	*n*
Quality of PD (*n* = 9)	Participant Reactions	7
	Implementation of PD	5
	Content of PD	2
Impact on Teacher Learning (*n* = 13)	Content knowledge	11
	Pedagogical knowledge	6
	Preparedness	8
	Beliefs/Attitude	2
Impact on Classroom Teaching (*n* = 14)	Use of new learning	14
	Use of new material	4
	Classroom Culture	3
Impact on Student (*n* = 12)	Student knowledge	9
	Student attitudes	5
	Student skills	5
Organizational Change (*n* = 11)	Support	6
	Resources/Materials	5
	School Priority	5
	Sustainability	3
	School Policy	2
	School Vision	2
Teacher Leaders/Coaches (*n* = 3)	Reactions to PD	2
	Teacher Leaders' Experiences Leaders	2
	Influence on Teacher	1
Reach (*n* = 2)	Participation	2

Standards

Information on performance standards on each sub-criterion was collected. As discussed in the introduction section, the two coders looked for any system, representing standards, including rubrics, which clearly specify performance levels and corresponding values. To measure performance on all the 139 sub-criteria coded, 62 quantitative methods (45%), 48

mixed-methods (35%), and 29 qualitative methods (21%) were utilized. In all the quantitative methods identified, statistical significance was the only standard applied (n = 36). There were no standards identified in the qualitative or mixed-methods methods (Table 2.3).

Table 2.3

Number and Percentage of Standards Coded

Types of Methods	Number of Methods	Number of Standards	Percentage of Standard Use
Quantitative	62	36	45%
Mixed Methods	48	0	0%
Qualitative	29	0	0%

Without standards or any system for data interpretation, how do evaluators describe and evaluate data? The two coders examined how performance was evaluated by seeking to identify patterns of how performance was described. Table 2.4 summarizes the identified themes and provides examples to illustrate each theme.

In the quantitative methods, statistical significance, absolute value, and relative value emerged. Statistical significance was coded when any statistical procedure was conducted. Effect sizes were discussed in four cases to describe the magnitude of the PD impact on teachers. Absolute value does not involve any comparison, but a judgment was made based on an absolute high or low score. There was no description of how high or low a score should be to draw a conclusion. Relative value indicates a difference between pretest and posttest or between a treatment and comparison group without any statistical test. There was no description of how large the difference needed to be to draw a conclusion.

Table 2.4

Major Themes Identified as Performance Descriptors

Method	Theme	Example
Qualitative (29)	Typical Response (7)	In interview, 7 out of 10 participating teachers indicated that they learned a lot from PD.
	Quote (11)	The quote below illustrates positive experiences by teachers who participated in PD. "PD was well organized and very informative."
	Summary (9)	Based on interview data, PD was very successful to increase teachers' perceptions of science.
	Comparison (2)	PD participating teachers encouraged student engagement than non-participating teachers.
Quantitative (62)	Statistical Significance (36)	A statistically significant increase was observed in teachers' science knowledge after PD.
	Absolute Value (16)	On a survey, the preparedness of participating teachers to practice new teaching strategies was 4.5 on the average out of 5.
	Relative Value (10)	On a pre/posttest, participating teachers' pedagogical knowledge was increased by 10 points.
Mixed-Methods (48)	Triangulation (16)	A statistically significant change in teacher knowledge was confirmed by teacher interviews.
	Elaboration (30)	Teacher survey showed that PD was useful for participating teachers and teacher interview elaborated on the usefulness, explaining how PD was useful for them.
	Exploration (15)	Teacher survey indicated that participating teachers were overall satisfied with the quality of PD, but teacher interview indicated some weaknesses and areas for improvement in PD.

Within the qualitative methods, typical responses referred to common themes identified by analysis of interview data. Common themes were utilized as evidence, but there was no discussion of how common a theme would need to be in a response set to count as adequate evidence. Quote indicates the use of quotes to illustrate findings from interviews. Expert review

occurred when science experts evaluated students' work in science class, but there were no descriptions of how good students' work should be. Comparison indicates a description of the difference by making a comparison, but there was no description of what difference should be observed to draw a conclusion. Summary refers to a summary statement of interview data without the identification of themes and the use of quotes.

In the mixed methods reports, qualitative data typically served to supplement to quantitative data; that is, qualitative data were presented to confirm, elaborate, and explore quantitative data. Triangulation was coded when quantitative data were confirmed by qualitative data from interviews. Elaboration is a code to indicate the use of qualitative interview data to elaborate on findings from quantitative data. Exploration indicates the use of qualitative interview data to further explore what quantitative data did not capture. Multiple codes were used in the qualitative and mixed-methods. For instance, when data from interviews were interpreted by describing common themes with quotes, the codes, typical response, and quote, were collected.

Designs and Methods to Measure Performance

Quasi-experimental designs were most frequently utilized in the evaluation reports (n = 15; 83%), followed by experimental (n = 2; 11%) and non-experimental (n = 1; 6%) studies. All evaluation reports indicated the use of mixed-methods, incorporating both quantitative and qualitative methods. Even the two experimental studies incorporated qualitative data into their quantitative evaluation design to investigate the implementation of a PD program. Table 2.5 summarizes the methods used to measure performance on each criterion. Overall, data on about half of the criteria were collected via mixed-methods. This trend is evident even in outcome evaluations of impact on teacher learning. Impact on teacher learning was coded in 13

evaluation reports. Of these, 8 reports utilized mixed methods to collect data on the criterion. Interviews were typically incorporated into a quantitative method to measure an increase in participating teachers' knowledge.

Table 2.5

Methods at Criteria Level

	Methods		
Criteria	Mixed Methods	Quantitative	Qualitative
Quality of PD (*n* = 9)	4 (44%)	2 (22%)	3 (33%)
Impact on Teacher Learning (*n* = 13)	8 (62%)	4 (31%)	1 (0.8%)
Impact on Classroom Teaching (*n* = 15)	10 (66%)	3 (20%)	2 (13%)
Impact on Student (*n* = 12)	3 (25%)	6 (50%)	3 (25%)
Organizational Change/Support (*n* = 11)	8 (73%)	0 (0%)	3 (27%)
Teacher Leaders/Coaches (*n* = 3)	2 (66%)	0 (0%)	1 (33%)
Reach/Participation (*n* = 6)	0 (0%)	6 (100%)	0 (0%)
Total (*n* = 69)	35 (51%)	21 (30%)	13 (19%)

Synthesis Methodology

Criteria level synthesis method. When a criterion was comprised of sub-criteria, the two coders examined how a criterion-level conclusion was reached. A criterion-level conclusion describes the conclusion regarding a criterion by evaluating the sub-criteria consisting of the criterion. Among the 30 conclusions identified at the criterion level, no synthesis method was identified. In most cases, a criterion-level conclusion was supported by sub-criterion level conclusions that comprised the criterion. For instance, one evaluation report indicated that the PD had positive impacts on students (a conclusion on a criterion) because most students reported a high level of interest in science (a sub-criterion conclusion 1), they showed positive attitudes toward science (a sub-criterion conclusion 2), and they increased their scores on science (a sub-

criterion conclusion 3). Additionally, when a qualitative study or a mixed-method study was conducted, quotes and themes were inserted into a criterion-level conclusion to add more descriptions to the conclusion. Typically, strengths and limitations were depicted to illustrate a big picture of the criterion-level conclusion. However, any procedure to reach a criterion-level conclusion was not identifiable. Clear value words, such as good, bad, excellent, and poor, were not utilized.

Final synthesis method. Of the 18 evaluation reports, half included PD level conclusions. A PD level conclusion is a summary statement of whether PD initiatives as a whole were effective. Of those 9 conclusions, no synthesis procedure was identified. Therefore, it was unclear how the conclusions at the program level were reached. A similar pattern to the one observed in the criterion-level synthesis was identified. A program level conclusion was supported by the conclusions on the criteria. One evaluation report, for example, indicated that the PD program played a key role in assisting participating schools in making progress toward improving their science programs (a program level conclusion) because the PD was well received by participating teachers (a criterion conclusion 1), those teachers learned a lot from the PD (a criterion conclusion 2), they improved their teaching skills (a criterion conclusion 3), and they had positive influences on their students' learning (a criterion conclusion 4). Yet, a clear synthesis procedure was not identifiable from the evaluation reports. As in the conclusions at the criterion level, obvious value words, such as good, bad, excellent, and poor, were not utilized in the conclusion sentences.

Discussion

Scriven's logic of evaluation is an important concept in evaluation practice, and presumed to underline all evaluation activities (Shadish, 1998). This study was intended to

examine whether Scriven's logic of evaluation can be identified by content analysis of evaluation reports. Findings suggest little evidence that the logic of evaluation was applied explicitly in the sample of the products. Specifically, standards were typically not clearly specified and procedures for synthesis were not documented. Although content analysis cannot reveal all the evaluation process, the findings of this study correspond with Stake et al. (1994), stating that "Seldom are the criteria seen as direct criteria of merit but rather information categories from which interpretations of merit are made. Standards are seldom explicitly identified. Personal judgment is common" (p. 92).

However, it is not possible to draw a conclusion that Scriven's logic of evaluation was not practiced; his logic might have been applied but was not reported in the evaluation reports. In addition, evaluators might not recognize and intentionally apply the logic of evaluation in their evaluation practice because Scriven's logic of evaluation may not be a logic to follow explicitly, rather one that is used implicitly in conducting evaluation with or without any conscious awareness. Shadish et al. (1991) consider the logic as a meta-theory of evaluation, saying that "His logic of evaluation—selecting criteria of merit, setting standards, and assessing performance—is always implicit in evaluation but rarely appreciated by evaluators" (p. 94). Thus, evaluators are unlikely to be aware of the logic and purposefully apply it to their evaluation practice. Scriven's logic of evaluation can be considered as a general logic of reasoning that underlies any evaluation practice, which is applied to different evaluation contexts as working logic (Fournier, 1995). The working logic is a logic that appears in different forms when applying the general logic. Thus, evaluators might be aware of the working logic because it is specific to the type of evaluation they are conducting. However, they might not be aware of

the general logic because it is the logic behind the working logic they use in their evaluation practice.

Although the lack of using Scriven's logic of evaluation does not indicate any sign of good or bad evaluation practice of the evaluators in the reports, it is important to consider potential benefits of explicit use of the logic of evaluation, particularly explicit standards and synthesis procedures. As many evaluation scholars consider evaluation as involving determination of value (Christie & Alkin, 2008; Schwandt, 2015; Scriven, 1990), the process of determining values is one of the central components of evaluation. To convert performance data into value statements, standards of performance (or certain rules to determine values), are required (Davidson, 2005). Whether or not this process involves personal judgment, it is important to reveal this valuing process. For the advancement of the field of evaluation, evaluators should be proudly and willingly involved in valuing practices for the betterment of evaluands, which potentially impact human functioning. The use of Scriven's logic of evaluation does not need to be explicitly reported all the times; a low stake, simple evaluation might not require the documentation of a detailed evaluation process. However, it is important for the evaluation community to focus on a valuing process. As Shadish (1998) discussed in his presidential address, one important way to differentiate professional evaluators from other professions in the social sciences is that professional evaluators have knowledge of evaluation theory. It is essential for professional evaluators to have knowledge of evaluation theories to advance the evaluation profession and discipline. Scriven's logic of evaluation can become such an evaluation theory that professional evaluators need to know and should apply consciously into their practice, regardless of whether they report on it. Without any licensure to be a professional evaluator, the current status of evaluation as a profession is not well established. Although it

remains unknown whether so called “evaluation-specific methodology” is necessary to be a competent evaluator, the field of evaluation will benefit from more research on evaluators’ valuing practices, which evaluators should conduct if they are doing evaluation.

Limitations

One limitation of this study is that the sample was intentionally drawn from the gray literature. Unlike published articles in academic peer-reviewed journals, evaluation reports are not stored at a certain website in a systematic way. Although every effort was made to obtain as many evaluation reports as possible by contacting key informants of the reports, there are probably some evaluation reports that would have met the inclusion criteria but were not available for this study. It is possible that characteristics of evaluation reports not publically available online are different from those of the sample of the evaluation reports examined in this study. Thus, the findings of this study are not generalizable to other evaluation reports.

Another limitation involves bias regarding coding. Although efforts were made to minimize coding bias by going through the calibration process and calculating interrater agreement, coding bias is inevitable. Coding bias also occurred due to the diversity in the structures of the evaluation reports because different evaluation reports interpret criteria and sub-criteria are in different ways. Although some criteria and sub-criteria were easier to code than others, it was not easy to code all the variables in a clear-cut way. For instance, the criterion, impact on teacher learning, was relatively easy to code, but quality of PD and organizational change were sometimes difficult to code due to confusion in what was measured on those criteria.

Future Research

Although this study did not provide evidence that evaluators applied Scriven's logic of evaluation, it does not indicate that his logic has been ignored in evaluation practice. As Shadish et al. (1991) indicated, the logic might not be explicitly applied and is implicit in nature. Thus, one area for future research is to investigate whether evaluators are familiar with the logic of evaluation and how they perceive and apply the logic. In addition, because Scriven's logic pertains to valuing practices, another area to investigate is how evaluators are involved in valuing. For instance, it is unknown how evaluators consider making evaluative conclusions. If they perceive it as unnecessary, they are unlikely to conduct any synthesis methodology to reach evaluative conclusions. Furthermore, it is not unknown whether and how evaluators establish performance standards. Examining issues related to standard setting is an interesting area to investigate because it will reveal how evaluators assign values. Due to a context-dependent nature of evaluation practice, it would be worthwhile to examine contextual influences on valuing practices.

References

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Brandon, P. R., & Fukunaga, L. L. (2013). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation, 35,* 26-44.

Christie, C. A., & Alkin, M. C. (2008). Evaluation theory tree re-examined. *Studies in Educational Evaluation, 34*, 131-135.

Christie, C. A., & Fleischer, D. N. (2011). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation, 31*, 326-346.

Coryn, C. L. S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32*, 199-226.

Cronback, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation, 68*, 15-32.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, CA: Corwin Press.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels*. San Francisco, CA: Berrett-Koehler.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation, 27*, 296-319.

Schwandt, T. A. (2015). *Evaluation foundations revisited: Cultivating a life of the mind for practice*. Stanford, CA: Stanford University Press

Scriven, M. (1980). *The logic of evaluation*. Inverness, CA: Edgepress.

Scriven, M. (1990). The evaluation of hardware and software. *Studies in Educational Evaluation, 16*, 3-40.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1-19.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4-14.

Stake, R., Migotsky, C., Davis, R., Cisneros, E. J., DePaul, G., Dunbar, C., Jr., . . . Chaves, I. (1997). The evolving synthesis of program value. *Evaluation Practice, 18*, 89-103.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159-176). New York, NY: Russell Sage Foundation.

CHAPTER III

AN EXAMINATION OF EVALUATORS' PERCEPTIONS AND APPLICATION OF SCRIVEN'S LOGIC OF EVALUATION

Scriven's (1980) logic of evaluation is considered a meta-theory of evaluation, providing a fundamental reasoning process of conducting evaluation. Therefore, evaluators are likely to follow Scriven's logic of evaluation in their evaluation practice whether consciously or not. However, there is no complete agreement among evaluation scholars on whether evaluators should follow his logic and, if so, to what extent. In addition, few empirical studies examined whether evaluators actually utilize Scriven's logic in their evaluation practice. The purpose of this study was to examine evaluators' perceptions and application of Scriven's logic of evaluation. Findings revealed that many of the respondents were unfamiliar with Scriven's logic of evaluation and did not always follow it, although they considered it important in their evaluation practice.

Introduction

Evaluation is a relatively young discipline and has been expanding in the recent years. The number of evaluation jobs and professional evaluation organizations has been on the rise in recent years (Donaldson & Christie, 2006). Along with this expansion of evaluation as a profession, evaluation practice has become diverse because of unique characteristics related to evaluation. First, there is no official professional credential to work as a professional evaluator in the United States (Worthen, 1999). Without any required training to be an evaluator, evaluators are likely to have different academic backgrounds. For instance, evaluators with a psychology

degree are likely to have different academic experiences from those with an education degree. The difference in academic experiences can lead to a difference in evaluation practice. Additionally, due to lack of an official license to be an evaluator, people without any or little training in evaluation can work as an evaluator (Altschuld, 1999). Those evaluators might need to learn how to conduct evaluation through their practical evaluation experiences on the job. Therefore, diversity in evaluation practice can be created among those with different types of evaluation jobs. Evaluators in education are likely to practice evaluation differently from those who work in an international development context.

Second, different evaluation practices emerge to accommodate various evaluation circumstances. Unlike research that is usually conducted in a controlled setting, evaluation is typically conducted in a real-world setting (Mathison, 2007). Even evaluators with identical evaluation training and experiences would practice evaluation differently to deal with external contextual demands. Contextual factors, such as budgets, time, data constraints, and political influences, influence evaluation practice (Bamberger, Rugh, & Mabry, 2012). Evaluation budgets usually impact evaluation practice because evaluation activities cost money. Evaluators might have to change their evaluation design to reduce the costs of data collection. Time is another important factor that influences evaluation practice because evaluation typically has a deadline and needs to be completed within a certain time frame. Data collection methods that require a long time might need to be avoided because of a time constraint. Stakeholder influence is not avoidable because evaluation typically involves stakeholders. Using a simulation study, Azzam (2010) indicated that evaluators are more likely to modify their evaluation designs in response to the influence of stakeholders with greater political power than those with less power. Thus, evaluators might have to change evaluation methods due to stakeholders' requests. In this

way, evaluators need to consider various external factors and evaluation practice changes to deal with them in a timely fashion.

Third, there are other factors contributing to diversity in evaluation practice. Evaluators' internal factors, such as evaluators' preferences for certain data collection methods or certain evaluation approaches, can increase diversity in evaluation practice. Azzam (2011) revealed that evaluators with quantitative-orientation are more likely to use quantitative methods than those with qualitative-orientation, suggesting the possibility of evaluators' internal factors influencing evaluation practice. Additionally, evaluators' preferences for a certain evaluation model influence evaluation practice because different evaluation models require different evaluation activities. For instance, an evaluator who utilizes a theory-driven evaluation needs to collect data to test whether a theory behind an intervention is true. Therefore, evaluators' internal factors lead to variability in evaluation practice. Furthermore, evaluation practice needs to be modified based on the maturity of an evaluand. Differences in evaluands require different types of evaluation (Rossi, Lipsey, & Freeman, 2003). For instance, a new program tends to require more outcome evaluation than a well-established program because the outcome of a new program is typically unknown and needs to be investigated, while a well-established program calls for implementation evaluation because the outcome has been already established by other locations (Hansen, 2005).

Shared Logical Process

Even with existing diversity in evaluation practice, evaluators may share a certain procedure when conducting evaluations.

Scriven (1980) proposed that evaluation practice should involve establishing criteria, setting standards, and collecting data relative to standards on the criteria. Regardless of

differences in evaluators' academic backgrounds, evaluation experiences, contextual factors, and/or evaluators' preferred evaluation approaches, evaluators should decide criteria on which values are based, establish standards of performance to judge the performance on the selected criteria, and measure performance and compare it with the determined standards. In addition, Scriven (1994) discusses the importance of synthesizing results from the previous steps into a single value judgment of merit and worth of an evaluand. These procedures are called "the logic of evaluation" consisting of the four steps of conducting evaluation: (1) establishing criteria, (2) setting standards, (3) comparing measured performance with standards, and (4) synthesizing the results into a value judgment (Fournier, 1995). After a detailed analysis of Scriven's logic of evaluation, Shadish, Cook, and Leviton (1991) claim that Scriven's logic of evaluation provides a basis framework for evaluation practice. Additionally, there are other evaluation scholars who discuss the importance of Scriven's logic of evaluation in evaluation practice. Owen (2006) indicates that Scriven's logic is involved in the process of making a value judgment of an evaluand. Fitzpatrick, Sanders, and Worthen (2011) agree that determining criteria and standards and applying the standards to form a judgment are essential features of evaluation. Furthermore, Shadish (1998) states that this basic logic should be utilized in evaluation as long as evaluation involves determining value. Thus, Scriven's logic of evaluation is an essential and fundamental logic that should be utilized in evaluation practice (Fournier, 1995). However, Scriven's logic of evaluation has not been examined empirically; some scholars question certain parts of the logic. For instance, the last step of synthesizing the results into a single evaluative conclusion is questionable (Shadish et al., 1991). In addition, Stake et al. (1997) raise issues of following Scriven's logic of evaluation to make value judgments. Therefore, Scriven's logic of evaluation does not guarantee high-quality evaluation. Although Scriven's logic of evaluation might provide

a sound logic, few empirical studies investigated whether evaluators follow his logic in their actual evaluation practice. In addition, no study has investigated evaluators' familiarity and perceptions of Scriven's logic of evaluation. Therefore, it remains unknown whether evaluators actually follow Scriven's logic of evaluation.

Purpose and Research Questions

As previously discussed, evaluation theorists and scholars extensively discussed Scriven's logic of evaluation, but it was unknown whether evaluators actually know and apply his logic into their evaluation practice. Therefore, the purpose of this study was to investigate whether evaluators knew Scriven's logic of evaluation and to what extent they used his logic in their evaluation practice. Additionally, this study explored evaluators' perceptions of Scriven's logic to gain a better understanding of how important and useful evaluators considered the logic. With this purpose, the focal questions investigated in this study were as follows:

1. To what extent do evaluators apply Scriven's logic of evaluation?
    a. How often do evaluators perform each step of his logic?
    b. How difficult it is to perform each step of his logic?
2. What are evaluators' views on Scriven's logic of evaluation?
    a. How familiar are evaluators with his logic?
    b. How useful do evaluators consider his logic?
    c. How important do evaluators consider his logic?

Method

Sample

An AEA member email list was obtained following an application procedure with the AEA Research Request Task Force and its approval. The list had a total of N = 7231 individuals who were an AEA member as of March 2016. With a bound on the error of estimation of ±5% and the assumption of a population proportion of p = 00.50, a random sample of n = 365 AEA members was estimated for this study. To accommodate potential nonresponse, a 20% oversample (n = 73) was taken resulting in a total sample size of n = 438 AEA members. During the administration of the surveys, 4 of the AEA members selected for inclusion in the sample were excluded due to their undeliverable email addresses (n = 2) and their request to opt out (n = 2), resulting in the final sample of n = 434. From this final sample, 130 AEA members initiated the survey. Furthermore, because the purpose of this study was to understand evaluators' perspectives and practices, 12 respondents who indicated their no current evaluation activity were excluded, leading to an analysis of the results from 108 AEA members with a response rate of 24.88%. To generalize the results of this study to the AEA population, demographic data on AEA members were requested. However, comparable data were not available as requested. Thus, this study was unable to be generalized to the whole AEA population, although it provided valuable insights from experienced evaluators with the average evaluation experience of 10.18 years. In Table 3.1, the demographics on the sample of the evaluators included in this study are summarized. As shown in Table 3.1, 70.41% of the respondents are female and 91.84% of the sample indicated that they conducted program evaluation most regularly in their evaluation practice. Many of the respondents completed graduate schools, with 51.02% holding a doctoral degree and 43.88% holding a master's degree.

Table 3.1

Demographics of the Sample

Demographic Information	*N*	Percent of Total
Gender		
Female	69	70.41%
Male	26	26.53%
Prefer not to answer	3	3.06%
Highest Level of Education		
Doctorate	50	51.02%
Masters	43	43.88%
Bachelors	4	4.08%
Other	1	1.02%
Role as an Evaluator		
External	51	52.04%
Internal	20	20.41%
Mix of Both	27	27.55%
Country of work setting		
United States	83	84.69%
Other	15	15.31%
Primary Work Setting		
Private Business	26	26.53%
College/University	22	22.45%
Nonprofit organization	21	21.43%
Local Agency	9	9.18%
Federal Agency	8	8.16%
State Agency	7	7.14%
School System	3	3.06%
Other	2	2.04%
Type of Evaluation Regularly Conducted		
Programs	90	91.84%
Portfolios	3	3.06%
Policies	2	2.04%
Other	4	4.08%
Number of Years Conducting Evaluation (*M*, *SD*)	10.18 (7.36)	

Note. Due to nonresponse, not all the respondents answered the demographic questions.

Instrumentation

Using the Qualtrics web-based survey system, an online survey was created to answer the focal research questions. The survey focused on questions related to Scriven's logic of evaluation and its components. In the survey, respondents were asked about the frequency to which they performed each step of Scriven's logic and the difficulty in completing each step. Then, they were asked about their familiarity with and their views on Scriven's logic of evaluation. In addition to the questions about Scriven's logic, the participants were required to answer demographic questions, such as their evaluation experiences, gender, their highest degree, and their work setting. The survey consisted mostly of closed-response and partially closed-response items. Open-response items were used to follow up with particular responses to closed-response items. For instance, when a participant chose "difficult" or "very difficult" to the question of "How difficult is it to perform this task?" the follow-up question to understand the nature of the difficulty appeared. Additionally, the skipping patterns were utilized to reduce the burden of response when applicable.

Results

Frequency and Difficulty of Performing Each Step of Scriven's Logic

The respondents were asked about their frequency of performing each step of Scriven's logic of evaluation in their typical evaluation practice. Each step was described without mentioning Scriven's logic of evaluation to minimize a potential risk of bias. If they indicated "never" or "infrequently" to any of the steps of Scriven's logic of evaluation, the respondents were asked to briefly provide their reasons for not frequently conducting the step. In addition, unless they indicated "never" performing each task, respondents were asked about their perceived difficulty in performing the task. Table 3.2 summarizes the percentages of the

respondents who performed each step “frequently” or “very frequently” and those who perceived each step “difficult” or “very difficult.” As shown in Table 3.2, the respondents indicated that they performed all of the steps relatively frequently, although standard setting and comparing with standards were not performed often. Overall, not many respondents perceived each task difficult, with criteria-identification being the most difficult task.

Table 3.2

Frequency and Difficulty of Performing Each Step

	Criteria	Standard	Compare	Synthesis
Performing				
Never	3%	6%	7%	9%
Infrequently	15%	30%	31%	16%
Frequently	60%	53%	45%	58%
Always	22%	10%	17%	17%
Difficulty				
Not at all	4%	5%	12%	6%
A little	46%	47%	61%	53%
Difficult	45%	44%	24%	36%
Very difficult	6%	5%	3%	6%

Criteria determination was the most frequently practiced task: 82% of the respondents (n = 89) indicated that they performed this task (60%, frequently; 22%, always). The most frequently described (7 out of 15 responses) reason for performing this task “never” or “infrequently” was lack of necessity of identifying criteria because they were typically pre-determined by grant requirements or stakeholders. Statements such as “Many are based on grant/funder requirements or fidelity reviews” and “Evaluation criteria are typically set by the

democratic process in the context of my work" exemplify this point. The second most frequently practiced task was synthesis; 73% of the respondents practiced this task (60%, frequently; 13%, always). As for reasons for not performing this task often (i.e., never or infrequently), the most frequently described reason (11 out of 15) was because they usually provided evaluation findings to their clients without any synthesis procedure. This is illustrated by statements such as "I don't tend to make single judgments, but instead identify areas of improvement and areas of high performance" and "I provide impact data so others can judge." The frequencies of performing the other two tasks are identical: 61% of the respondents (n = 66) conducted standard setting frequently (51%) and always (10%), while 61% of the respondents (n = 64) practiced the task of comparing performance with standards frequently (44%) and always (17%). As for reasons for "never" or "infrequently" setting performance standards, the most frequently described reason (25 out of 32) for not setting performance standards was irrelevance of setting standards in their evaluation practice. Quotes such as "Mostly formative evaluation for continuous improvement of training or services, so don't set performance standards" and "My evaluation tasks don't often require setting performance standards of an evaluand" exemplify this point. The most frequently described reason for not comparing performance (20 out of 26) with standards was similar to the ones with standard settings: "Many programs do not have a standard to compare against." and "We have found it difficult to identify similar system level types of programs and standards."

Reasons for Difficulty

If they perceived any of the tasks as "difficult" or "very difficult," the respondents were prompted to briefly present their reasons why they considered the task difficult. One major theme that emerged across all the steps was challenges due to stakeholder involvement.

There were 40 responses that described reasons for difficulty in determining criteria. Of those, 24 indicated that their challenge was associated with stakeholders, such as working with stakeholders to agree on the criteria or lack of stakeholders' involvement or their knowledge in identifying criteria, noting, "Agreeing on appropriate evaluation criteria among stakeholders sometimes poses challenges" and "In many situations, the client is needed to help us determine these criteria and often, they themselves, have not fully considered outcomes and goals of their programming to develop evaluation criteria." Many of the other respondents indicated their difficulty without a specific example, including, "It isn't difficult in that I struggle to complete it. It is difficult in that you need to exercise care with question structure, terminology, administration method, etc., and sometimes you must do so in complex environments." However, there were some specific reasons, such as issues with how to define an evaluand, saying, "One of the main challenges is making sure that you have adequately defined the features and important aspects of the program or activity you're evaluating," and "Lack of clearly defined goals and priorities and benchmarks to measure those goals." Additionally, there are a few comments that identified measurement of criteria as a reason for difficulty in criteria determination. Statements such as "determining the measurements" and "Getting content experts to clearly articulate how they would measure good/ know what counts" illustrate this difficulty.

There were 33 respondents who stated their difficulty in setting standards. Of those, the most frequently described reason (n = 19) was difficulty in finding appropriate standards often due to the complex nature of evaluation. For instance, one described "Complex interventions cannot be benchmarked or compared easily on performance." Another stated that

> Often there is no basis upon which to make the performance standard. Some outcomes require a change in something (e.g., a percentage change in a measure), but it is not rarely clear what is a practically significant change. Statistical significance, while important, does not tell the whole story.

Similar to criteria, stakeholders was a factor influencing difficulty in setting standards (n = 11). Statements such as “Settling on a common definition of progress amongst all stakeholders.” and “Determining what is realistic for the population, getting everyone to agree . . .” illustrate this point. As for comparing performance with standards, the most frequently mentioned difficulty (12 out of 22) was to make right comparisons, reflected in statements such as “Sometimes finding the appropriate data to compare to can be challenging. While some evaluands may appear similar on the surface, they may be very different, thus running the risk of comparing apples to oranges” and “Lack of consistent literature to which to compare findings.” In terms of difficult with synthesis, most respondents (20 out of 25) expressed that it was hard to integrate findings, sometimes conflicting results, from multiple criteria into a conclusion, illustrated by statements such as, “Hard to decide the weight to give to each criterion, and how to reconcile differences between results on different criteria” and “Can be difficult to make a fair statement based on disparate streams of information and occasionally conflicting findings.”

Evaluators’ Familiarity with and Views on Scriven’s Logic of Evaluation

When evaluators were asked about their familiarity with Scriven’s logic of evaluation, 76% of the respondents indicated that they were not familiar with Scriven’s logic (38%; not familiar at all, 34%; a little familiar). Only 28 % were familiar (20% were familiar and 8% were very familiar).

Usefulness and importance of Scriven’s logic. After answering the question about familiarity with Scriven’s logic, the four steps of Scriven’s logic were described and presented so that the unfamiliar respondents would understand what each step entailed. Then, whether familiar or not, respondents were asked to rate the usefulness and importance of the whole idea of Scriven’s logic in terms of evaluation planning, implementation, and reporting. The results

show that 81%, 71%, and 77% of the respondents considered Scriven's logic useful ("useful" or "very useful") in evaluation planning, implementation, and reporting, respectively. Therefore, Scriven's logic was viewed as being more useful in evaluation planning than in evaluation implementation and reporting. As for the importance of Scriven's logic, evaluation planning was the stage in which Scriven's logic was considered most important (78%), followed by evaluation implementation (77%) and reporting (77%). The results are summarized in Table 3.3.

Table 3.3

Usefulness and Importance of Scriven's logic of evaluation

	Planning	Implementation	Reporting
Usefulness			
Not at all	3%	3%	5%
A little	14%	26%	18%
Useful	51%	49%	50%
Very useful	32%	21%	27%
Importance			
Not at all	3%	3%	5%
A little	17%	20%	18%
Useful	39%	47%	44%
Very useful	40%	30%	33%

Evaluators' Views on Scriven's Logic of Evaluation

After rating the usefulness and importance of Scriven's logic of evaluation, the respondents were asked to briefly explain their views on Scriven's logic. There were positive and negative comments on various aspects of Scriven's logic. On a positive side, there were many respondents (32 out of 52) who indicated the centrality of Scriven's logic in their evaluation

practice. Several statements illustrate this point, including, “It (Scriven’s logic model) still can be and it is important to make sure that the flow of the evaluation makes sense” and “It provides the basic framework for the evaluation of organizational programs.” In addition, one respondent further indicated that Scriven’s logic is important to distinguish evaluation from social science research, noting,

> The logic of evaluation underlies all evaluation inquiry. If the logic is not used as a guide, then the most sensical attempts to do evaluation just end up being social science research, and therefore we only learn the existing situation, not what it means.

Eleven respondents, while expressing their agreement on the concept of Scriven’s logic of evaluation, raised the issue of the narrow perspective of Scriven’s logic. Statements such as “This is a very narrow perspective of evaluation. Evaluation should consider much more than whether something is performing as desired” and “Scriven’s approach provides a defined and clear course of action, so it can be useful. However, the results obtained may reduce the complexity of reality” highlight this point. Of those, 3 respondents specifically mentioned lack of necessity of setting performance standards in their evaluation practice, noting, “I cite Scriven's logic of evaluation a lot, but the performance and standards part I have always found difficult given the type of work I tend to do.”

On a negative side, several respondents (n = 9) indicated the lack of utility of Scriven’s logic due to its inability to deal with the complexity of evaluation they face. “It is hard to argue with such global statements, but they don’t provide sufficient guidance to implement the notions well. You could follow his rubric and still do a very impoverished evaluation,” “In terms of utility when actually conducting one, reality is often not as accommodating to the Scriven’s construct as we might like,” and “The Scriven logic of evaluation may work well in academia,

but is not very useful in the real world," exemplify lack of utility of Scriven's logic in practical evaluation.

Familiarity and Evaluators' Practices and Perceptions

Because Scriven's logic of evaluation is not an explicit logic to follow (Fournier, 1995; Shadish et al., 1991), it was speculated that evaluators' knowledge of Scriven's logic of evaluation did not have a large influence on their evaluation practice. To explore this notion, an association between evaluators' familiarity with Scriven's logic and their evaluation practice was investigated. In addition, a relationship between evaluators' familiarity and their perceptions about Scriven's logic was examined because the familiarity could lead to their favorable views on his logic. To examine these associations, two groups regarding familiarity with Scriven's logic were created. The familiar group consists of those who were not familiar at all or a little familiar, while the unfamiliar group was composed of those who were familiar or very familiar. With the familiar and unfamiliar group, Table 3.4 summarizes the number and percentage of performing each step of Scriven's logic, and those of perceiving it important and useful in evaluation planning, implementation, and reporting.

As shown in Table 3.4, descriptive statistics show the largest difference in the perceptions of Scriven's logic as useful in evaluation reporting (familiar, 93%; unfamiliar, 70%) and as important in evaluation implementation (familiar, 96%; unfamiliar, 73%). In other words, the familiar group tended to perceive Scriven's logic as more useful in evaluation reporting and more important in evaluation implementation than the unfamiliar group did. To examine whether there is a statistically significant different between whether evaluators were familiarity with Scriven's logic, and whether they frequently performed each step of his logic, and whether they perceived his logic important and useful, the chi-square test of independence was administered.

Following the recommendation that the test should be avoided when an expected value in any cell is less than 5 (Agresti, 2013), the Fisher's exact test was utilized when the chi-square test was not appropriate. In Table 3.4, the results of the chi-square test are summarized. When a statistically significant result was identified, the odds ratio was calculated to measure the magnitude of the association. The chi-square test found a statistically significant difference between familiarity and three variables: performing the synthesis procedure, perceiving Scriven's logic useful in evaluation reporting and important in evaluation implementation. The odds ratios for each of these variables are 3.82 for the synthesis procedure, 5.31 for perceiving Scriven's logic as useful in evaluation reporting, and 9.55 for perceiving Scriven's logic as important in evaluation implementation.

Table 3.4

Descriptive Statistics, Chi-square Test and Odds Ratio

	Familiar *n* (%)	Unfamiliar *n* (%)	Chi-square Test *p*-value	Odds Ratio
Performing				
Criteria	24 (84%)	61 (88%)	1[a]	
Standards	18 (62%)	46 (62%)	0.993	
Comparison	18 (62%)	45 (63%)	0.9677	
Synthesis	25 (89%)	48 (69%)	0.0408	3.82 (LL=1.04 to UL=14.01)
Perceived as Useful				
Planning	25 (93%)	53 (79%)	0.1402[a]	
Implementation	21 (78%)	46 (68%)	0.467	
Reporting	25 (93%)	47 (70%)	0.0398	5.31 (LL= 1.15 to UL = 24.62)
Perceived as Important				
Planning	25 (93%)	50 (75%)	0.0932	
Implementation	26 (96%)	49 (73%)	0.0247	9.55 (LL=1.21 to UL=75.62)
Reporting	24 (89%)	48 (72%)	0.1291	

[a]These numbers were calculated using the Fisher's exact test.

Discussion

This study investigated evaluators' applications of and views on one of the most important concepts in evaluation, namely, Scriven's logic of evaluation. The results suggest that each step of Scriven's logic of evaluation was practiced relatively frequently (70% on the average), while approximately three quarters of the respondents were not familiar with Scriven's logic of evaluation (43%, familiar at all; 31%, a little familiar). Therefore, findings are likely to support the notion that Scriven's logic of evaluation is implicit in nature (Shadish et al., 1991). Indeed, some comments indicated this point:

> Although I'm not familiar with this work, the steps described above (each step of Scriven's logic of evaluation) make sense and are important to remember during planning and implementation phases. I think more info is needed beyond what is described here for the reporting phase, but the literature may have more details that I'm not familiar with.

> The four steps are very typical within my evaluation work, I just have not hear(d) them call(ed) Scriven's logic. The steps are very straight forward I can't think of a time when they wouldn't be very useful or very important!

These comments may shed light on the fact that Scriven's logic was utilized implicitly. However, the results of the chi-square test showed a statistically significant difference in performing the synthesis procedure between the familiar group and the unfamiliar group. In other words, evaluators who knew Scriven's logic of evaluation were likely to conduct a synthesis procedure than those who do not. It is possible that knowledge of Scriven's logic inspires evaluators to be more conscious about making a value judgment.

In addition, the results of this study suggest that standard setting and comparing performance was not practiced as frequently as criteria determination and synthesis. Specifically, performance standards were not used as regularly. For instance, one respondent said, "Mostly formative evaluation for continuous improvement of training or services, so don't set performance standards." Another mentioned, "I do mostly impact evaluation in which targets are

set or absent. I provide the impact data but do not judge." Therefore, performance standards are not so relevant to evaluators' actual practice. It is interesting to note that most respondents (91%) indicated they evaluate programs most frequently in their evaluation practice. As Stake et al. (1997) described, Scriven's logic of evaluation might not be as relevant to program evaluators as to product evaluators. Indeed, one AEA member concurred with this point, noting, "It (Scriven's approach) can be quite appropriate when referred to a purchase, when you have to choose between different options of a product, but not for most of the programs, and even less for policy evaluation." Therefore, Scriven's logic of evaluation is probably not so relevant to program evaluation. The results of this study might have been different if the sample had been taken from product evaluators.

Furthermore, this study investigated evaluators' perceptions of Scriven's logic of evaluation. Findings suggest that evaluators' familiarity with Scriven's logic might be related to their views on it in some aspects. Specifically, the more familiar evaluators were with the logic, the more useful they found it in evaluation reporting and the more important in evaluation implementation. According to his presidential address, Shadish (1998) advocates that evaluation theory should become part of the evaluator's identify, indicating that knowledge in evaluation theory distinguishes evaluators from social research scientists. If the field of evaluation supports this idea, it is important that evaluators are more familiar with evaluation theories because knowledge in evaluation theories defines who evaluators are. As this study shows, although many evaluators considered Scriven's logic of evaluation important regardless of their knowledge of it, many evaluators were not familiar with it. If the field of evaluation needs features that evaluators proudly embrace as their identity, evaluators should strive to acquire such a unique knowledge base.

Limitations

One limitation of this study is that this study cannot be generalized to the AEA population. This study was intended to generalize the outcomes to the AEA population with a bound of error of 95% confidence intervals using a random sample from the AEA population. Although a random sample was drawn from the AEA member list, comparable demographic data of the AEA population were not available. Thus, it was not possible to generalize the outcome of this study to the AEA population. Another limitation of this study lies in its focus on the examination of general evaluation practice. Rather than inquiring about specific evaluation practice under particular circumstances, this study investigated evaluators' typical approach to their evaluation irrespective of contextual differences. Evaluation practice is heavily influenced by contextual factors (Bamberger et al., 2012); it might be difficult to discuss evaluation practice without particular evaluation contexts. Although this study did not attend to evaluation contexts to reveal what works for whom under what circumstances, this study was an initial step toward understanding more about the frequency of evaluators' performing each step of his logic and their perceptions of it. Because there were few empirical studies on Scriven's logic of evaluation, this study provided important insights into evaluators' practice in terms of Scriven's logic. Third, as the results indicate, many of the evaluators in this study were not familiar with Scriven's logic of evaluation. Because survey questions were created around his logic, some of the questions might have been confusing for respondents to interpret. For instance, the word "criteria" might have been not recognized among evaluators. Indeed, a few respondents expressed their confusion in understanding this term. Therefore, the evaluators' unfamiliarity with Scriven's logic and lack of clarity of some of the questions create an additional limitation of this study. Lastly, it is desirable for two researchers to code written responses to minimize potential biases. However,

one coder conducted data analysis of the open-ended responses. Therefore, there is a certain amount of risk for potential biases in the analysis of the open-ended responses.

Future Research

This study focused on Scriven's logic of evaluation because there were few empirical studies on it. Specifically this study investigated how frequently evaluators went through each step of the logic and how they perceived it. The results indicated that performance standards were not set as frequently as it was assumed. Therefore, one area for a further investigation is how various contextual factors influence setting performance standards. For instance, product evaluation might involve more standard setting than program evaluation does. In addition, developmental evaluation might not require fixed performance standards due to its evolving nature of developmental evaluation. It would be interesting to investigate how different contextual factors lead to different methods to set performance standards.

Another topic of future research is value statements because Scriven's logic of evaluation is typically considered as a basic structure for making value statements. As Scriven's logic of evaluation indicates, data need to be compared with some type of standards in order to attach values to the data. If no comparisons are made, data simply indicate what they are without any indication of whether something is sufficient, good, or bad. It would be valuable to examine whether evaluators' practice involves value statements and, if so, how they reach evaluative conclusions. It is possible that some evaluators do not need to make any evaluative conclusions in their evaluation practice.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Altschuld, J. W. (1999). The case for a voluntary system for credentialing evaluators. *American Journal of Evaluation, 20,* 507-517.

Azzam, T. (2010). Evaluator stakeholder responsiveness. *American Journal of Evaluation, 31,* 45-65.

Azzam, T. (2011). Evaluator characteristics and methodological choice. *American Journal of Evaluation, 32,* 376-391.

Bamberger, M., Rugh, J., & Mabry, L. (2012). *RealWorld evaluation: Working under budget, time, data, and political constraints*. Los Angeles, CA: Sage.

Donaldson, S. I., & Christie, C. A. (2006). Emerging career opportunities in the transdiscipline of evaluation science. In S. I. Donaldson, D. E. Berger, & K. Pezdek (Eds.), *Applied psychology: New frontiers and rewarding careers* (pp. 243-259). Mahwah, NJ: Lawrence Erlbaum Associates.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2011). *Program evaluation: Alternative approaches and practical guidelines*. Upper Saddle River, NJ: Pearson Education.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation, 68,* 15-32.

Hansen, F. H. (2005). Choosing evaluation models: A discussion on evaluation design. *Evaluation, 11*(4), 447-462.

Mathison, S. (2007). What is the difference between evaluation and research and why do we care? In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 183-196). New York, NY: Guilford Press.

Owen, J. M. (2006). *Program evaluation: Forms and approaches*. New York: Guilford Press.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach.* Thousand Oaks, CA: Sage.

Scriven, M. (1980). *The logic of evaluation.* Inverness, CA: Edgepress.

Scriven, M. (1994). The final synthesis. *Evaluation Practice, 15*(3), 367-382.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1-19.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Stake, R., Migotsky, C., Davis, R., Cisneros, E. J., DePaul, G., Dunbar, C., Jr., . . . Chaves, I. (1997). The evolving synthesis of program value. *Evaluation Practice, 18,* 89-103.

Worthen, B. R. (1999). Critical challenges confronting certification of evaluators. *American Journal of Evaluation*, *20*, 533-555.

CHAPTER IV

AN INVESTIGATION OF EVALUATORS' VALUING PRACTICES

The field of evaluation has grown in recent years. Various evaluation approaches or models have been developed to deal with the complexity of today's evaluation needs, leading to a wide range of evaluation practices. Even with the diversity in evaluation practice, there is a certain amount of agreement in the evaluation community that evaluation typically involves the determination of values. However, little is known about how evaluators conduct valuing practices. This study investigated valuing practices with its focus on standard setting and evaluative conclusions. Findings suggest that many of the evaluators in this study engaged in making evaluative conclusions, but were unlikely to create a single evaluative conclusion. Most of them considered the task of making evaluative conclusions difficult yet important.

Introduction

Historically, evaluation was conducted to measure whether evaluation objectives were achieved in educational contexts. Due to the expansion of the field of evaluation, however, evaluation practices have become diverse with a wide variety of different models or approaches of evaluation today (Stufflebeam, 2001). Different evaluation approaches require various evaluation activities because of their different focuses. Those evaluators who endorse a theory-driven evaluation approach collect data to test the theory behind an intervention (Donaldson, 2007), while evaluators using empowerment evaluation are likely to teach evaluation skills to stakeholders so that they will be able to conduct evaluation and make decisions on their own (Fetterman & Wandersman, 2005). In addition, evaluation practices are likely to differ if

evaluation purposes are different. Scriven (1967) indicates that evaluation purposes are broadly divided into two types: formative and summative. The purpose of formative evaluation is to improve an evaluand based on evaluation results, while summative evaluation is conducted to inform decision-making about the continuation or termination of an evaluand. Therefore, evaluation with a formative purpose may focus on particular aspects of an evaluand to find areas for improvement, while evaluators who conduct evaluation for a summative purpose are likely to focus on the overall function of the evaluand to decide whether funding should continue. Based on the evaluation purpose, evaluation practice is likely to differ. In addition to these basic purposes, there are other purposes of evaluation. For instance, evaluation can be conducted to produce knowledge by systematic inquiry of an evaluand (Owen, 2006). In this case, evaluators strive to produce new knowledge and insights into an evaluand, but stakeholders, such as funders and program managers, will decide how to use evaluation findings. With a variety of evaluation approaches and different purposes, evaluation practice today is very diverse.

Definition of Evaluation

If different evaluators practice evaluation differently, what is it that evaluators are doing as a profession? What is the definition of evaluation? To understand what it is that evaluators are doing, it is important to define what evaluation means. However, there is no agreed-upon definition of the term *evaluation* (Schwandt, 2002). It is possible that evaluators have different ideas of what evaluation means to them. Michael Scriven, one of the founders of evaluation, defines evaluation as the determination of merit, worth, and significance (Scriven, 1967). Other scholars define evaluation in similar ways. According to Fitzpatrick, Sanders, and Worthen (2002), evaluation is defined as "the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit) in relation to those criteria"

(p. 7). Additionally, the Joint Committee (1994) defines evaluation as "the systematic assessment of the worth or merit of an object" (p. 3). What these definitions have in common is a value feature of evaluation. Although a unanimous agreement on the exact definition of evaluation may not exist, there is a general agreement that evaluation involves the determination of value (Schwandt, 2002). If the definition of evaluation includes the determination of values, evaluators should conduct evaluation in a way that provides evaluative conclusions about an evaluand. Shadish, Cook, and Leviton (1991) argue that Scriven's logic of evaluation is a meta-theory of valuing in evaluation because it provides a logical basis for creating value judgments. According to Fournier (1995), Scriven's logic of evaluation is summarized as four steps of conducting evaluation: (1) establishing criteria, (2) setting standards, (3) measuring performance and comparing it with standards, and (4) synthesizing the results into a judgment of an evaluand. These procedures are important in evaluation practice because evaluators need to follow this procedure if evaluation involves values (Shadish, 1998). Therefore, it is generally agreed that evaluation involves values. In order that evaluation provides values, Scriven's logic becomes important because it offers a basic reasoning process for valuing.

Valuing practices in evaluation are important because whether evaluation should involve the determination of values is an issue pertaining to evaluation as a distinct profession. Currently, there is no licensing or credential to work as an evaluator in the U.S. (Worthen, 1999). Thus, compared with other well-established professions such as medical doctors, lawyers, psychologists, and counselors, evaluators might not have a shared core knowledge and skills to conduct evaluation. If valuing is not incorporated into evaluation practice, it would be difficult to distinguish evaluators from social scientists because both utilize methods and tools for conducting social science research. Thus, if the field of evaluation is to be a unique profession

different from other social scientists who provide similar services, valuing practice should become one of the essential unique components that evaluators, not social scientists, should practice. Scriven (2004) states, “It is taking the extra step, from empirical or merely factual research to an evaluative conclusion that marks the evaluator as a practitioner working, at least partly, in a different discipline” (p.235). This statement implies that evaluation as a profession needs to involve valuing practice. Other scholars typically agree on this point. Fournier (2005) states that the value feature of evaluation distinguishes evaluation from other types of inquiry. However, it remains unknown how professional evaluators are involved in valuing practices to reach evaluative conclusions. Although Scriven (1986) argues that the evaluator should determine the value of an evaluand, Alkin, Vo, and Christie (2012) indicate that other evaluation scholars have different views on the degree of evaluators’ involvement in valuing. Additionally, it is not clear whether evaluative conclusions are required in evaluation practice. Some evaluators might simply provide data to stakeholders so that they can make their own conclusions. Due to lack of empirical research into valuing practices, it has not been revealed how evaluators proceed with valuing.

Performance Standard Setting

An important topic related to valuing practices is performance setting. Various evaluation scholars indicate the importance and the necessity of setting performance standards if evaluation involves values. Cousins and Shulha (2007) argue that making judgments about values requires comparison between data and a set of standards. Data need to be compared with a certain set of standards to assign meaning to the data. Without any standards to refer to, it is not easy to explicitly make a value judgment of an evaluand because data need to be blended with performance standards to interpret data (Davidson, 2005). However, setting performance

standards is not a simple task to accomplish. There is no uniform procedure to follow in order to set performance standards. Some evaluation scholars advocate for making a clear set of performance standards, while other scholars raise issues of explicit performance standards. For instance, Davidson (2005) advocates for setting clear standards by using a rubric, in which different performance descriptions correspond to different degrees of value terms such as superior, good, average, poor, and so on. If a certain performance of an evaluand falls in a certain performance description, the value term attached to the particular performance description can be obtained. This way, it is possible to make value judgments about certain performance of an evaluand in a clear and straightforward way. In addition, Patton (2004) recommends setting explicit standards before conducting evaluation. However, it is not always possible to determine clear standards. Based on interviews with experienced evaluators, Stake et al. (1997) argue that standards have not been set as clearly as Scriven suggested. Stake and Schwandt (2006) raise the issue of making value judgments using performance standards because explicit standards of performance can seldom be made in real-world complex evaluation. Although they indicate the importance of comparing with standards, they warn against simple use of standards, noting, "When the comparisons are acknowledged to come from tentative and evolving views of the evaluand and when the multiple realities of participating stakeholders are taken into account, then the misuse of comparison is diminished" (p. 412). As reflected in Weiss's (1998) definition of evaluation as "the systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards . . . ," evaluators do not always determine a clear set of standards in their actual evaluation practice. It may not be easy to set clear performance standards because evaluation practice is a very complex endeavor involving consideration of various conditions of an evaluand and its stakeholders.

As discussed above, different evaluation scholars have discussed valuing practices in evaluation, but valuing has received little attention and there are few empirical studies on it. Coryn et al. (under review) indicate that as small as 4% of all the studies on research on evaluation in the last decade dealt with issues of valuing. Therefore, more research is required to investigate valuing practices, one of the central components of evaluation. It remains unknown how evaluators set performance standards and make evaluative conclusions.

Purpose

As discussed previously, valuing is a key feature of evaluation that can distinguish an evaluator from other social scientists. However, little empirical research has examined evaluators' valuing practices. Therefore, the purpose of this study was to explore valuing practices of evaluators. This study focused on how evaluators performed tasks of setting performance standards and reaching evaluative conclusions. Specifically, this study intended to reveal different aspects of performance standard setting, including whether evaluators typically utilized performance standards in their evaluation practice, who was involved in standard setting, what type of performance standards were frequently used, and what challenges they faced in setting performance standards. In addition, this study explored evaluative conclusions, such as who made evaluative conclusions, how difficult it was to do so, and what challenges they had in the process of making evaluative conclusions.

Research Questions

With the purpose previously discussed, the focal questions investigated in this study were as follows:

1. How do evaluators make evaluative conclusions?

   a. How often do they make evaluative conclusions?
   b. Who is responsible for making evaluative conclusions?
   c. What are challenges in making evaluative conclusions?
2. How do evaluators establish standards?
   a. Do they use standards for evaluative conclusions?
   b. Who are involved in establishing standards?
   c. When do they establish standards?

Method

Sample

An AEA member email list was obtained following an application procedure with the AEA Research Request Task Force and its approval. The list had a total of $N = 7231$ individuals who were an AEA member as of March 2016. Two separate studies were conducted using this AEA list. The first study utilized a sample of 438 from this list, resulting in a total of $N = 6793$ for this study. With a bound on the error of estimation of ±5% and assuming a population proportion of $p = 00.50$, a random sample of $n = 364$ AEA members was estimated to address the focal research questions. To accommodate potential nonresponse, a 20% oversample ($n = 73$) was taken, resulting in a total sample size of $n = 437$ AEA members. During the administration of the surveys, 4 of the AEA members selected for inclusion in the sample were excluded due to their undeliverable email addresses ($n = 2$) and their request to opt out ($n = 1$), resulting in the final sample of $n = 430$. From this final sample, 115 AEA members responded to the survey. Furthermore, because the purpose of this study was to understand evaluators' perspectives, 9 respondents were excluded from the study because of their no current evaluation activity, leading to an analysis of the results from 106 AEA members with a response rate of 24.65%. To

investigate the generalizability of this sample to the AEA population, demographic data on AEA members were requested but were not available. Although this study was unable to generalize the results, this study will provide insights into evaluation practice from experienced evaluators with the average evaluation experience of 12.45 years. Table 4.1 summarizes the demographic information on the AEA member sample included in this study. As shown in Table 4.1, approximately 70% of the respondents are female and 90.32% of the sample indicated their focus on program evaluation in their evaluation practice. Many of the respondents were highly educated, with 46.24% holding a doctoral degree and 48.39% holding a master's degree.

Instrumentation

Using the Qualtrics web-based survey system, an online survey was created to answer the focal research questions. The survey was constructed to answer questions about evaluative conclusions and performance standard setting. To clarify that evaluative conclusions do not indicate just facts from data, evaluative conclusions were described as conclusions involving value judgements about an evaluand. Survey questions were constructed to explore different aspects of performance standard setting, including whether they utilized performance standards in their evaluation practice, who was involved in standard setting, what type of performance standards were frequently used, and what challenges they encountered in setting performance standards. In addition, there were questions to explore the nature of evaluative conclusions, such as who made evaluative conclusions, how difficult it was to do so, and what challenges they had in the process of making evaluative conclusions. In addition to the questions about Scriven's logic, the participants were required to answer demographic questions, such as their evaluation experiences, gender, their highest degree, and their work setting.

Table 4.1

Demographics of the Sample

Demographic Information	*N*	Percent
Gender		
Female	65	69.15%
Male	26	27.66%
Prefer not to answer	3	3.19%
Highest Level of Education		
Masters	45	48.39%
Doctoral	43	46.24%
Bachelors	4	4.30%
Other	1	1.08%
Role as an Evaluator		
External	42	44.68%
Internal	27	28.72%
Mix of Both	25	26.60%
Country of Work Setting		
United States	71	76.34%
Other	22	23.66%
Primary Work Setting		
College/University	30	31.91%
Private Business	25	25.53%
Nonprofit organization	20	21.28%
Federal Agency	6	6.38%
Local Agency	5	5.32%
School System	4	4.26%
State Agency	2	2.13%
Other	2	2.13%
Type of Evaluation Regularly Conducted		
Programs	84	90.32%
Policies	3	3.23%
Portfolios	3	3.23%
Products	1	1.08%
Proposals	1	1.08%
Other	1	1.08%
Number of Years Conducting Evaluation (*M, SD*)	12.45 (9.54)	

Note. Due to nonresponse, not all the respondents answered the demographic questions.

The survey consisted mostly of closed-response and partially closed-response items. Open-response items were used to follow up with particular responses to closed-response items. For instance, when a participant indicated that performance standards usually changed during evaluation, the follow-up question to investigate reasons for the change appeared. Additionally, the skipping patterns were utilized to reduce the burden of response when applicable.

Results

Nature of Evaluative Conclusions

One purpose of this study was to explore the nature of evaluative conclusions, such as whether evaluative conclusions were made, who was primarily responsible for evaluative conclusions, and whether a single conclusion was made. Approximately 90% of the respondents indicated that making evaluative conclusions are part of their typical evaluation practice (frequently, 60.64%; always, 28.72%). However, only 14% of the respondents frequently made a single conclusion in their evaluation practice (frequently, 14%; always, 1%), while the majority of them did not made it often (never, 37%; infrequently, 48%). To make evaluative conclusions, approximately 59% of the respondents indicated their frequent use of performance standards (always, 7%; frequently, 52%), while 42% of them indicated their little use (never, 7%; infrequently, 37%). Therefore, although many evaluators were typically involved in making evaluative conclusions, they were less likely to use performance standards for the conclusions and were very unlikely to make a single evaluative conclusion of an evaluand. The results are summarized in Table 4.2.

Table 4.2

Percentages of Respondents Conducting the Followings Tasks in Evaluation

	Never	Infrequently	Frequently	Always
Evaluative conclusions	2%	8%	57%	27%
A single conclusion	33%	43%	13%	1%
Performance standards used for evaluative conclusions	7%	37%	55%	7%

When asked who was primarily responsible for making evaluative conclusions, 56.04% respondents indicated that reaching evaluative conclusions was the joint work between stakeholders and evaluators, while 43.96% indicated that it was primarily the evaluator's job. As for the responsibility for a single evaluative conclusion, 55% indicated that evaluators had a major responsibility for creating it, while 41% indicated that it was the joint work between evaluators and stakeholders.

Evaluators' Perceptions about Making Evaluative Conclusions

The respondents were asked about their perceived importance and difficulty of making evaluative conclusions. The results indicated that many of the respondents considered this task difficult (70%) or very difficult (3%), yet important (41%) or very important (51%).

The respondents who considered the task of making evaluative conclusions difficult or very difficult were asked to briefly provide their reasons for the difficulty, while those who did not perceive it as difficult were asked to provide their strategies. Of those who provided written responses (n = 47) for their difficulty, three themes regarding challenges were identified. The most frequently discussed theme (n = 16) was a theme related to data. This theme indicates insufficiency of data used for evaluative conclusions and difficulty with making a judgment

about data. This point is illustrated by written responses such as “Evidence is not always fully conclusive so judgments have to be made” and “Exercising judgment over and above what the data tell us. Documenting a transparent pathway between diverse types of data and the evaluative conclusions.” The other themes (n = 12, respectively) were stakeholders and complexity of evaluation. The theme “stakeholder” includes difficulty in involving stakeholders in the process of making evaluative conclusions. This theme was illustrated by statements such as “Evaluators and stakeholders sometimes don’t agree when making evaluative conclusions often because of the priorities each have to/for their respective communities (e.g., scientific community)” and “it’s sometimes hard to determine conclusions that the stakeholders don’t already know or don’t currently see in a particular way.” The “complex” theme indicated the difficulty in making evaluative conclusions due to the complex nature of evaluation. Statements were given such as “Understanding the context for a result is complex. There are sometimes policy changes or priorities that take place over evaluative conclusions” and “There are often many factors contributing to an impact, making it difficult to conclude the extent to which any one factor contributes.”

There are only 13 written responses regarding strategies in dealing with evaluative conclusions. Of those, 7 respondents indicated the importance of involving stakeholders in reaching evaluative conclusions, with a statement such as “We solicit stakeholder feedback on our findings and the standards by which we judge.” The other responses indicate the adequacy of expertise in the intervention evaluated (n =3), and the importance of having clear evaluation targets (n = 3) and meaningful evaluation questions (n = 2).

Nature of Performance Standard Setting

Another purpose of this study was to examine the nature of performance standard setting in the process of reaching evaluative conclusions because setting performance standards is a critical aspect of making evaluative conclusions (Davidson, 2005). Various aspects of setting performance standards were examined, such as whether evaluators use performance standards to make evaluative conclusions, what type of standards they use, when evaluators set them, whether performance standards change during evaluation and, if so, why. As discussed previously, approximately 59% of the respondents indicated their regular use of performance standards (always, 7%; frequently, 52%). To avoid collecting data from those evaluators who did not use performance standards, if they indicated they "never" used performance standards, the subsequent questions relating to standard setting were skipped. Descriptive statistics of characteristics of performance standards are summarized in Table 4.3. As shown in Table 4.3, performance standards were typically set by evaluators and stakeholders collaboratively (79.38%) with 10.31% and 9.28% indicating the sole responsibility of evaluators and stakeholders, respectively. The most commonly used type of performance standards was a mixture of absolute and relative standards (60.82%), followed by relative standards (22.68%) and absolute standards (13.40%). Many of the respondents (87.76%) indicated that performance standards were typically set in the stage of evaluation design and a few respondents seemed to set them after designing evaluation, such as data analysis (3.06%), data collection (2.04%), and data interpretation (2.04%). In addition, it seems that performance standards tended not to change frequently, with 27.83% indicating frequent changes in performance standards (frequently, 26.80%; always, 1.03%). When changes in performance standards occurred, it seems that stakeholders typically request changes (59.38%).

Table 4.3

Descriptive Statistics of Characteristics of Performance Standards

Characteristics of Standard Setting	*N*	Percent
Who sets performance standards		
Evaluator(s) and Stakeholder(s)	77	79.38%
Evaluator(s)	10	10.31%
Stakeholder(s)	9	9.28%
Other: Existing Standards	1	1.03%
Type of performance standards		
Both absolute and relative standards	59	60.82%
Relative standards	22	22.68%
Absolute standards	13	13.40%
Not sure	3	3.09%
When to set up performance standards		
Evaluation design	86	87.76%
Data analysis	3	3.06%
Data collection	2	2.04%
Data interpretation	2	2.04%
Don't know	5	5.10%
Frequency of Change in Performance Standards		
Never	13	13.40%
Infrequently	57	58.76%
Frequently	26	26.80%
Always	1	1.03%
Who requests changes?		
Stakeholder	38	59.38%
Evaluator	10	15.63%
Both	10	15.63%
It depends	6	9.38%

Reasons for Changes in Performance Standards

When they experienced changes in performance standards, the respondents were asked to briefly provide the reasons for the changes. A wide variety of contextual factors leading to

changes in performance standards were described. While many responses simply indicated a broad contextual factor, such as “realization that a performance standard doesn’t meet the context of the evaluation” and “Based on what has been learned in the process of evaluating the program,” several responses specified certain factors affecting changes in performance standards. For instance, 6 respondents indicated the initial unrealistic expectations on performance standards, noting “Initial performance standards identified turn out to be unrealistic.” In addition, there were 5 written responses indicating changes due to a potential negative consequence of poor performance of the evaluand, such as “In order not to make the evaluation results (rating) look worse than they expect” and “A desire to demonstrate to funder that certain goals were attempted or achieved.” Furthermore, 8 respondents indicated the developmental nature of the programs they evaluated. They stated that performance standards needed to be modified as the goals of the evaluand changed: “Changes in program direction from the time of inception to the beginning or middle of the evaluation period are the most common cause,” and “developing situation/initiative, goals and priorities change.”

Discussion

Valuing practices are essential part of evaluation practice. Evaluation should incorporate value judgments (House & Howe, 1999), but little research has been conducted to understand valuing practices. Although there are limitations to this study, this study can provide interesting insights into how professional evaluators perceive and conduct valuing practices. First, the results of this study revealed that evaluation practice of many of the respondents involved evaluative conclusions, but only 27% of the respondents indicated “always.” Because the field of evaluation is relatively new without any formal credential to be a professional evaluator, it was unknown whether evaluators provide evaluative conclusions in their practice. This result

suggests that there some occasions when evaluators simply provide data without any value judgments. In addition, it was found that many evaluators in the sample perceived creating evaluative conclusions as difficult yet important. This finding is important because valuing practices are one unique feature that distinguishes evaluators from other social scientists (Fournier, 2005). Evaluators should consider valuing practice an essential part of evaluation. However, valuing practices are not straightforward due to various external factors that influence evaluation practice. It is, therefore, important to conduct more research on valuing practices so that evaluators will find better strategies to deal with their difficulty and will be able to skillfully conduct valuing practices.

Second, this study explored a single evaluative conclusion. Results show that many evaluators did not frequently make a single evaluative conclusion. Among evaluation scholars, there was no unanimous agreement on whether a single conclusion integrating all the data on different criteria was required. For instance, Scriven (1967) advocated for making a single evaluative conclusion, while other scholars, such as Cronbach and Shadish, Cook, and Leviton (1991) did not admit the necessity for a single value claim about an evaluand. This study revealed that many evaluators did not typically create a single conclusion in their evaluation practice. However, it does not mean that other evaluators do not deal with a single evaluative conclusion because most of the evaluators in this sample were program evaluators. Scriven's logic of evaluation might be more relevant to product evaluation than it is to program evaluation. It is possible that evaluators who specialize in product evaluation regularly create a single judgment of products under evaluation. On the other hand, unless a summative evaluation is specifically required, program evaluators might not have to create a single evaluative conclusion.

Furthermore, this study revealed that reaching evaluative conclusions was typically the joint work of evaluators and stakeholders, as suggested in evaluation approaches focusing on collaboration. Historically, Scriven (1986) maintains that evaluators should value, but it seems that many evaluators are likely to work collaboratively with stakeholders in making evaluative conclusions.

Limitations

One limitation of this study is generalization of this study into the AEA population. Drawing a random sample from the AEA population, this study intended to generalize the outcomes to the AEA population with a bound of error of 95% confidence intervals. However, comparable demographic data on the AEA population were not available. This result can shed light on how frequently evaluators make evaluative conclusions. Therefore, any generalization cannot be made from this study, although this study provides insights into valuing practices of experienced evaluators. Another limitation of this study stems from the examination of typical evaluation practice. Evaluation practice is affected by external demands to deal with real-world contexts (Bamberger, Rugh, & Mabry, 2012). Rather than inquiring about specific valuing practices under particular circumstances, this study investigated how evaluators conducted valuing practices in their typical practice. Thus, when answering questions of this study, the respondents might have considered different evaluation contexts, which could lead to different responses to the questions. Although most of the evaluators in this sample were program evaluators, there is still diversity within program evaluation. Therefore, it was unknown how contextual factors influenced their responses, although it was important to investigate typical evaluation practice. Lastly, to guard against potential biases, it is recommended that two coders analyze written responses. However, two researchers did not conduct data analysis of the open-

ended responses. Therefore, there is a risk for potential bias in the analysis of the open-ended responses.

Future Research

Because of scarcity of empirical investigations of valuing practices, this study intended to explore evaluators' valuing practices. Although it provides important insights into how evaluators engaged in valuing practices, this study could not provide full details on valuing practices. Due to the complex nature of evaluation due to various contextual factors, valuing practices differ in different evaluation contexts. Thus, it is important to investigate valuing practices in a certain evaluation context. In addition, it is interesting to investigate evaluators' perceptions of their identity as an evaluator. Evaluation is similar to social research in that both fields utilize methods from social science (Mathison, 2007). There is an overlap between evaluators and social scientists. Many evaluation scholars indicate that valuing is a key feature that distinguishes evaluation from other similar inquiry. Therefore, evaluators' identity may be associated with various aspects of their valuing practices.

References

Alkin, M. C., Vo, A. T., & Christie, C. A. (2012). The evaluator's role in valuing: Who and with whom. In G. Julnes (Ed.), Promoting valuation in the public interest: Informing policies for judging value in evaluation. *New Directions for Evaluation, 133,* 29-41.

Bamberger, M., Rugh, J., & Mabry, L. (2012). *RealWorld evaluation: Working under budget, time, data, and political constraints*. Los Angeles, CA: Sage.

Coryn, C. L. S., Westine, C. D., Wilson, L. N., Ozeki, S., Fiekowsky, E. L., & Hobson, K. A. (under review). *A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014.* Manuscript submitted for publication.

Cousins, J. B., & Shulha, L. M. (2007). Complexities in setting program standards in collaborative evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation.* New York: Guilford Press.

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fetterman, D. M., & Wandersman, A. (2005). *Empowerment evaluation principles in practice.* New York: Guilford Press.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2011). *Program evaluation: Alternative approaches and practical guidelines*. Upper Saddle River, NJ: Pearson Education.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation, 68,* 15-32.

House, E. R., & Howe, K. R. (1999). *Values in evaluation*. Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs*. Thousand Oaks, CA: Sage.

Mathison, S. (2007). What is the difference between evaluation and research and why do we care? In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation*. New York: Guilford Press.

Owen, J. M. (2006). *Program evaluation: Forms and approaches*. New York: Guilford Press.

Patton, M. Q. (2008). *Utilization-focused evaluation*. Los Angeles, CA: Sage.

Schwandt, T. A. (2002). *Evaluation practice reconsidered.* New York: Peter Lang.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Scriven, M. (1986). New frontiers of evaluation. *Evaluation Practice, 7,* 7-44.

Scriven, M. (2004). Logic of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1-19.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Stake, R., Migotsky, C., Davis, R., Cisneros, E. J., DePaul, G., Dunbar, C., Jr., . . . Chaves, I. (1997). The evolving synthesis of program value. *Evaluation Practice, 18,* 89-10

Stake, R., & Schwandt, T. A. (2006). On discerning quality in evaluation. In I. Shaw, J. C. Greene, & M. M. Mark (Eds), *The Sage handbook of evaluation: Policies, programs and practices*. London, UK: Sage.

Stufflebeam, D. L. (2001). Evaluation models. *New Directions for Evaluation, 89*, 7-98.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies*. Upper Saddle River, NJ: Prentice Hall.

Worthen, B. R. (1999). Critical challenges confronting certification of evaluators. *American Journal of Evaluation, 20*, 533-555.

CHAPTER V

CONCLUSION

This final chapter reviews the main findings from each study in order to summarize conclusions regarding Scriven's logic of evaluation. Next, the limitations of the three papers are collectively discussed. Finally, areas of future research based on the results of these papers will be discussed.

Review of Main Findings

The three studies in this dissertation explored Scriven's logic of evaluation. In Study One, Scriven's logic of evaluation was investigated through a content analysis of evaluation documents in educational evaluation. To investigate whether and how the use of Scriven's logic of evaluation was documented in evaluation reports, Study One content-analyzed the four components of Scriven's logic of evaluation: (1) criteria, (2) performance standards, (3) methods to collect performance and comparisons between the measured performance and the standards, and (4) synthesis procedures. The findings from this study indicate that the process of reaching evaluative conclusions was not documented in the sample of the evaluation reports. Specifically, most performance standards were not explicitly established and any synthesis methodology was not identified. These findings raised the question of whether practicing evaluators knew Scriven's logic of evaluation and, if so, to what extent they applied the logic in their evaluation practice. Study Two investigated these questions. Additionally, conceptualized as a meta-theory of valuing practices, Scriven's logic of evaluation provides an essential logical sequence in the process of valuing (Shadish, Cook, & Leviton, 1991). In other words, if evaluation involves

value judgments, evaluators are supposed to follow Scriven's logic of evaluation (Owen, 2006). Study Three examined valuing practices with its focus on performance standard setting and evaluative conclusions. Two different samples were drawn from a list of the entire AEA members; one sample is for Study Two and the other is for Study Three. Both studies utilized a survey to answer focal questions for each study. The samples of the evaluators in both studies were experienced evaluators with an average of 10 years conducting evaluation. Approximately 90% of them evaluated programs or projects in their regular evaluation practice. The findings from Study Two revealed that many evaluators were not familiar with Scriven's logic of evaluation, yet found it important and useful in conducting evaluation. In addition, Study Two revealed that evaluators did not always follow all the steps of Scriven's logic of evaluation. Specifically, approximately 40% of the respondents did not frequently set performance standards in their evaluation practice. Study Three explored valuing practices with its focus on performance standards and evaluative conclusions. The findings suggest that many evaluators were likely to make evaluative conclusions, but were unlikely to create a single evaluative conclusion. Most of them considered the task of making evaluative conclusions difficult yet important in their evaluation practice. In addition, about 80% of the respondents indicated that setting performance standards was the joint work of evaluators and stakeholders, but fewer respondents (44%) indicated that reaching evaluative conclusions was the collaboration between evaluators and stakeholders.

Conclusions

Various evaluation scholars argue that valuing is considered one of the key features of evaluation (Fournier, 1995; House, 1986; Schwandt, 2015; Scriven, 2004). However, valuing has received little attention in research on evaluation (Coryn et al., under review). This dissertation

attempted to shed light on valuing. Specifically, Studies One and Two focused on examining the use of Scriven's logic of evaluation using a content analysis and a survey method. It was concluded that evaluators might not follow all the steps of Scriven's logic of evaluation in their evaluation practice. Particularly, evaluators are unlikely to create a clear set of performance standards to create value judgments. This is probably because of the fluctuating nature of evaluation; an evaluand develops, data collection is modified, and stakeholders' opinions change. Explicit performance standards may be unnecessary to deal with the changing nature of evaluation. Another conclusion regarding Scriven's logic of evaluation is that evaluators consider Scriven's logic of evaluation important in conducting evaluation regardless of their previous knowledge of it. Overall, it seems that the previous knowledge does not much influence whether evaluators follow Scriven's logic. Study Three focused on further details of performance standards and evaluative conclusions. It examined various characteristics of performance standards, such as who set them, when they were set, what types of standards were set, and so on. It was found that the nature of performance standards varies even among evaluators who usually conducted program evaluation. In addition, it was concluded that evaluators typically dealt with evaluative conclusions and perceive valuing an important aspect of evaluation. This finding is important because valuing is an essential aspect of evaluation. It is one of the features that help to distinguish evaluators from other social scientists, both of whom rely on similar research methods (Mathison, 2007). Therefore, it is likely that evaluators have a sense of identity as a professional evaluator by conducting valuing practices and considering them important.

Limitations

Throughout the dissertation, several limitations were identified. In Study One, a content analysis was utilized to identify characteristics of Scriven's logic of evaluation in evaluation

documents. Three major limitations were identified. First, evaluation documents of educational evaluation were collected from gray literature. Not all the evaluation reports were systematically stored nor available. It is unknown how representative the sample in Study One was of all the evaluation documents. Second, because a content analysis involves coding by two coders, human errors when coding were unavoidable. In order to minimize coding bias, the two coders conducted a calibration procedure in which any coding issues were resolved. However, coding bias and errors were inevitable. Study Two and Study Three were survey studies, in which the respondents were asked about various questions regarding Scriven's logic of evaluation and valuing practices. First, both studies drew a random sample from the population of the AEA members so that demographics of the samples would be compared with that of the population. However, demographic information of the AEA members was not available, which made it impossible to generalize the findings from both studies. Thus, although these studies provide interesting insights into evaluation practice, no generalization can be made. Second, Study Two and Study Three explored evaluators' typical evaluation practice without much contextual information. Evaluation is a highly contextualized practice, influenced much by external factors. In other words, evaluators are constantly required to deal with those factors (Bamberger, Rugh, & Mabry, 2012). Examining usual evaluation practices might not be as informative as examining specific evaluation practices in certain contexts. Due to lack of empirical research on Scriven's logic of evaluation and valuing practices, Study Two and Study Three can make important contributions to research on evaluation by yielding interesting findings into evaluators' practices. However, it might have been difficult for the respondents to discuss their usual evaluation practices without much evaluation context. Lastly, Study Two and Study Three provided a conflicting result in terms of the frequency of performing the last step of Scriven's logic. Study

Two indicates that almost three quarters of the respondents frequently conducted the last synthesis procedure, while Study Three revealed that not many evaluators made a single evaluative conclusion about an evaluand in their regular practice. One potential reason for this conflicting result is because of unfamiliarity of the respondents with Scriven's logic of evaluation. To reduce the burden of the respondents in order to increase response and completion rates, the questions in the surveys used in Study Two and Study Three were kept as short and concise as possible. Although an effort was made to minimize the confusion by making the questions clear, the descriptions of Scriven's logic of evaluation were minimal. The short descriptions, coupled with the respondents' unfamiliarity with Scriven's logic, might have caused the conflicting result. Lastly, the open-ended questions in Study Two and Study Three were analyzed by one coder. There is a risk of misunderstandings and bias in interpreting the written responses in those questions.

Future Research

The main focus of this dissertation was on Scriven's logic of evaluation and valuing practices. There are several directions for future research to further investigate Scriven's logic of evaluation. One potential area is to further explore the nature of performance standards and valuing practices. Although this dissertation explored characteristics of performance standards used by evaluators and certain aspects of valuing practices, it could not provide a full picture of evaluators' valuing practices. A qualitative study is a suitable method to explore more details on this topic. It would be worthwhile to conduct a qualitative study of performance standards by interviewing selected evaluators to capture the details of setting performance standards. This would help to identify important contextual factors that influence performance standard setting. Similarity, a qualitative study can be utilized to explore various features of valuing practices.

When do valuing practices become difficult? How do evaluators overcome difficulty in valuing practices? Answering these questions would contribute to the field of evaluation by accumulating empirical knowledge on valuing practices. Additionally, the samples used for this dissertation were evaluators who regularly practiced program evaluation. It would be worthwhile to explore evaluation practices of evaluators who specialize in other types of evaluands. For instance, how would product evaluators conduct evaluation differently from program evaluators in terms of Scriven's logic of evaluation? As discussed, is Scriven's logic of evaluation more relevant to product evaluators than evaluators who conduct other types of evaluation? Are there any challenges in conducting evaluation that are specific to the types of evaluands? These questions are interesting because contextual factors need to be identified to answer these questions. Revealing those factors will help to develop contingency theories of evaluation, in which various contingencies are specified, which help to make decisions about evaluation practices that accommodate those contingencies (Shadish, 1998).

References

Bamberger, M., Rugh, J., & Mabry, L. (2012). *RealWorld evaluation: Working under budget, time, data, and political constraints.* Los Angeles, CA: Sage.

Coryn, C. L. S., Westine, C. D., Wilson, L. N., Ozeki, S., Fiekowsky, E. L., & Hobson, K. A. (under review). *A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014.* Manuscript submitted for publication.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation, 68,* 15-32.

House, E. R. (1986). In-house reflections drawing evaluative conclusions. *American Journal of Evaluation, 7*(3), 35-39.

Mathison, S. (2007). What is the difference between evaluation and research and why do we care? In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 183-196). New York: Guilford Press.

Owen, J. M. (2006). *Program evaluation: Forms and approaches*. New York: Guilford Press.

Schwandt, T. A. (2015). *Evaluation foundations revisited: Cultivating a life of the mind for practice.* Stanford, CA: Stanford University Press.

Scriven, M. (2004). Logic of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 235-238). Thousand Oaks, CA: Sage.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation*, *19*(1), 1-19.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice.* Thousand Oaks, CA: Sage.

Appendix

Human Subjects Institutional Review Board
Approval Letters

Approval Letter for Study One

WESTERN MICHIGAN UNIVERSITY

Human Subjects Institutional Review Board

Date: January 20, 2016

To: Chris Coryn, Principal Investigator
Satoshi Ozeki, Student Investigator for dissertation

From: Daryle Gardner-Bonneau, Ph.D., Vice Chair Daryle J. Gardner-Bonneau

Re: Approval not needed for HSIRB Project Number 16-01-20

This letter will serve as confirmation that your project titled "The Logic of Evaluation in Professional Development Evaluation Practice" has been reviewed by the Human Subjects Institutional Review Board (HSIRB). Based on that review, the HSIRB has determined that approval is not required for you to conduct this project because you are not collecting personal identifiable (private) information about individual and your scope of work does not meet the Federal definition of human subject.

45 CFR 46.102 (f) Human Subject

(f) *Human subject* means a living individual about whom an investigator (whether professional or student) conducting research obtains

(1) Data through intervention or interaction with the individual, or
(2) Identifiable private information.

Intervention includes both physical procedures by which data are gathered (for example, venipuncture) and manipulations of the subject or the subject's environment that are performed for research purposes. Interaction includes communication or interpersonal contact between investigator and subject. *Private information* includes information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (for example, a medical record). Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects.

Thank you for your concerns about protecting the rights and welfare of human subjects.

A copy of your protocol and a copy of this letter will be maintained in the HSIRB files.

1903 W. Michigan Ave., Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276
CAMPUS SITE: 251 W. Walwood Hall

Approval Letter for Study 2 and Study 3

WESTERN MICHIGAN UNIVERSITY

Human Subjects Institutional Review Board

Date: January 25, 2016

To: Chris Coryn, Principal Investigator
Satoshi Ozeki, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 16-01-24

This letter will serve as confirmation that your research project titled "An Empirical Examination of Evaluation Practices" has been **approved** under the **exempt** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may **only** be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., ***you must request a post approval change to enroll subjects beyond the number stated in your application under "Number of subjects you want to complete the study***)." Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: **January 24, 2017**

1903 W. Michigan Ave., Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276
CAMPUS SITE: 251 W. Walwood Hall