Reporting Credibility in Educational Evaluation Studies that Use Qualitative Methods: A Mixed Methods Research Synthesis

A dissertation presented to

the faculty of

The Patton College of Education of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Hongjing Liao

May 2015

This dissertation titled

Reporting Credibility in Educational Evaluation Studies that Use Qualitative Methods: A Mixed Methods Research Synthesis

by

HONGJING LIAO

has been approved for

the Department of Educational Studies

and The Patton College of Education by

Krisanna Machtmes

Associate Professor of Educational Studies

Renée A. Middleton

Dean, The Patton College of Education

Abstract

LIAO, HONGJING, Ph.D., May 2015, Educational Research and Evaluation, Social Studies Education

Reporting Credibility in Educational Evaluation Studies That Use Qualitative Methods: A Mixed Methods Research Synthesis

Director of Dissertation: Krisanna Machtmes

Qualitative methods have increasingly been applied in program evaluations of policies and interventions (Patton, 2003; Sandelowski & Barroso, 2007), and establishing the quality of evaluation findings is of essential importance. Yet little is known about how qualitative quality criteria of evaluation reports have been applied in practice (Brandon & Singh, 2009; Gephart, 2004). One of the ways to assess such practice can be to examine the reporting of credibility techniques in published evaluation reports that use qualitative methods. This study examined the practice of establishing qualitative credibility in the context of program evaluation by carrying out a research synthesis of the related studies published in six leading evaluation journals (from 2003 to 2012). Mixed methods are used to identify key credibility techniques, document the frequency of the techniques, and to describe their use and properties. There were 118 articles found eligible for analysis. The finding suggests that the reporting of credibility techniques has been relatively steady over the past decade. The majority of articles addressed basic methodological quality of the study, but most authors are not sensitive in addressing credibility of their qualitative findings with credibility techniques.

Dedication

To my loving parents, Wei, and Spring: thank you for being there for me.

To my five-years-ago self: good choice, and you made it.

Acknowledgments

I would like to thank my dissertation committee members who were more than generous with their expertise and precious time. Special thanks to Dr. John Hitchcock, for being such a great advisor, mentor, and dear friend. I would like to thank him for his countless hours of reading, editing, and discussing with me in person and on-line throughout the entire process. His guidance and encouragement have made this a pleasant and rewarding journey. I would like to extend my sincere appreciation to Dr. Krisanna Machtmes, my Chair, for taking me on and providing all support for me to move forward; to Dr. Gordon Brooks for his detailed feedback, kind advice, and patient explanations on statistics and everything else; to Dr. Adah Ward Randolph for listening, caring, and sharing, and setting a great example of a strong mind; to Dr. Valerie Conley for broadening my horizons even further and offering me inspiring opportunities.

I would like to thank my fellow doctoral students and friends at OU for their support, feedback, and friendship. To Leslie and Moira for spending time helping me with my codebook and providing feedback to my ideas and drafts. To Chris for his super efficient help with journals and articles. To Gab for proofreading several of my drafts. To Dianne, Weiyu, and Yanju for valuable suggestions and boosting my confidence when thing were difficult.

Last, but certainly not least, thanks to my family, for the love and support that can only come from a great mother and father; for the personal support and tolerance from my husband, and for the unconditional love from my daughter, who constantly reminds me to strive to become the best I can be.

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Chapter 1: Introduction

Program Evaluation is a systematic endeavor that aims to assess the merit of a program and provide useful feedback (Chen, 2005; Scriven, 1996; Shadish, Cook, & Leviton, 1991). The goal of program evaluation is to assist with making sense of and improving policies and programs to achieve social betterment. Formative evaluation and summative evaluation are two broad categories of evaluations, and the purposes and practice of these two types of evaluations could be different (Mark, Henry, & Julnes, 2000). Summative evaluations are designed to yield evidence pertaining to the overall effectiveness of programs, which is often conducted after completion of a program. In contrast, formative evaluation is improvement oriented, where continuous quality improvement is the primary agenda (Patton, 2008; Scriven, 1996). A vivid metaphor is offered by Robert Stake in describing the distinction between the two types of evaluation: “When the cook tastes the soup, that’s formative; when the guests taste the soup, that’s summative” (quoted in Scriven, 1991, p. 169).

Program evaluation is “a very young discipline -- although it is a very old practice" (Scriven, 1996, p. 395). The history of evaluation practice that is comparable to contemporary approaches can be traced back to 1792, when William Farish assessed student performance using quantitative marks, was considered as the first documented formal use of evaluation (Hogan, 2007; Hoskins, 1968). In the 1930s, Ralph Tyler, who was often credited as being the founding father of evaluation in the education field, directed the Eight-Year Study (1932-1940) that evaluated program outcomes in 15 progressive high schools and 15 traditional high schools. Tyler’s contribution to

evaluation was well remembered for his use of an objective-oriented approach to compare outcomes with pre-defined behavioral objectives (Hogan, 2007; Stufflebeam, Madaus, & Kellaghan, 2000).

It was not until the 1970's that evaluation emerged as a profession and began its expansion as an established field of study (Hogan, 2007). Evaluation journals, such as *Educational Evaluation and Policy Analysis*, *Evaluation Review*, and *Evaluation and Program Planning* were published, many of which are still leading journals in the field today. Universities, recognizing the importance of evaluation research, began to offer courses in evaluation methodology (Stufflebeam et al., 2000). Professional associations like the American Evaluation Association were developed, and evaluation standards were established, such as the Joint Committee's Standards for Educational Evaluation.

Today, evaluation is a distinct discipline that emphasizes the practical, and has its own set of rules and approaches (Patton, 2008; Shadish et al., 1991). Methods developed for program evaluation are abundant in literature (Chen, 2005; Rossi, Lipsey, & Freeman, 2004; Shadish et al., 1991), and there has been growing interests among policymakers, the general public, and other stakeholders in understanding the experience of different target groups in the programs, and what actually happens in the field (Wholey, Hatry, & Newcomer, 2010). In recent years, there has been increased use of qualitative methods in program evaluation in order to explore various facets of the programs and hear the voices of the participants, and qualitative methods have been widely applied in evaluations of policies and interventions (Patton, 2003; Sandelowski & Barroso, 2007). For example, they are often used to:

- identify contextual factors that contribute to the success of delivery of a program,
- identify both intended and unintended outcomes,
- study the process of program implementation,
- examine program improvement opportunities, and
- to generally explain problems that are poorly understood (Chen, 2005; Spencer, Ritchie, Lewis, & Dillon, 2003).

In addition, findings from qualitative evaluations are used for making decisions about programs and to advance policy development (Brandon & Singh, 2009; Spencer et al., 2003). Qualitative methods, such as open-ended interviews, analyses of written documents, and direct observations, are often used for the above purposes in evaluations largely because of their capacity to generate emergent, detailed descriptions and in-depth insights about human experiences. The holistic approach of qualitative research often presents multiple standpoints of a program and its impact on stakeholders, which can provide different perspectives to facilitate government based investigations and policymaking (Lichtman, 2010; McDavid & Hawthorn, 2006).

Concurrently rising with the increased number of qualitative evaluation studies are concerns about the quality of evidence generated by this type of research (Barusch, Gringeri, & George, 2011; Lincoln, 1995; Lincoln & Guba, 1985). Such concerns focus on the rigor of qualitative evaluation and the credibility of findings. Results of an evaluation are often used for making decisions about a program, which makes the quality of evaluation findings even more important (Rossi et al., 2004). There is a substantial

body of literature on criteria of "good" qualitative research, as well as instructional texts on rigorous use of qualitative methods, including discussions and debates about a variety of techniques developed to help promote credibility, that is, the degree to which evaluators have faithfully described the program and generated valid conclusions (Bochner, 2000; Brantlinger, Klingner, Richardson, & Taylor, 2005; Cho & Trent, 2006; Lewis, 2009; Lietz & Zayas, 2010; Sandelowski, Voils, & Barroso, 2006; Tracy, 2010). One of the aspects that has not been fully addressed in the literature is how criteria for judging quality have been applied in practice (Brandon & Singh, 2009; Gephart, 2004). Furthermore, with rapidly increased number of publications reporting on qualitative (and mixed methods) evaluation findings (Dixon-Woods, Booth, & Sutton, 2007), little is known about what approaches are used and with what frequency. One of the ways to assess such practice can be to examine the reporting of credibility techniques in published qualitative evaluation reports. Credibility techniques refer to strategies and methods developed to document accuracy and promote quality. A more complete definition of this concept is offered later in the chapter.

Another notable movement, which influenced the methodology of this study as a research synthesis, is the lack evidence showing the current status of the field, which can be provided by synthesis research. Scholars have pointed out the shortage of systematic evaluations of program evaluation approaches, and the lack of systematic evidence to guide the practice of evaluation (Henry & Mark, 2003; Miller & Campbell, 2006). Synthesis research is needed to help improve evaluation work, and to help enhance utility of research (Maxwell, 2004; National Research Council, 2002; Sandelowski et al., 2006).

Against the general backdrop of the proliferation of qualitative research in the evaluation field, concurrently occurring with the rising concern about the quality and explicitness of practice of qualitative methods, this research synthesis study was conducted to systematically examine the techniques employed to enhance credibility of qualitative methods in educational program evaluation.

To confirm the point that there does not appear to be any systematic examination of educational program evaluation studies that use qualitative methods, major databases of educational research and evaluation have been searched on October 20, 2012 and January 6, 2013, including EBSCO, Eric, ISI, Educational Research Complete as well as major publishers websites such as Sage and Elsevier. This search revealed no syntheses that focus on quality checks of empirical educational program evaluations that use qualitative methods. This represents a gap because the field of educational evaluation has its own established standards by the *Joint Committee on Standards for Educational Evaluation* (Yarbrough, Shulha, Hopson, & Caruthers, 2011), and credibility techniques are an integral part in promoting these Standards. Evaluation standards such as the ones established by AEA are expected to be followed, and when qualitative approach is applied, credibility techniques can help a design to meet the standards. Thus, having systematic evidence about the current status of this topic is not only in need, but also of great significance in understanding the field and in informing practice. Therefore this study addressed this gap and examined the practice of establishing qualitative credibility in the context of program evaluation by carrying out a research synthesis of the selected

studies published in six leading evaluation journals (from 2003 to 2012), with a focus on credibility techniques.

Purpose of the Study

A broad goal of this study was to advance the understanding of how credibility techniques were applied and reported in educational evaluation journals. To achieve this goal, this study examined the reporting of credibility checks presented in empirical educational program evaluation studies that use qualitative methods (including both qualitative stand-alone studies and qualitative components in mixed method studies) published in six selected leading peer reviewed evaluation journals in education. A mixed method research synthesis was conducted to identify, describe, and evaluate key techniques used by evaluators to enhance credibility of their qualitative work. A mixed methods synthesis protocol was developed to specify data collection procedures, selection of studies and analytic strategies. As an essential part of the synthesis protocol, a communicable codebook was constructed building on the existing credibility criteria and techniques, and adjusted according to the particular need and purpose of this study (Liao & Hitchcock, 2012). Both quantitative and qualitative data were collected, including numerical scores, and coder's notes documenting the use of credibility techniques and the coder's appraisal process. Descriptive statistical methods were used to indicate frequencies of credibility techniques used, and thematic analysis was adopted to identify patterns in the use of credibility techniques.

Research Questions

This study aims to examine the state of credibility techniques in education evaluation practice by addressing two main questions: (a) to what extent are credibility techniques reported in selected journal articles? (b) What are the features that can be observed in the reporting of credibility techniques? This leads to a series of sub-questions:

1. Of the selected published evaluation studies that use qualitative methods, how many of them clearly applied credibility techniques? How many explicitly described the credibility techniques adopted, and how many made only vague references to these techniques?
2. What are the credibility techniques used in practice? What are the most and least commonly used techniques?
3. Are there distinct variations in the vocabulary and terms used to describe credibility techniques?
4. Comparing the practice of credibility techniques in qualitative stand-alone and mixed method studies, does use of credibility techniques vary as a function of the methodologies?
5. When credibility techniques are used, are they used to directly support evidence? If so, how is the quality of the evidence evaluators provide influenced by these techniques?
6. Are there different trends across the ten-year period of time?

By answering the above questions, this study can offer a comprehensive description of the use and presentation of credibility techniques in qualitative evaluation

reports. The first group of sub-questions covered the types, frequencies, and comparison of reporting of these techniques across time and methodological approaches; the second group of sub-questions focused on application details of different techniques, and the last group of sub-questions is about terms and language use of credibility technique reporting.

Significance

The significance of this study lies in its contribution to the current literature in the qualitative evaluation field, its practical application, and its methodological significance in using a mixed methods synthesis.

First of all, there is no apparent research synthesis on empirical educational program evaluations found that explicitly focuses on the application of credibility techniques. This study addresses this gap and ventures to examine the practice of establishing qualitative credibility in the context of program evaluation.

Secondly, a brief review of relevant literature shows that the majority of writing focuses on the definition of good quality research, criteria and techniques to promote quality, but little attention has been paid to developing a summary of current status or providing practical guidelines on reporting credibility. In other words, if credibility techniques are useful as discussed in literature, then it is worth determining whether they appear to be used and reported in evaluation studies that use qualitative methods.

The number of studies devoted to methodological rigor of evaluation studies that use qualitative methods are very limited. Shek, Tang, and Han (2005) examined the quality of evaluation studies using qualitative methods in social work literature. Based on criteria documented in previous literature, Shek, Tang and Han combined both criteria of

post-positivist and constructivism paradigms to examine aspects like the philosophical bases, sampling procedures, reflexivity, and reliability of selected social work studies. Their examination also included strategies for credibility checks, such as triangulation, thick description, and member checking. The results of the study show two major findings: first, there is lack of clarity and details on qualitative approaches and procedures. Second, not many credibility strategies are found in the selected social work studies. For example, many credibility strategies, such as reflexivity and bias control, are highly valued in qualitative methodology, but they are not commonly described in practice. Other essential techniques, such as member checking, negative case analysis, and analyst triangulation, were seldom reported. The authors summarized the results as showing that many social workers seem to be not very conscious about the importance of different credibility strategies in conducting qualitative evaluation. Shek et al.'s study provided valuable information about the extent to which credibility techniques are present in qualitative evaluations in social work literature, but it would be more helpful for them to go further and explain the implication of the absence of strategies. Yet like the authors pointed out in the title, it is a "wake up call," which sets the purpose of the study as calling for awareness. Comparatively, Barusch, Gringeri and George's (2011) study also conducted a review on credibility strategies used to enhance the rigor of qualitative work in social work literature. The authors of the study developed their own template for the review, which incorporated important hallmarks of good qualitative research methodology, such as sampling rationale, clear theoretical framework, and specification of limitation, together with key credibility strategies to form the criteria of evaluation.

The results show that sampling procedures and triangulation are popular strategies used in more than half of the sample articles, but other key credibility strategies, such as peer debriefing and member checking, are not commonly used. Similar to Shek et al.'s study, Barusch et al. contributed to the field by identifying basic methods and credibility strategies qualitative social workers used in practice. The findings seem satisfactory in terms of the basic methods, yet disappointing in terms of the degree to which key credibility strategies were used. However, authors of both studies did not take their studies further to explore the issue regarding when credibility strategies are present and the rigor of the applied strategies. This is one of the things this dissertation will add to the practice of credibility techniques literature.

In addition to studies that give specific emphasis to credibility techniques, there are several that examine the methodological soundness of evaluation studies. Brandon and Singh (2009) reported on the methodological soundness of evaluation studies cited in five reviews of literature. The study examined methodological strength of both quantitative and qualitative methods in evaluation studies, including survey, simulation studies, case study, and narrative studies, and put the focus on the examination of content validity. The results show the current state on educational evaluation as "dearth of validity information" (Brandon & Singh, p. 134), lacking details about methods and self-report bias information. Price et al. (2005) assessed methodological rigor of evaluations studies published between 1980 and 2003 that evaluated training programs of cultural competence. The systematic review sample included both quantitative and qualitative studies, but the authors adopted a set of more quantitatively oriented criteria to examine

evaluation studies in five domains: representativeness (whether settings and subjects were described), intervention description (whether enough details were provided about the intervention), bias and confounding (whether bias was minimized using rigorous strategies), outcome assessment (whether outcomes were validated), and analytic approach (whether analytic approach was reported). The results on the methodological rigor of the evaluations about cultural training programs are not encouraging, and the authors called for more attention to issues like proper design, evaluation, and reporting of these training programs.

Apart from methodological review of evaluation studies, there are also a few reviews that examined quality of qualitative research in different research designs in both qualitative primary research and qualitative synthesis. For example, Dixon-Woods, Booth, and Sutton (2007) evaluated the quality of qualitative research syntheses in healthcare between 1988 and 2004, and Hannes and Macaitis (2012) provided an update of Dixon-Woods et al.'s review between 2005-2008. Both reviews described the methods used for qualitative synthesis, and particular attention is placed on the quality appraisal process of the qualitative studies being synthesized. The results show growing value and importance being attached to the methodological quality of qualitative work, and also the need to improve transparency of qualitative methods used in both appraisal work and empirical research. Furthermore, methodological soundness of qualitative research is also examined in different research fields and theoretical groundings. Koro-Ljungberg and Douglas (2008) examined the methodological rigor of qualitative research in engineering education by carrying out a systematic review on articles published in the *Journal of*

Engineering Education. Other authors focused on qualitative studies based on a particular theoretical perspective, such as De Witt and Ploeg's (2006) methodological appraisal of interpretive phenomenological studies in nursing literature, and Denk, Kaufmann and Carter's (2012) quality assessment of supply chain management (SCM) research that use grounded theory.

The above studies on appraisal of qualitative research and evaluations demonstrate an expansion of the use of qualitative methods in different research areas and formats, and reveal a variety of criteria created for appraisal of qualitative methods. Additionally, the work related to quality assessment of evaluations and qualitative studies in general share a few concerns in common. All relevant reviews call for more informed use of qualitative methods. There are mainly two reported reasons for such a concern: first, reviewers encountered difficulties in deciding the quality or characteristics of the studies due to the lack of information about the methods in the evaluation studies. Therefore, insufficient information makes it hard to conclude the methodological soundness of the study under review. Secondly, based on the information shown in the reports, quality criteria are not adequately addressed in the current qualitative research. Findings reported a common concern of lacking explicitness of specific qualitative approach, research procedures, participants, and preoccupations of the researchers. Inappropriate use of credibility techniques is another frequently raised problem. Shek, Tang, and Han (2005) directly pointed out that credibility techniques are not commonly used, and presentation of some techniques they examined even raise doubts about whether these techniques were thoroughly understood. To address such concerns, this

research synthesis contributes to the gap in literature and serves the needs for a comprehensive understanding of the field in recent years. This study allowed detailed description and analyses of how credibility strategies have been presented and articulated in qualitative evaluation reports. It contributes to the knowledge of how quality principles and key credibility techniques have been incorporated into methodological practice in qualitative evaluation, and provides a better understanding of the reporting of evaluation work that use qualitative methods.

Another significance of this study is that a detailed evaluation of how well evaluators present their chosen approaches to promote credibility could also provide important information for the discussion of evaluation strategies, and help provide practical instructions, recommendations and guidelines for quality assessment of evaluation work. The findings of this study therefore have practical use for the training of evaluators.

This study also has methodological significance. For one thing, this study developed a systematic codebook to systematically gather data. Compared with coding tools of similar studies, this codebook has a few features that are not only tailored to the purpose of this study, but can also be used to evaluate credibility techniques in future reports. For example, the codebook combines structured scoring with flexible open-ended comments; it focuses exclusively on assessing the presence and reporting of credibility techniques, and provides comprehensive categories of credibility techniques. More details about the codebook is provided in Chapter 3.

Finally, this study adopted a mixed methods synthesis design. Scholars (e.g., Harden, 2010; Hogan, 2007) have identified the trend of adopting mixed methods program evaluation, and there is increased acceptance of for multiple-method research. As an emerging methodology field, the basic concepts, approaches, and procedures for conducting mixed methods research synthesis have been developed by theorists, but not much has been provided in literature on how they have been applied in practice (Chen, 2005; Harden, 2010; Wholey et al., 2010). This study applied these strategies, and at the same time participated in refining these techniques in practicing mixed method synthesis for this particular study. In addition, most of existing mixed methods systematic reviews focus on integrating findings. This study reviewed methodological soundness and applied mixed methods on the synthesis level by not only pooling the use of credibility techniques from studies, but also described and analyzed the characteristics of such use.

In short, the research questions posed here have thus far not been asked, and establishing empirical answers are of interest to program evaluators, standards developers, journal reviewers, and consumers of program evaluation findings and results. This study could add new knowledge to the field of evaluation by summarizing key credibility techniques used in qualitative evaluation, and provide a better understanding of the reporting of these techniques in practice.

Basic Assumptions

This study examines the use of credibility techniques in qualitative evaluation, so there are several basic assumptions to be clearly stated. The first assumption is about researcher bias. Since qualitative research promotes a level of self-revealing and

reflexivity, researcher-bias inevitably permeates qualitative inquiry when conducting program evaluations. Having said that, the presence of bias does not necessarily represent an inherent shortcoming of qualitative work; rather, accounting for it and dealing with the likelihood of bias undermining the capacity to draw conclusions is something the program evaluator should pursue if attempting to meet the evaluation standards. Therefore, the goal is not to eliminate bias, because it is assumed to be ever-present as the researcher serves as the instrument in qualitative research. Instead, the presence of bias should be openly discussed when assessing the credibility of evaluation studies.

Although bias and error in statistical endeavors are different constructs, there are some commonalities that can help explain this point. Quantitative researchers deal with error (e.g., Type 1, Type II, measurement error, non-response error, etc.) when engaged in inferential work. Error is an ever-present concept when conducting inferential statistical work and, like bias, it is unrealistic to assume it can be pre-detected or even eradicated from the process of inquiry. Instead, quantitative researchers struggle with the degree to which error is likely to be in sufficient amount to render findings to be hard to defend, should any findings be offered at all.[1] Qualitative program evaluators should likewise consider if bias was in sufficient amount during their inquiry that has become difficult to claim any findings are credible. Researcher bias, along with other problems when describing phenomena such as the completeness of data sources can oftentimes be accounted for by using credibility techniques.

[1] In full disclosure, bias is not limited to qualitative program evaluations, in part because the researcher rejects any notion of a qualitative-quantitative research dichotomy (Hitchcock & Newman, 2012; Newman & Hitchcock, 2011). With the potential exception of a double-blind randomized controlled trial, we assume bias is also present in so called quantitative program evaluation efforts and some training in qualitative inquiry can help researchers to account for bias.

Furthermore, the conceptual principle applied in this synthesis is a combination of predetermined research decisions and an element of flexibility throughout the design, implementation, and analysis stages. To elaborate, most research syntheses have a set of fixed stages including the establishment of an a priori methodology protocol that specifies pre-defined data collection procedures, inclusion and exclusion criteria of the primary studies (Cooper, Hedges, & Valentine, 2009; EPPI-Centre, 2006; Wholey et al., 2010). Syntheses also describe analytic and reporting plans at the outset of the work so as to address the research questions at hand. This is critical because following tools like protocols can promote an empirical basis for addressing research questions in a way that will generate new knowledge and understanding. The emergent design principle is however invoked here. According to Morgan (2008), emergent design is a flexible approach that allows data collection and analysis procedures to evolve in response to what is learned in the research process. Although procedures of emergent design are often applied within the framework of qualitative research, in this research synthesis, within the pre-specified frame of the synthesis protocol, rather than to solidify every aspect of the methodological plan, this synthesis may alter the analysis approaches depending on what is learned during early stages of the synthesis. In this study, changes were made on the scope of sampling due to the number of relevant reports identified in the target journals, the codebook was altered to fit better with the data, and analytic approaches of both quantitative and qualitative methods were used in an interactive manner. However, it should be noted that any change made and rationales for making such changes were explicitly reported and done with committee approval.

In addition, the same principle applies to the codebook construction and its calibration. As Schreier (2012) suggested that coding frames are always partly data-driven, the researcher of this study has been open to emergent changes during the search and identification of the evaluation articles, and tailored the codebook to the selected materials in the actual coding. Credibility techniques such as inter-coder reliability check, triangulation, and audit trial were also be used to help enhance the reliability of the codebook and the overall quality of this study. Details of these techniques and procedures are discussed in Chapter 3.

Definition of Terms

Some of the terms frequently used in this dissertation are defined in this section. These terms are chosen because they are key words for this study; some of the terms have rich connotations that are conceptually defined differently in literature.

Program evaluation: Program evaluation is defined by Chen (2005) as "the application of evaluation approaches, techniques and knowledge to systematically assess and improve the planning, implementation, and effectiveness of programs" (p. 3). The purpose of evaluation is to judge the merit or worth of a program, or to gather data that informs program improvement (Patton, 2003; Scriven, 1996; Stufflebeam & Shinkfield, 2007). Programs, which are often referred to as interventions, are defined as organized efforts to improve human wellbeing (Chen, 2005; Rossi et al., 2004). This study focused on educational evaluation studies, so the program included education-related interventions, products, policies and other forms of practice.

Credibility: Credibility is considered one of the most important indicators of quality in qualitative research (Brantlinger et al., 2005; Cho & Trent, 2006; Creswell, 2009, 2012b; Cutcliffe, 2001; Glensne, 2011; Hannes, Lockwood, & Pearson, 2010; Johnson & Christensen, 2012; Lather, 1991; Lincoln & Guba, 1985; Lincoln, Lynham, & Guba, 2011; Maxwell, 2005; Onwuegbuzie & Leech, 2006; Patton, 1999, 2003). This concept has multiple definitions in the literature. Credibility is the degree to which the phenomenon under study is faithfully described, and the degree to which defensible information, convincing arguments, and interpretations are provided. Efforts to promote credibility should also be made throughout the research process. A detailed definition and features of credibility are discussed in Chapter 2.

Credibility techniques: Credibility techniques are strategies and methods developed to "operationalize" (Lincoln & Guba, 1985, p. 301) quality standards such as credibility in qualitative research. Credibility techniques have systematic procedures or explicit principles to help qualitative researchers validate their research claims and strengthen the credibility of their findings. Examples of credibility techniques include member check, triangulation, and negative case analysis.(Brantlinger et al., 2005; Onwuegbuzie & Leech, 2006; Whittemore, Chase, & Mandle, 2001). Major credibility techniques are described and discussed in Chapter 2.

Research synthesis: Research synthesis refers to a process of scientific inquiry with the primary goal of systematically assessing and integrating empirical research studies relating to a particular question (Chalmers, Hedges, & Cooper, 2002; Cooper & Hedges, 2009; Sandelowski & Barroso, 2007). As an inclusive concept, research

synthesis encompasses a variety of methodological approaches. Depending on different purposes and methods, it can be further categorized into more specific synthesis types, such as meta-analysis, meta-synthesis, and mixed methods synthesis (Cooper et al., 2009; Sandelowski et al., 2006; Suri & Clarke, 2009; Tashakkori & Teddlie, 2003). Although research synthesis has overlapping features with other forms of inquiry, it has long been established and recognized as a type of research on its own, and has played an important role in enhancing utility of knowledge and shaping further research, policy, and practice (Chalmers et al., 2002; Cooper & Hedges, 2009; Sandelowski & Barroso, 2007; Suri & Clarke, 2009). Major steps of an evaluation synthesis often include specifying the topic area, develop a search strategy, inclusion and exclusion criteria for studies under review, a scheme for coding studies, management strategies, analysis strategies, and interpret and report the results (Boruch & Petrosino, 2010; Cooper et al., 2009; Gersten & Hitchcock, 2009). More details about the differences between mixed methods synthesis and other types of synthesis, as well as the current status, features, and methods of research synthesis are provided in Chapter 3.

Mixed method research synthesis: In general, mixed methods research synthesis is to systematically review data and formally summarize knowledge applying mixed methods principles (Sandelowski, Voils, Leeman, & Crandell, 2012). According to Sandelowski et al. (2006), the rise of mixed method synthesis is the result of both the turn to evidence-based practice and the growth of qualitative research in the last two decades. Syntheses of empirical research are produced to address practice problems and are viewed to have potential in enhancing the utility of research, especially qualitative

findings, and the effectiveness of practice. As this study aims at an important aspect of methodological practice, this research synthesis used both qualitative and quantitative methods for an integration of the application of credibility techniques in a shared domain of empirical evaluation research.

Codebook: A codebook is a tool that serves as a guide for reviewing reports. It provides information on the structure and definitions of codes, and documents the link between the text and numeric values assigned to the data (Bourque, 2004; Schreier, 2012). A codebook was built particularly for this research synthesis as a frame to systematically describe selected evaluation reports, and serves as the tool for developing a database of evaluation studies that use qualitative methods. Apart from indicating numerical scores assigned to each items, the codebook also includes explanations about the layout and meaning of codes in order to facilitate coder's recognition of credibility technique featured in the text; and it provides examples that illustrate how different credibility constructs might appear in text to help the coder to link features of the text to the constructs, or coding categories in this case. On the whole, the codebook is not only an instrument for mapping the informational terrain of the text, but also documents a theoretical lens for analyzing and evaluating the practice of credibility techniques. A complete description of the development of the codebook is provided in Chapter 3.

Limitations

Like all studies, this research synthesis has its limitations. They relate to the construction of the sample of program evaluation articles. Although an exhaustive search was conducted within each journal to find all articles that meet the selection criteria, only

a limited number of journals were covered, and examples of qualitative evaluation work published elsewhere or unpublished studies in this area are not included.

In addition, publication bias limits findings of this synthesis. Originally, publication bias was often defined in meta-analysis as that the results of reviews might be biased toward positive results because studies with statistically significant and more positive results are easier to be published (Rothstein, Sutton, & Borenstein, 2005). In other words, what appears in the published literature may not be representative of all the completed studies (Cooper et al., 2009; Sandelowski & Barroso, 2007). Suggestions are made to minimize publication bias in research design. For example, multiple groups of researchers can agree to combine their findings prior to knowing the results of their meta-analysis studies (Berlin & Ghersi, 2005), or to locate and retrieve grey and unpublished literature (Hopewell, Clarke, & Mallett, 2005). Various techniques are also developed to detect publication bias (e.g. the funnel plot), to assess the sensitivity of conclusions, and to adjust for possible effects of publication bias (Rothstein et al., 2005). To put it in the context of this study, what is shown in the six evaluation journals may not be the case for all qualitative evaluation studies. However, the six journals, as a sample of leading evaluation journals that cover educational studies, are purposefully selected. One of the objectives this synthesis study is to inform intended audience of the use of credibility techniques, therefore, it is the purpose of this study to provide a review of evaluation studies published in leading evaluation journals, as they are generally more accessible to a broad audience, and have relatively strong impacts.

Publication time is yet another limitation. The current plan is to examine only articles published after the year 2003. This cut off was established to cover evaluation studies published in the most recent decade, but clearly, examining earlier publications may alter overall findings.

In addition, the analysis and even understanding of the practice of credibility techniques may be constrained by what is presented in published reports without taking account of authors' experiences of producing such reports. This is inherent to the type of text analysis research such as research synthesis, in which data are only written reports. In other words, this study aims to answer the question of what are the credibility techniques and how are they presented in published articles, but not "why" they are presented that way. The "why" questions could become the concentration for further studies on this topic.

Delimitations

To set the boundaries of this study, there are two major characteristics that need to be made explicit. First and for most, it is important to point out that there are arguments against judging qualitative research along a series of universal criteria (Denzin & Lincoln, 2005). The creativity required to address emergent designs and highly contextualized findings can by themselves undermine such attempts. But this study is not promoting universal criteria for judging the broad expanse of qualitative research. Rather, the assessment is limited to the program evaluation world, which can simplify matters given the expectation that such work is to judge the merit or worth of a program, and to gather data that inform program improvement (Patton, 2003). Put another way, since program

evaluation oftentimes pursues fairly concrete goals, this may simplify efforts to judge the application of criteria that can be used to judge the quality of research methods.[2] This is especially the case if evaluators are expected to establish and assess the accuracy of their conclusions, which is a requirement of the *American Evaluation Association's* program evaluation standards (Yarbrough et al., 2011). Moreover, this study focuses on qualitative credibility although the research is open to qualitative methods performed in a mixed methods setting. Both qualitative stand-alone and mixed methods studies are included in the sample, but qualitative criteria are used to examine the qualitative component in the mixed method works.

Another delimitation of this study is that the analysis and assessment are limited to published research articles, not including unpublished information, nor the actual conduct of the research itself. It is to be recognized that many factors contribute to what journal articles may include, published evaluation reports nevertheless reflect and should accurately indicate the research practice and credibility of the findings (Yarbrough et al., 2011).

To summarize, this chapter includes an introduction to the topic of evaluation studies that use qualitative methods, and a read of AEA standards suggest that credibility techniques should be used, but there has been no empirical information found on the degree to which they are used. This study would like to find out the use of credibility techniques in evaluation studies published starting with top evaluation journals. The

[2] The discussions are limited to program evaluation efforts that use qualitative methods to directly comment on the merit or worth of a program, and/or offer direct advice on program improvement (i.e., formative program evaluation). This is to be distinguished from efforts that might inform program evaluation work but are otherwise distinct steps. Examples include formative research, rapid reconnaissance, and so on.

finding would have implications for both journal practices and the application of qualitative methods in program evaluations.

Chapter 2: Review of Literature

This dissertation aims to understand the use of credibility techniques in program evaluations that apply qualitative methods. The purpose of this chapter is to review pertinent literature so as to describe what is currently understood about the application of credibility techniques in the context of program evaluation, and features of mixed methods synthesis. Specific topics to be covered are program evaluation standards, debate pertaining to the merit of promoting standards in the context of qualitative inquiry, how credibility techniques can be used, what is mixed methods synthesis research and to what criteria can such mixed methods research be assessed.

Program Evaluation Standards

The terms program evaluation, evaluation, and evaluation studies are often used interchangeably (Rossi et al., 2004). They all share the common idea of using systematic social research procedures to delineate and explain the merit or worth of a program's planning, operation, effects and social implications (Chen, 2005; Mark et al., 2000; Scriven, 1996; Stufflebeam & Shinkfield, 2007). For sake of convenience, all these terms are referred to in this study as program evaluations.

Evaluation standards are important as they are shared understandings of what quality evaluations are and how they should be conducted (Yarbrough et al., 2011). Evaluation standards provide guidance that address different dimensions of planning, implementation, and utilization of evaluations. In the past 30 years, professional evaluation standards were developed and applied in a wide range of professional societies and disciplines worldwide. In this section, three major evaluation standard systems are

discussed, namely the *Joint Committee Evaluation Standards* in North America, the European adaptation of the Joint Committee standards in national evaluation societies, and the evaluation standards endorsed by large international evaluation communities. The North American Joint Committee on Standards for Educational Evaluation is undoubtedly one of the earliest and most influential organizations when it comes to evaluation standards. Founded in 1975 and supported by 17 sponsoring organizations including the *American Evaluation Association* (AEA), the Committee's evaluation standards not only take the lead in establishing professional evaluations in the U.S., but also have become the most important driving force for the development of evaluation studies. By now the Committee has produced *The Program Evaluation Standards* (1981, 1994, 2011), *The Personnel Evaluation Standards* (1987, 2009), and *The Student Evaluations Standards* (2004). These standards have been widely adopted and applied for guiding and assessing evaluations in the U.S. and Canada, and adapted and used in countries outside of North America and beyond the education discipline (Stufflebeam, 2004).

Apart from the Joint Committee Standards, there are two other important standard systems that worth mentioning. One system consists of the various national evaluation societies in Europe, such as evaluation societies established in Switzerland, Germany, France, Italy, Slovenia, and the UK, each of which has developed their own guiding principles and standards (European Evaluation Society, 2012). The other important evaluation standard system is large international organizations like the United Nations

and the European Union. However, as this synthesis study focuses on standards of evaluation in the United States, the emphasis is placed on the Joint Committee Standards.

The Joint Committee Evaluation Standards

The Joint Committee Standards are used as an important source of this study, especially in constructing the coding categories of the codebook. These standards have been widely referenced in the program evaluation literature and are extensively used in practice (Russ-Eft, Bober, de la Teja, Foxon, & Koszalka, 2008; Stufflebeam, 2004).

The third edition of *The Program Evaluation Standards* (published in 2011) is organized by five major attributes with thirty individual standards. The five major attributes are briefly introduced in this section. Examples of sub-sets of standards under each attribute are discussed and compared to credibility techniques of qualitative research in later sections of this chapter. The five attributes are:

Utility: Utility standards are created to help ensure the evaluations conducted are serving the needs of intended stakeholders, so the evaluations can have positive outcomes and substantial influence. Utility standards define the use, misuse, and influence of evaluations, and highlight the competence of evaluators, engaging diverse stakeholders and their changing needs, and providing information to help participants and users of evaluations to be more confident working in their programs.

Feasibility: Feasibility standards describe the factors to be considered before implementation of the evaluation, and to maintain and improve the efficiency and effectiveness of evaluations.

Propriety: Propriety standards address the ethical and legal concerns in evaluations. These standards require the evaluation to be conducted with due regard for the welfare of those involved and those that could be affected by the results of the evaluation.

Accuracy: Accuracy standards are intended for judging and increasing the accuracy of findings and conclusions. These standards comprehensively include criteria for credibility/validity and reliability of evaluation representations, propositions, interpretations, and reporting.

Accountability: Accountability standards discuss adequate documentation and reflection of evaluation process and products.

The different components of the Joint Committee Standards suggest that evaluation has matured into a well-established field that is becoming more open and inclusive, which is best illustrated in the following features brought by the standards. To start with, the Standards defined a common evaluation language as well as conceptually agreed-upon general guidelines for educator and evaluators to follow. Yet more importantly, these standards broadened the very concept of quality evaluation. By including major attributes like feasibility, utility, and propriety, the perception on evaluation assessment is no longer limited to a judgment on internal and external validity of the studies (Campbell & Stanley, 1966; Stufflebeam, 2004). Similarly, evaluation standards explicitly require evaluator credibility, contextual information, and viability. Such standards thus include not only guidance pertaining to typical experimental designs, but also a qualitative and mixed method approaches. Recent years have seen two

emerging trends: an expanded use of qualitative methods, and rather than depending exclusively on one type of approach, there has been a trend of combining quantitative and qualitative methods in one program evaluation (Worthen, Sanders, & Fitzpatrick, 2004). The changing landscape of the field generated the needs to incorporate the new trends into the existing standards. However, it is necessary to note that general evaluation standards are expected to be followed regardless of the methods used, and when qualitative approach is applied, credibility techniques can help a design to meet the standards. More detailed discussion and examples are provided in later sections.

Qualitative Criteria: Concepts, Terms and Connotations

Two arms make up the conceptual basis for this study which aims at examining credibility techniques in evaluation work that used qualitative methods. One of them is evaluation standards, and the other is the notion of credibility as a quality criterion for qualitative research. Before going straight to the topic of credibility, it is necessary to first introduce the general background of quality criteria in qualitative research as a whole.

Given the complexity and the dynamic nature of qualitative research, there have been debates about the appraisal of qualitative research, such as whether criteria are necessary, and what kind of criteria should be established (Holloway & Wheeler, 1996; Perakyla, 1997). As literature on the concepts, terms, and procedures of quality criteria for qualitative research accumulated, the dialogue on qualitative criteria continued. Until now, no consensus of any kind regarding quality criteria has been made, and it remains to be a question whether a consensus should ever be made.

As some scholars have pointed out (Dixon-Woods et al., 2007; Hannes et al., 2010; Walsh & Downe, 2006), it is not surprising that debate about quality criteria continues. The reasons are largely because of the naturalistic or context-dependent nature of qualitative research, in addition to the current state of methodological pluralism in qualitative research. First, brought by fundamental epistemological difference,[3] there is inevitable tension between some of the researchers who use qualitative research and those who embrace the concept of appraisal criteria. This is in part because, in qualitative research, knowledge and evidence are considered as a socially produced construct (Bryman, 2004; Denzin & Lincoln, 2011). Thus, the interaction and dynamics that are most valued and pursued in qualitative research are going against the "truth-seeking" standardizations implied in the idea of appraisal and evaluation (Easterby-Smith, Golden-Biddle, & Locke, 2008; Walsh & Downe, 2006). For instance, some researchers argue that a particular story is situated in a specific context, and many believe the essence of qualitative research is to compel the authoritative version and be open to more than one version of the phenomenon, which is exactly opposing the implication of criteria that aim to establish the "legitimate" or the "right" (Rolfe, 2000; Walsh & Downe, 2006).

Second, qualitative methodology values variety and plurality, and covers a broad range of philosophical positions in a wide spectrum. As a result, current qualitative research has a large range of types and forms of methods and perspectives (Buchanan &

3 Epistemological difference here refers to the difference in philosophical stances that define the nature of evidence and knowledge produced by an approach. For example, the positivist epistemology believes true and objective knowledge. In contrast, interpretivist and constructivist believe constructed knowledge and multiple realities. Generally, the philosophical stances of those who apply qualitative research reject the acceptance of one standard version as authoritative, which is in distinct contrast to the positivist stance.

Bryman, 2007; Easterby-Smith et al., 2008; Patton, 2003). It is therefore only natural that it is difficult to reach agreement on the issue of appraisal.

Faced with various and sometimes contradictory perceptions on the issue of qualitative criteria, it is not efficient to list all of them here. Thus the approach taken in this chapter is to present this issue in two parts: First to be presented is the evolution of the terms constructed for qualitative quality so as to review the general development of qualitative research. Next, current trends in defining the soundness of qualitative research are summarized.

Terms. The progression of terms used to describe the soundness of qualitative research largely reflected the development of qualitative research itself. Concerns with what is good qualitative research were raised in the 1980s, when qualitative methodology gradually gained power to become more visible in the social sciences (Barusch et al., 2011; Denzin & Lincoln, 2005; Lincoln & Guba, 1985).(Barusch et al., 2011; Denzin & Lincoln, 2005; Lincoln & Guba, 1985). But at the early stage of the dialogue on qualitative criteria, quantitative terms, such as reliability, validity, and generalizability, were borrowed to assess qualities of qualitative research (Marshall & Rossman, 2011), and qualitative research was evaluated with standardized criteria to minimize "human limitations" or "subjectivity" (Barusch et al., 2011; Breuer, Mruck, & Roth, 2002). Scholars have, however, questioned the appropriateness of using criteria based on post-positivist assumptions to assess qualitative research, as these assumptions "undermine the purpose and essence of qualitative research" (Cutcliffe, 2001, p. 376). [4] If standards based

[4] It is necessary to clarify that post-positivism is one of the theoretical traditions of qualitative inquiry, and post-positivists often adopt a "reality-oriented approach" to qualitative research (Patton, 2003, p. 94); thus, there is nothing

on post-positivist assumptions alone cannot align with an approach as broad as qualitative research, what could be used that recognizes the naturalistic axioms and interactive dimension of qualitative inquiry? To address such concerns, scholars like Lincoln and Guba (1985) discussed alternative constructs such as: Credibility, dependability, conformability, and transferability as quality indicators for qualitative research. Although Lincoln and Guba's terms are more or less parallel concepts of post-positivist concepts of reliability, validity, and objectivity, these new terms enacted the discussion of the basic concepts of criteria in the new context of a qualitative paradigm, and offered useful vocabulary to describe qualitative perspectives. In the past three decades, there has been a substantial body of literature on criteria of qualitative research, and a large array of terms created to indicate good qualitative research, such as "validity," "credibility," "trustworthiness," "rigor," "authenticity," "validation," and "goodness" (Lincoln & Guba, 1985; Marshall & Rossman, 2011; Maxwell, 1996; Patton, 2003). Developed from the basic concept of producing credible work using qualitative methods, these notions carry different but often overlapping connotations. To take a few examples, Patton (2003) adopted Lincoln and Guba's (1985) term "credibility" and developed it into an overarching concept for quality assessment of qualitative evaluations. Maxwell (1996), by contrast, retained the quantitatively oriented term validity. Incorporating validity with qualitative features, Maxwell refined criteria and made them applicable to qualitative research. Marshall and Rossman (2011) used the term "trustworthiness" as an umbrella concept to cover credibility, validity, and dependability all in one. The concept of

---

inherently wrong with post-positivists using qualitative methods. However, post-positivist assumptions alone, or for that matter any single philosophical orientation cannot capture all of what qualitative research purports to do.

credibility used in this study, which will be defined and discussed in more detail later in the chapter, is an overarching concept that has connotations drawn from many existing quality criteria, both from what qualitative criteria should be and should not be.

What qualitative criteria are not. In the following section, I will examine what is not desired in appraising qualitative research since it can offer a clear view of the trends that shape qualitative criteria.

First, leading qualitative scholars argued against universal criteria (Guba & Lincoln, 2005). This means that there should be no overriding criteria that fit every particular situation (Lichtman, 2010; Parker, 2004).[5] In a broader sense, universal criteria also refers to seeking one agreement or consensus on quality criteria (Bochner, 2000). Some scholars even went more extreme, pointing out that establishing regulative norms of good qualitative research is problematic (Guba & Lincoln, 2005; Schwandt, 1996). Yet despite the different levels of arguments against universal criteria, qualitative researchers generally agreed that it is critical for researchers to be clear about the criteria they adopted in adjudication and report such criteria explicitly.

Second, many qualitative researchers argue against fixed or static standards in judging qualitative research. This trend could have been triggered by the Scientifically Based Research (SBR) movement at the beginning of this century. Adopting an evidence-based epistemology, SBR promoted methodologies like experimental causal models, data replication, and generalization of results (Maxwell, 2004). Consequently, scientific based methods "include(s) the expectation that the studies are replicable," and experimental

[5] As indicated earlier, this synthesis does not intend to promote universal criteria either, but rather to understand what credibility techniques are used to enhance quality in program evaluations that use qualitative methods.

practice techniques and model building are seen to deserve federal funds and hold high value in social policy-making (AACTE, 2002; National Research Council, 2002). Historically speaking, this movement has, to a large extent, confronted the growing momentum of qualitative research being more inclusive and diversified. Qualitative researchers resisted the initiative by restating that the boundary of what is scientific research should not be mandated and hardened (Lichtman, 2010), and the traditional empiricist criteria were not helpful for qualitative studies (Bochner, 2000; Parker, 2004). As an extension to the external political backdrop, qualitative researchers and journal editors called for further examination of the quality issue internally, and suggested "move away from employing listing of static criteria to adjudicate and develop qualitative research" (Easterby-Smith et al., 2008, p. 419). Some argue that fixed criteria could limit innovation and risk valuing certain types of qualitative research at the expense of others (Lichtman, 2010; Parker, 2004). In response to the problem, some scholars encouraged more emphasis on flexibility and innovation in criteria building, and gave more attention to the links between quality, research process, and context (Denzin, 2002; Fade, 2003; Seale, 2002; Tobin & Begley, 2004).

Criteria becoming specific and generic. Given the argument that qualitative criteria should not be universal or static, what criteria can be used? In this section the features discerned and advocated in recent years are organized into to two different trends: Criteria becoming more specific and more generic.

A number of leading qualitative scholars have suggested that criteria for good qualitative research should be tied to different paradigms (Patton, 2003; Tracy, 2010).

Qualitative research should be assessed on its own terms, which take into consideration of its purpose, nature, and conduct (Kushner, 2005). Thus, a review of literature on criteria shows the trend of criteria becoming more specific to different philosophical frameworks, theoretical traditions, and qualitative communities (Creswell & Tashakkori, 2007; Ellington, 2008; Golafshani, 2003; Guba & Lincoln, 2005). Patton (2003) put forward different evaluative criteria for positivistic, constructivist, artistic, and evocative paradigms, respectively. Creswell (2012b) tailored evaluative criteria for each of the five different qualitative approaches: Narrative, phenomenological, grounded theory, ethnographies, and case study. Even for each qualitative method, interview, observation, or document analysis, there is a different set of standards (Creswell & Tashakkori, 2007; Marshall & Rossman, 2011; Silverman, 1993). In addition, criteria are becoming more specific in relation to individual field, discipline, or individual studies in order to meet specific needs (De Witt & Ploeg, 2006; Kushner, 2005). This trend, in turn, led to a large number of checklists and frameworks for particular kinds of research approaches.

Although specific criteria have their own advantage to help researchers find the standards that fit their particular theoretical community, they also yield difficulties. For example, it could be confusing, exhausting and even intimidating for novice researchers as well as those who are not very familiar with qualitative research to align themselves with a particular type of criterion. Furthermore, a set of inappropriately selected criteria could cause misjudgments and undesirable consequences; it would after all be problematic when criteria generated from one context are used in another. For example, if criteria for grounded theory research are used to examine a phenomenological study,

evaluative questions will be raised such as whether concepts are generated, or whether strong theoretical links between categories are established. This quality assessment will be missing the point since phenomenology is to capture and describe how people experience and perceive a certain phenomenon (Van Maanen, 1988), and instead of developing a theory that explains the process, actions or interactions of a substantive topic (Creswell, 2008), a credible phenomenological study should provide "an accurate portrait of the common features and structural connections" of the phenomenon (Polkinghorne, 1989, p. 57). In response to these emerging issues, there is the concurrent trend of qualitative criteria becoming more generic. Therefore on the one hand, criteria tend to be developed more flexibly to leave the evaluative decision more "local" to fit into the specific contexts (Barusch et al., 2011), but on the other hand, qualitative criteria need to be broad and abstract in order to embrace the general principles of quality check (Creswell, 2012b).

Having presented many of these complexities in assessing qualitative research, it is also to be pointed out that when focusing on program evaluation only, some of the complexities may be less of a concern, especially given the emphasis on credibility techniques. It is true that various qualitative approaches can be used in program evaluations, but all summative program evaluations deal with the central goal of determining the merit or worth of a program, which require the evaluators to provide evidence to support their claims. This is the point where credibility and credibility techniques are emphasized.

Credibility: Definition and Approaches

Definition. The key concept for this study, credibility, after initially articulated as an alternative construct for validity in a quantitative approach, has always been considered as one of the essential indicators for a quality assessment of qualitative research (Brantlinger et al., 2005; Cho & Trent, 2006; Creswell, 2009, 2012b; Cutcliffe, 2001; Glensne, 2011; Hannes et al., 2010; Johnson & Christensen, 2012; Lather, 1991; Lincoln & Guba, 1985; Lincoln et al., 2011; Maxwell, 2005; Onwuegbuzie & Leech, 2006; Patton, 1999, 2003). Evolving together with qualitative research itself, the notion of credibility has also endured an on-going process of being constructed, challenged, debated, defined, and redefined (Cho & Trent, 2006; Lewis, 2009). The definition of credibility applied to this study is one that seeks commonality in the variety of definitions of credibility in literature, and emphasizes three prominent features of the notion as being faithful, believable, and systematic.

First and foremost, credibility in this study is defined as a faithful description of the phenomenon of interest. Credible description and analysis should present an authentic and vivid picture of what is seen happening, and an accurate account and presentation of the findings (Beck, 1993; Huberman & Miles, 1994; Joint Committee on Standards for Educational Evaluation, 1988). It also indicates the efforts to achieve a sense of contextual "real." In order to capture the often complicated "realness," researchers recommended paying attention to the social and cultural context, the multiple voices, positions, and dynamics of different social and cultural groups, so as to render the "reality" recognizable for the informants and readers who have similar experiences (Beck, 1993;

Richardson, 2000; Tracy, 2010). Contextual and cultural sensitivity is therefore an important factor in addressing credibility.

Secondly, credibility means research claims and statements need to be convincing and believable (Kvale, 1996; Marshall & Rossman, 2011). In other words, credibility is the degree to which both information and interpretation are defensible (Johnson & Christensen, 2012). In order for the research to be trusted, any argument inferred from data should be strong and convincing, and any claims or statements made should be logical and based on well-grounded premises (Kvale, 1996).

Thirdly, credibility checks are often included as a part of the research design, as a systematic process throughout all stages of research, and it should be continued during report writing after completion of the research (Creswell, 2009). There is no cut-off point for this standard, and it can never be accomplished with one stroke. Kvale (1996) saw credibility as more than mere accumulation of techniques, but rather craftsmanship of the researcher. It requires theoretical questioning, knowledge communicating, and the practical action of credibility checking. For Patton (1999, 2003), credibility consists of broad elements: Rigorous techniques, trustworthiness of the researchers, and philosophical beliefs of the reader or users. Various authors have also constructed diverse typologies of credibility (or as some of them call it, validity), such as Maxwell's (2005) five types that cover description, interpretation, theory, researcher bias, and reactivity; Lather's (1993) four framings of validity including ironic, paralogical, rhizomatic legitimation, and voluptuous validity, which largely emphasize multiple representations rooted in postmodernism; and Cho and Trent's (2006) five purposes indicate that validity

is used for "truth" seeking, for thick description of the unique perspectives constructed by participants, for developmental over time, for the purpose of researcher's personal interpretation, and for the purpose of praxis/social change. Researchers generally choose their own credibility process, and reference the types of credibility system they adopt (Creswell, 2012a). The credibility checking or validation process differs when the fundamental questions vary (Cho & Trent, 2006). The purpose of the credibility checking process is not to compare one's own study with a set of independent gold standards as suggested in some areas of inquiry (Campbell, 1988; Putnam, 1990), but rather to build the "possibility of testing these accounts against the worlds, giving the phenomena that we are trying to understand the chance to prove us wrong" (Maxwell, 2005, p. 106). To clarify, the goal is never right or wrong, but rather to enhance understanding through the process of checking credibility or validity (Hesse-Biber, 2010).

Credibility of the researcher is another critical element of credibility, and Patton (2003) suggested that since the researcher is the research instrument in qualitative studies, in a qualitative study, the researcher's training background, perspectives, and relevant experiences in the field should be reported as part of the study's methodology. Credibility of the researcher could be addressed by "revealing the self" and revealing the "other" (Lichtman, 2010, p. 224). Self-revealing is a process of self-reflexivity, and according to Patton (2003), the principle is to report any information related to the researcher that may have affected the research, both positively and negatively. Thus reflexivity is, on the one hand, an approach typically adopted by constructivist analysts to deal with the concern of bias. On the other hand, it is a way to enhance the competence of the researcher in a

particular setting, because the competence of the researcher as the instrument is directly related to the methodological rigor and credibility of the research. In addition to self-revealing, the researcher should also reveal the "other" (the participants), and the interaction of the self and the other (Lichtman, 2010). One important factor involved is the investigator effects, or reactivity, that is, how the presence of the researcher may have affected what was happening in the field.

The focus of this research synthesis is credibility techniques. Such techniques may be only optional for certain qualitative research to promote credibility, but they are vital for evaluation studies. This in part because formal educational program evaluation standards developed by the Joint Committee on Standards for Educational Evaluation (JCSEE) as introduced earlier, have explicitly required evaluation works to provide evidence of credibility checks in their reports (Yarbrough et al., 2011). For example, for some qualitative research, credibility is embedded in the analyses and interpretation of the findings that are only "true for this place at this time" (House, 2005, p. 1070). For program evaluation, due to its need to assist decision-making of interventions or policies for similar situations, information concerning credibility needs to be fully documented in the report according to accountability standards of JCSEE, and expect external reviews of their work (Yarbrough et al., 2011). However, in this synthesis study, the researcher also strives to follow a holistic approach to assessment (Chen, 2005). That is, focus of the study is placed on the use of credibility techniques, but contextual information and general quality of the evaluation are also taken into consideration, including credibility of

the research methods and analyses, and the degree of details included to promote transferability (Patton, 2003).

Approaches to promote credibility. The existing literature documents many approaches to promote credibility, and to be concise, they can be summarized into three families: Transactional, transformational, and the middle ground approach (Cho & Trent, 2006; Marshall & Rossman, 2011).

The transactional approach can be considered as technique-oriented. For this approach, credibility is a transactional process consisting of techniques or methods by which misunderstandings can be identified and explained (Cho & Trent, 2006). In other words, procedures or techniques are considered as the medium to promote credibility (Marshall & Rossman, 2011). Many traditional credibility strategies can be seen belonging to the transactional family, such as Guba and Lincoln's (1985) credibility techniques, Eisner's (1991) emphasis on corroboration, and Maxwell's (1992) 'descriptive' and 'interpretive' approaches for validity. The goal of the traditional credibility strategies from the transactional family is to "operationalize" (Lincoln & Guba, 1985, p. 301) the quality standards of qualitative research. Examples of some of these credibility techniques will be discussed in the next section of this chapter.

The second family, or the transformational approach, encourages researchers to express dynamics and complexities of the conceptualization process, and gives central attention to researcher reflexivity and participant interaction Transformational approaches have a strong interpretive lens. Credibility in the transformational or transgressive sense can hardly be achieved by way of concrete methods or be done once for all (Cho & Trent,

2006; Lather, 1993; Richardson, 1997). It is rather an ongoing open dialogue on the topic of credibility (Agen, 2000), and a continuous challenge on the ideas developed during research (Whittemore et al., 2001).

There are also scholars who would rather take the middle ground, to not merely rely on either techniques or self-reflexivity, and instead they sought for flexible, useful, and integrated credibility approaches (Cho & Trent, 2006; Tracy, 2010). The "middle ground" approach has a few key emphases. Firstly, this approach requires the qualitative researchers to be holistically engaged in the specific territory of their qualitative inquiry. In this sense, the specific research paradigm, research purpose, questions, and the actual research process should be considered for a specific understanding of credibility and the means to address it (Creswell, 2012a; Creswell & Miller, 2000). Tracy's (2010) contribution of a holistic judgment of qualitative work is to introduce a set of "big tent" criteria to have an overall structure for quality while still attending to complex differences of various paradigms and genres. Another set of criteria is the one put forward by the Interdisciplinary Qualitative Research Subcommittee (IQRS), which examines the contribution of qualitative research from a mixed methods perspective (Nastasi & Schensul, 2005).

Secondly, the middle ground approach emphasizes the transparent reflective process, the process to keep thinking out loud about the researcher's concerns, safeguards, and contradictions (Cho & Trent, 2006). This dissertation takes this methodological perspective to describe and assess the use of credibility techniques. Credibility is comprehensively viewed and approached in this dissertation, to not only examine the

techniques for credibility check, but also take into account the overall methodological soundness of selected studies.

Credibility Techniques

Credibility techniques are used "to document the 'accuracy' of their studies" (Creswell, 2012b, p. 250). Credibility can and should be checked throughout the course of qualitative research, and techniques for credibility should be addressed in research design, data collection, analysis, and reporting (Maxwell, 1996; Whittemore et al., 2001). In this section, a list of major credibility techniques in the methodological literature is described and discussed, and the categories developed for the section of credibility techniques in the codebook are also based on the following discussion.

Triangulation. Qualitative researchers define triangulation as a search for converging evidence from multiple data sources, methods, theories, and investigators, and triangulation needs to be explained both when it occurs and when it does not occur (Brantlinger et al., 2005; Nastasi & Schensul, 2005; Patton, 2003). As it is suggested in the definition, there are mainly four kinds of triangulation: Triangulation of sources, investigator triangulation, theory/perspective triangulation, and methods triangulation.

Triangulation of sources, or data triangulation, is to compare and cross-check the consistency of information collected from different sources (Patton, 2003). These sources could be multiple informants. In the case of a school program evaluation, informants could be stakeholders of the program, administrators, teachers, students, parents and so on. Data collected with each of these groups can help gain insight into their perspectives on the school program, and results could be compared for agreement and disagreement of

perceptions. In addition, these sources could be information derived at different times and occasions with the same group of informants, such as their attitude in public or in private, or the differences in the expression when the question is asked earlier and later during an interview.

Investigator triangulation is to use multiple investigators to reduce the intrinsic bias that come from a single analyst (Denzin, 1989; Patton, 2003). In the data collection stage, this strategy means the phenomenon under study being examined by more than one investigator with the same qualitative method. That is, having more than one observer or interviewer to provide a check on bias. In the data analysis stage, investigator triangulation means two or more researchers independently analyze the same data and compare their findings. Discussion on how different investigators view the issue can help develop a broader and deeper understanding of the phenomenon. If the findings from different investigators arrive at the same conclusion, then we can be more confident in the credibility of the findings.

Theory triangulation means using different theoretical frameworks and perspectives for the same data (Patton, 2003; Thurmond, 2001). For example, when a researcher conducts a study on faculties in higher education, he/she may draw on a variety of theoretical paradigms from sociology, organizational research, and economics. If similar interpretations are generated using perspectives or theories from different disciplines, credibility is greatly enhanced.

Finally, methods triangulation involves using different methods to investigate the same phenomenon. It could be the use of multiple methods within the qualitative

approach, such as checking interviews against observation and other written documents (Johnson & Christensen, 2012). It also involves methodological triangulation that integrates both quantitative and qualitative data (Patton, 2003). For example, results from surveys, statistical analysis, and interviews could be compared to see if similar results are being found. The pragmatic approach of reconciling quantitative and qualitative data comes from mixed methods, and cross-method triangulation (multiple methods) is a key concept in mixed-methods work (Johnson & Onwuegbuzie, 2004).

Triangulation has been perceived as an important and typical methodological strategy to improve credibility of findings in evaluations (Guba & Lincoln, 1981; Mathison, 1988). Although triangulation is often defined as the strategy to seek converging evidence, it is impractical to always assume a convergence, or "a single valid proposition" being constituted as a result of triangulation (Mathison, 1988, p. 15). It is certainly desirable to achieve corroboration, or to build each piece of evidence into an integrated claim about the phenomenon under study (Eisner, 1979), yet convergence is only one of the possible outcomes, and certainly not the sole purpose of triangulation. Mathison (1988) pointed out that other possible outcomes of triangulation included having inconsistent data or contradictory data. Compared with achieving convergence, these relatively more complex outcomes may, first of all, be the actual state of affairs, and they can also provide a rich picture of the program under study. Second, the researchers' effort in searching for and constructing explanations for the inconsistency may deepen their understanding on rival themes emerging from the data, and help provide meaningful information. After all, the true value of triangulation lies in the

knowledge and the understanding of when, where, and why there are such differences (Patton, 1980, 2003).

Member checks. In qualitative research, realities are best described by both the participants and the researcher (Cho & Trent, 2006; Creswell, 2012a, 2012b). Member checks, or informant feedback, is a systematic procedure to share with participants one's data, analysis, interpretations and sometimes conclusions and to obtain their feedback (Creswell, 2012a; Onwuegbuzie & Leech, 2006). The first level of member checks, which is the most common approach, is to have participants review interview transcriptions or observational field notes prior to analyses, and them to confirm the accuracy (or inaccuracy) of the data (Brantlinger et al., 2005; Doyle, 2007), and the purpose is to find out "whether the data analysis is congruent with the participants' experiences" (Curtin & Fossey, 2007, p. 92). However, Creswell (2009) recommended that member checks are best done with already interpreted themes and patterns emerging from the data, so the second level is to seek feedback about analyses and interpretations (Brantlinger et al., 2005). In member checks, the participants play a major role, because member checking is not only the process of engaging participants in making sure their realities correspond with the interpretations brought forth by the researchers, but also a way to give power and voice to the participants (Cho & Trent, 2006; Doyle, 2007). Participants may disagree with researcher's interpretations, or change their mind on certain issues, which often occur in practice, but the goal of member checks is not to confirm a fixed reality, rather it provides the opportunity to detect possible

misrepresentations, and to negotiate and co-create the meaning of the issue under study with the participants (Maxwell, 1996).

Peer debriefing. Peer debriefing is to engage professional colleagues in analytic discussions and data interpretations (Lincoln & Guba, 1985). It is fair to consider peer debriefing a combined process of collaborative work and external evaluation of the research (Glesne & Peshkin, 1992; Maxwell, 1996; Newman & Benz, 1998). The peers are recommended to be professional colleagues with a similar status. They can be insiders, who are individuals with prior experience on the topic of research, and provide insights, review perceptions of analysis, or help develop the next steps of the research (Brantlinger et al., 2005; Johnson & Christensen, 2012; Lewis, 2009). They can also be outsiders, who have little or no exposure to the topic, and provide a fresh look and bring more questions regarding the study. It is usually even better to have both types of peers to ensure the most beneficial feedback. The key for the peer is to be impartial. They should take the position of the devil's advocate, and seek to engage the researcher in discussions or even arguments for solid evidence of the research interpretations and conclusions (Johnson & Christensen, 2012). When the researcher and the peer debriefers disagree, it is suggested to discuss and resolve the differences through honest communication (Barber & Walczak, 2009, April). The ultimate goal is not for the researcher and the debriefer to reach an agreement over interpretations, but, in the process of discussion, to challenge the research assumptions and be alert to researcher bias and alternative explanations.

Audit trails. Apart from feedback of participants and peer colleagues, detailed documentation is another essential element to promote credibility in qualitative research.

Detailed methodological description enables the reader to determine how far the data and constructs emerging from the data may be accepted (Lewis, 2009). Similarly, audit trails allow for a step-by-step trace of research procedures and the decision-making process. Audit trails refer to systematic documentation of all procedures and data relevant to the study (Lincoln & Guba, 1985; Onwuegbuzie & Leech, 2006). Halpern (1983) identified six classes of raw record, and they are mainly in two categories: One type is the "data oriented" records (Lewis, 2009, p. 72), including descriptions of interviews and observations, field notes, process notes, and so forth. The other type is theoretical audit trails, such as reflexive journals, finding synthesis products, or notes that describe how concepts emerged and how questions are pursued, which are more related to the development of ideas (Halpern, 1983; Lewis, 2009).

Negative case analysis. Negative case analysis is a search for disconfirming evidence after the preliminary themes and analysis categories have been established (Brantlinger et al., 2005; Lincoln & Guba, 1985). Searching and accounting for negative cases can greatly support credibility because reality in the constructivist sense is complex and never singular (Creswell & Miller, 2000). Negative case is not an exclusively qualitative concept; rather it is similar to outliers in the quantitative sense. The purpose of doing so is to expand and revise one's interpretation until all outliers are explained (Creswell, 2012b; Huberman & Miles, 1994; Lincoln & Guba, 1985; Maxwell, 1996). When evidence inconsistent with the themes is found, researchers should carefully examine the meaning of such cases and try finding explanations about it, as negative cases can provide valuable insights to the underlying phenomena (Onwuegbuzie & Leech,

2006). In other words, negative cases should be recognized, explained, and, when necessary, become the reason for modification of existing themes. In the cases when no reasonable explanations are found, it usually suggests further investigation or more data collection. In both scenarios, negative cases need to be transparently documented and limitations acknowledged (Creswell, 2012b; Huberman & Miles, 1994; Lincoln & Guba, 1985; Maxwell, 1996; Onwuegbuzie & Leech, 2006).

Thick descriptions. Thick description means the use of low-inference descriptors when writing the research report. It helps enhance the accuracy in description of the study, as a way of achieving qualitative transferability (Lincoln & Guba, 1985). By describing the phenomenon in sufficient detail, the reader can have their own evaluation about the degree to which the findings can be applied to other times, people, settings, and situations (Lewis, 2009). Furthermore, in order for readers to understand participants' perspective, researchers should report sufficient quotations to bring readers the experience with the participants' "actual language, dialect, and personal meanings" (Brantlinger et al., 2005; Johnson & Christensen, 2012, p. 267). In addition, thick description includes not only detailed descriptions, but also rich and in-depth illustration of the often complex culturally situated meanings of the data, so readers can come to their own conclusion about the scene with enough details (Tracy, 2010).

Prolonged engagement and persistent observation. Prolonged engagement and persistent observation are two intertwined techniques often used together. Both of them require investment of sufficient time and in-depth investigation to ensure accurate understanding of the phenomenon (Lincoln & Guba, 1985; Nastasi & Schensul, 2005).

The premises for these two techniques are that qualitative research as a natural inquiry is to document what is happening in the field rather than what has been put there for the researcher's benefit (Scott & Garner, 2013). Prolonged engagement and persistent observation, therefore, can help researchers construct the context in its natural state. During this prolonged period of time, the researcher not only learn the norms, characteristics of the participants and the phenomenon, but also learn their own role as a researcher in the environment under study, earn trust of participants and construct a deeper understanding about what is being studied. Substantive time spent in the field often becomes the basis for deciding what is and what is not important and/or relevant, as well as the basis for interpretation of the meanings of events (Ely, Anzul, Friedman, Garner, & Steinmetz, 1991; Lewis, 2009; Onwuegbuzie & Leech, 2006; Scott & Garner, 2013). Advantages of prolonged engagement and persistent observation include establishing good rapport between researcher and participants, shedding new light on old observations, and identifying emerging new questions and themes (Ely et al., 1991). Disadvantages largely involve reactivity, which is discussed in later sections.

Reflexivity. Reflexivity is another key strategy for researchers to better understand themselves and their research. The core concept of the definition of reflexivity is the idea of critical self-awareness. Researchers should be aware of the influence they have on research, their personal constructions of the world, assumptions, their values, beliefs, strengths, and weaknesses, all of which mold the research journey and the choices made (Hardcastle, Usher, & Holmes, 2006). Effort should be paid to monitor a researcher's influence and control personal biases (Giddens, 1984; Johnson &

Christensen, 2012). Researchers should also be aware of the impact the research process casted on themselves. This critical reflexivity of self is a continuous challenge, because the researcher has to face unknown challenges as they move on (Cho & Trent, 2006). The research process may not only change the research design and approach, but also alter how the researcher perceives the world. Burns and Grove (2001) stressed that when new aspects of the world are unveiled, researchers should be ready to cope with the change.

Apart from the techniques discussed above, there are other techniques that may help promote credibility. To give a few examples, credibility can be promoted using outside experts to assess the quality of the study (external auditors); carefully examining rival explanations (ruling out alternative explanations); comparing a series of predicted and actual results (pattern matching); considering design constraints (design check); and giving more weight to strong data and less weight to weak data (weighing the evidence) (Brantlinger et al., 2005; Creswell & Tashakkori, 2007; Marshall & Rossman, 2011; Onwuegbuzie & Leech, 2006; Patton, 2003). These credibility techniques documented in literature form the basis for the assessment of credibility techniques to be conducted in this synthesis, and the major characteristics of these credibility techniques were incorporated into the codebook.

Combined Use of Credibility Techniques

With major credibility techniques described and discussed, it is also important not to look at these techniques as separate strategies, but to understand them in a broad and connected sense. In this sense, combined use of credibility is connected with the "middle-ground" approach of promoting credibility: Multiple techniques are comprehensively

considered to enhance credibility of the study. Some examples are provided in this section to illustrate the combined use of triangulation with negative analysis, member checks, and reflexivity. As explained earlier, triangulation of multiple sources is about converging findings from different persons, time, and places. In that sense, searching for disconfirming evidence (negative case analysis) can be considered as another form of seeking a particular kind of information for finding convergence. More importantly, combined use of triangulation and negative case analysis can effectively control bias and enhance credibility, given that the disparate data sources in triangulation may be biased in the same way (Newman & Hitchcock, 2011).

In addition, the role of investigator or analyst in investigator triangulation can also be more than just the researcher. Patton (2003) suggested participants and audience be the analysts under the context of program evaluation. Participants' feedback on the research can not only confirm their own perceptions being accurately represented, but also contribute to the research findings, especially in collaborative and participatory inquiry (Patton, 2003). Audience, on the other hand, plays a critical role for establishment of credibility as they are the ultimate user of the findings, and naturalistic evaluations especially rely on the audience to reach their own conclusions and interpretations. As a result, triangulating the understandings of researchers, participants and the audience constitutes reflexive triangulation" (Patton, 2003, p. 561), which is also a demonstration of the close connections between triangulation, reflexivity, and member checks. Moreover, the analysts could also be research colleagues and external auditors. Both of these roles involve a disinterested expert to make judgments about the quality of the

research. This is naturally connected to the techniques of peer debriefing and external auditing.

Another example of combined use of credibility techniques could be the integrated use of reflexivity and member checks. Cho and Trent (2006) summarized this integrated relationship as reflexive member checking. In their point of view, member checking and reflexivity occur throughout the inquiry, and if member checking is a one-way process of bringing data and analyses back to the participants for perceived accuracy and reactions, then reflexive member checks refers to "the constant backward and forward confirmation between the researcher and the participants under study in regard to re/constructions of constructions of the participants" (Cho & Trent, 2006, p. 332). Similarly, reflexivity is not only a technique for researchers to illuminate a better representation of the lived experience of the participants, a way for researchers to openly express how their assumptions have been challenged and transformed as they collaborate and interact with participants for their construction and reconstruction, but also a way for the participants to differently perceive and impact their own lives (Cho & Trent, 2006).

Credibility Techniques and Evaluation Standards: Connections and Examples

So far, this chapter has reviewed a few key concepts: Program evaluation standards, quality criteria for qualitative research in general, and more specifically, the definition, approach, and techniques of credibility. In this final section, examples about the connections between evaluation standards and credibility techniques are provided. Both program evaluation standards and credibility techniques identify ways to enhance quality of findings, interpretations, and conclusions. However, evaluation standards

categorize and describe the dimensions of quality, but not the methods that can be used to address these dimensions. Complementarily, credibility techniques specify the strategies and procedures to be followed to check and promote quality. Therefore, in evaluation studies that use qualitative methods, evaluators should consider adopting various kinds of credibility techniques. Of course, credibility techniques offer relatively concrete outlines and procedures, but these techniques cannot define the exact practice in a specific setting. On-site response and researcher judgment are still required. Examples in this section (and in Appendix A) are presented to explain the dynamics between evaluation standards and credibility techniques, and to illustrate how credibility techniques can effectively support the qualitative research process in order to address different evaluation standards.

Among the five major attributes of the Evaluation Standards, the Accuracy Standards are designed to increase credibility of program evaluations. There are several shared connotations between the Accuracy Standards and credibility in qualitative research. First, accuracy in evaluation standards is defined as the truthfulness of evaluation findings and interpretations, as well as judgments about the quality of the evaluation studies (Yarbrough et al., 2011). Second, both qualitative credibility and Accuracy Standards emphasize the dependability of findings. Rather than to be achieved in a universal sense, truthful knowledge and evaluation propositions are to be established in a specific context that might change over time, place, and audience. Finally, in terms of the checking process, both concepts stress truthfulness to be achieved through "sound theory, methods, designs, and reasoning" (Yarbrough et al., 2011, p. 158). Therefore, given the largely integrated connotations, credibility techniques fit readily into the theme

of the Accuracy Standards and can serve as useful tools to identify and reduce inconsistencies, distortions, and misconceptions. For example, one of the supporting standards of the Accuracy attribute is A1 Justified Conclusions and Decisions. In the following paragraphs, different layers of meaning of this individual standard are explained to show the associations between quality goals required by the Standard, and the credibility techniques that can be applied.

The Justified Conclusion and Decision Standards require evaluators to keep in mind three aspects: The quality of information, the soundness of the logic that leads from information to findings, interpretations, and conclusions, and the plausibility of alternative interpretations (Yarbrough et al., 2011). Examples of the use of credibility techniques are to be explained in these three aspects.

First, with regard to the quality of information, there are at least three essential perspectives to approach quality data. To begin with, it is critical to include different types of data (Yarbrough et al., 2011). Different types of data can be collected through multiple data sources or by multiple methods (data and methods triangulation) to portray multiple realities and hear voices of different groups (multivocality). Furthermore, when diverse perspectives are collected, not all of them should be treated as equally useful. Evidence needs to be constantly compared to determine its relevance and quality (comparison and contrast). Evaluators also need to weigh participants' knowledge, social positions, and their experiences, as well as the circumstances of data collection to make decisions about the importance of data (weighing the evidence). Last but not least, many other credibility techniques can also be used to check quality. According to Miles and

Huberman (1994), there are good reasons to believe some data are usually stronger and more trusted than others. Such data include triangulated information, data collected through substantive observations, and data collected after spending sufficient time in the field (prolonged engagement and persistent observation).

The second aspect of the A1 standard is to have sound logic that leads from information to the findings, interpretations, and conclusions. Credibility techniques facilitate internal and external checks of the data processing logic. Internally, evaluators need to critically reflect on assumptions and interpretive frameworks applied in analysis, and disclose anything that might affect the evaluation results (reflexivity). Externally, logic and reasoning can be discussed with and checked by colleagues (peer debriefing), experts (expert review), and participants (member checks) to contextually define and justify the evaluation results.

Closely related to a sound logic of findings is the third aspect: The plausibility of alternative interpretations. When data suggest more than one explanation, or when there are exceptional cases that do not fit into any pattern or trends, credibility techniques such as negative case analysis and exploring rival explanations can help evaluators validate these alternative explanations and "deviant" cases. Patterns and cases that do not fit the majority of data could present challenges as well as fresh insights for the justification of evaluation conclusions.

Although the Accuracy Standards epitomizes the main contribution of credibility techniques, these techniques can be widely applied to the rest of the standard attributes. A typical example could be the *U5 Relevant Information* under the Utility Standards. To

optimize the relevancy of information, Yarbrough et al. (2011) indicated two aspects: The credibility of information and the ways through which information is obtained. In order to produce relevant and useful information, evaluators need to weigh the value of information. For example, data or information from an authoritative source may not necessarily be the most needed information, and triangulated data can provide more than one type of data to assist the decisions about data value. Similarly, easily accessible data are often not the most relevant data, and techniques like prolonged engagement and persistent observation can help evaluators gain access to information below the surface. Moreover, relevance and utility of information and findings are closely related to the perceptions of stakeholders and evaluation users. Therefore, the relevance of information should also be assessed by continued negotiation and discussion with different groups involved in the evaluation (e.g., member checks, external auditor, peer debriefing, investigator triangulation). Apart from paying attention to the credibility of information, another important but often neglected aspect of the *Relevant Information Standard* is the quality of the evaluation process, or the procedures adopted to collect and analyze the information. In this regard, the very basic credibility techniques concerning different research stages of appropriate design, explicit sampling, data collection, data analysis, and presentation of finding can be used to promote quality of the evaluation process. More examples of the use of credibility techniques to support evaluation standards can be found in Appendix A. The examples included in this section and in Appendix A are not meant to be exhaustive, but are presented with the intention to illustrate the key points.

To summarize, there are two key features in the connections between qualitative credibility techniques and the Joint Committee Evaluation Standards. First, multiple credibility techniques can be used to support one or more evaluation standards, such as above examples of standards A1 and U5. Conversely, single credibility techniques, such as triangulation, member checks, and peer debriefing, are often repeatedly used to promote different evaluation standards. Second, the connections between credibility techniques and the Evaluation Standards can also be viewed in a more holistic perspective. Credibility is an integrating theme across different evaluation attributes and their supporting standards. In addition, some fundamental credibility techniques regarding appropriateness and explicitness of evaluation design, data collection, analysis, and contextual descriptions are essential throughout the evaluation process, and may greatly affect the overall quality.

With some understanding of credibility and credibility techniques, the next section shifts the focus to methodology of this study. The following section describes the characteristics of the methodology, or the features of mixed methods synthesis.

Mixed Methods Synthesis

Methods of research synthesis were developed as one of the significant methodological advancements since the 1970s (Sandelowski & Barroso, 2007), and they have been used to synthesize findings of primary studies (Whittemore & Knafl, 2005). During the past decade, there has been renewed interests in synthesis research, partly in response to the proliferation but under-utilized qualitative study findings (Sandelowski & Barroso, 2007), and also largely due to the Evidence-Based Practice (EBP) movement,

which highlighted the need and value for synthesized research evidence to inform policy making and practice (Denzin & Lincoln, 2011; Heyvaert, Maes, & Onghena, 2011). Synthesis as a scientific research process plays an important role in understanding and disseminating current research knowledge and shaping future research (Cooper, 1982; Cooper & Hedges, 2009; Suri & Clarke, 2009). As the knowledge base of empirical studies expands over time, more researchers today rely heavily on syntheses to keep up with the current state of knowledge, and get directions for future research (Ahn, Ames, & Myers, 2012; Cooper, 1982). Furthermore, the role syntheses play in the decision-making/social policy domain is also becoming larger and larger (Cooper & Hedges, 2009). In the education realm where evidence is embraced as the basis for practice, useful syntheses provide evidentiary support for schools' adoption of educational programs and practices, and inform educational policy makers with research and evaluation evidence. Chatterji (2008) has summarized several efforts for educational program syntheses in different aspects. These efforts include U.S. Department of Education–sponsored initiatives like What Works Clearinghouse (WWC), Comprehensive School Reform Quality Centre (CSRQ), and the Best Evidence Encyclopedia (BEE); British government-sponsored Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre); the international Campbell Collaboration (C2); and academic journals that focus on reviews and synthesis research such as the *Review of Educational Research*.

A search of literatures reveals a variety of terms used to define different types of synthesis efforts, such as meta-analysis, meta-synthesis, systematic review, exemplary synthesis, meta-ethnography, and integrative review (Harden, 2010; Heyvaert et al., 2011;

Sandelowski et al., 2012; Suri & Clarke, 2009; Thorne, Jensen, Kearney, Noblit, & Sandelowski, 2004; Whittemore & Knafl, 2005). While all review approaches share some commonalities, they are disparate with different paradigmatic origins, designs, procedures, techniques, and distinct purposes, which have fully reflected the complexity and diversity of doing a thorough review (Whittemore, 2005). For example, meta-analysis emphasizes statistical integration of quantitative results, typically to increase the generalizability of the data (Suri & Clarke, 2009; Whittemore, 2005), while qualitative synthesis methods generate new insights and understanding from ideographic qualitative knowledge in order to enhance the utilization value and cross-case transferability of qualitative findings (Schofield, 1990; Smaling, 2003; Suri & Clarke, 2009). Recent developments in research synthesis also include integrative review that combines both theoretical and empirical data sources that are inclusive of both experimental and non-experimental research (Whittemore, 2005). The approach adopted for this dissertation is also one of the latest developments in mixed methods research and systematic review: The mixed methods research synthesis (Heyvaert et al., 2011; Sandelowski et al., 2006; Suri & Clarke, 2009). It well serves the purposes of this study to describe the current status of the field, and an integrated combination of both qualitative and quantitative approach allows comprehensive answers to the research questions.

Mixed methods research synthesis is, to use Heyvaert et al.'s (2011) definition, "a systematic review applying the principles of mixed methods research" (p. 4). This type of research synthesis is to systematically review the data and formally summarize the knowledge, and it entails approaches associated with both quantitative and qualitative

research (Sandelowski et al., 2006; Sandelowski et al., 2012). Therefore, designed from a mixed method perspective, mixed methods synthesis denotes features of a synthesis and mixed methods research, which will be discussed respectively in the following subsections.

Synthesis features. A synthesis is a particular kind of literature review that brings together existing studies on a specific question (Harden, 2010; Tashakkori & Teddlie, 2003). It is also widely identified as its own type of primary research that contributes to a given knowledge base via synthesizing existing studies and re-analyses (Cooper et al., 2009; Sandelowski & Barroso, 2007). As early as in 1971, Kenneth Feldman (1971) has stated that systematic integrative review (a particular type of synthesis research) "may be considered as a type of research in its own right—one using a characteristic set of research techniques and methods" (p. 86). In contrast to primary research in which data subject to analysis are collected from participants, data in a synthesis are often results and findings of empirical research on a particular topic (Tashakkori & Teddlie, 2003). There are two distinct features of a synthesis, which are also reflected in this study: A synthesis protocol and the principle of transparency. First, a protocol is established and made explicit to map the steps to be taken (more detail in Chapter 3, and see Appendix B for the protocol). A synthesis protocol establishes and documents, in advance, the methods that will be used to undertake the review. A pre-defined protocol is necessary prior to knowledge of the available studies to help reduce research biases and potential duplication, and promote transparency of methods (EPPI-Centre, 2006; Higgins & Green, 2011; Kitchenham, 2004). A synthesis protocol usually includes synthesis questions, the

standard stages and procedures of searching and screening for studies (sampling), data extraction and coding (data collection stage), and the data analysis and synthesizing stage (Albainese & Norcini, 2002; EPPI-Centre, 2006; Harden & Thomas, 2010; Whittemore & Knafl, 2005). A protocol also includes descriptions and explanations of the conceptual framework and methods used to conduct the review. The second key principle synthesis research follows is the transparency approach: Every component of the synthesis protocol is explicitly described, such as the search strategy, selection criteria, analysis process and synthesis methods (EPPI-Centre, 2006; Whittemore & Knafl, 2005), and characteristics of each identified study is critically documented on a common scale (Thorne et al., 2004).

Keeping details of a planned review transparent can encourage informed criticism and transparency. The use of systematic procedures can also help assess bias in individual primary studies that are included in the review, as well as to enhance credibility of the overall synthesis project (EPPI-Centre, 2006). For example, when review methods are defined and put forward at the start of the review, the actual review process is less likely to be overly influenced by the results (EPPI-Centre, 2006). When a comprehensive search or even an exhaustive search is performed to collect primary studies on a particular topic, the conclusion of the review is less likely to be overly influenced by the knowledge of a particular group of authors, or by the most accessible research (EPPI-Centre, 2006). These features, or core principles (the scientific rules, transparency, and emphasis on structured approaches to minimize bias), have defined the research synthesis as a piece of research on its own. Sandelowski and Barroso (2007) also provided explanations of research synthesis as "a form of scientific inquiry" (p.19) by itself, regardless of the

overlaps it has with many other forms of inquiry, and how it is different from other entities like literature review or secondary analysis. For example, research synthesis is an "inherently different" approach (Thorn et al., 2004, p. 1346), if not advancement from a traditional literature review (Harden, 2010; Thorn et al., 2004). On the surface, similar procedures are applied to both a literature review and a synthesis, such as searching, selecting, and summarizing research findings. Yet there are critical differences in terms of research goals and procedure requirements between the two kinds of reviews. One of the differences is that the transparency requirement to reduce bias in research syntheses presents how the work is done, which allows more details for the reader to make judgment about the quality of the work (EPPI-Centre, 2006). Another major difference is that although both literature reviews and synthesis are considered as reliable sources of research evidence that present the accumulation of knowledge, the standard rules that guide systematic review procedures make this approach not only more reliable, but also replicable and updatable. Finally, and perhaps most importantly, compared to literature reviews that largely summarize knowledge, synthesis research creates new knowledge. Specifically, they address a formal research question and add new understanding to the phenomenon. For example, integrative reviews not only pull together effect sizes of existing studies on a topic, but also describe how the topic is conceptualized in literature, and analyze how such conceptualization has shaped the scholarship. A methodological synthesis can take the issue of methodology beyond a summary of methods and procedures employed in literature, and explore how methods constrain or open up opportunities for the issue under investigation (The Review of Evaluation Research,

2013). The journal *Review of Evaluation Research,* which publishes critical reviews of research literature, has explicit standards on the quality of the literature in systematic reviews and the quality of analysis (The Review of Evaluation Research, 2013). Syntheses of different kinds are required to go beyond descriptions and include analyses and critiques, and add new findings to the empirical knowledge base.

The mixed method research features. In this study, mixed methods perspectives are applied in the synthesis level, and the purpose of adopting such an approach is to combine the strength of both qualitative and quantitative approaches to provide evidence of research practice, and to identify strong and weak points in the reporting of credibility techniques in written qualitative evaluation reports. The study is "mixed" in terms of both "the objects of synthesis and the mode of synthesis" (Sandelowski et al, 2012, p. 317-318). That is, for one thing, evaluative studies that were included as data were either primary qualitative stand-alone research or mixed methods studies. For another, both qualitative and quantitative approaches were used to describe and analyze the data for different research questions and sub-questions, while results were integrated to draw the final conclusion. In addition, it is to be made clear that although research syntheses are more often viewed as post-positivist than constructivist in their philosophical position (Harden, 2010; Suri & Clarke, 2009), the author takes a pragmatic stand on this study. It is argued that pragmatism offers a useful middle position in terms of philosophical stands and methodologies (Johnson & Onwuegbuzie, 2004). Embracing compatibility of both quantitative and qualitative paradigms and a mixture of methods and procedures, taking a pragmatic perspective could be productive for this synthesis research to not only allow

practice-oriented methods and yield more immediate and practical outcomes, the mixed approaches can also facilitate fundamental understanding of the credibility issue from the perspective of cross-method triangulation (Johnson & Onwuegbuzie, 2004; Johnson, Onwuegbuzie, & Turner, 2007). In the context of this synthesis, a pragmatic perspective means that this study should lead to conclusions that, to some extent, can serve as the evidence for a better understanding of the field, inform, and improve the current practice (Sandelowski et al., 2012). At the same time, I also want this synthesis to be a departure point of a journey of getting to know better about the qualitative evaluation domain, and to have an interpretive side of stimulating debates and conversations. Therefore, as the philosophical stance being reflected in the design, the synthesis had both a pre-defined instrument and elements of emergent design; and the analysis included a standard scoring scale with numerical results and thematic coding for interpretations.

The mixed method research quality criteria. As discussed in earlier sections, this study is characterized as a mixed methods study, it is then also necessary to introduce quality criteria and strategies to promote methodological rigor in the mixed methods setting. In this section, different sets of criteria for assessing mixed methods research are briefly summarized to serve two major purposes: First, mixed methods research as a growing field encounters more complexities in assessing the quality of combined qualitative and quantitative approaches than each of these approaches assessed alone (Onwuegbuzie & Johnson, 2006), and constant attentions need to be paid to validities of both qualitative, quantitative approaches, and the integrated inferences in mixed methods research. For this reason, current discussions on mixed methods quality standards were

introduced. With this context in mind, the researcher examined the implementation and the overall outcome of this synthesis study with these criteria of mixed methods research, and relevant strategies were adopted to promote the practice and quality of this study. Second, these quality criteria are not only standards that the researcher used to assess and enhance her own mixed methods research, they also inform the analyses of mixed methods studies in the synthesis sample. Although this study focuses on qualitative methods in program evaluations, the sample selection was open to qualitative methods performed in a mixed methods setting. The focus of the analyses were placed on credibility of qualitative applications, but qualitative methods used in a mixed methods study were analyzed slightly differently from qualitative stand-alone studies. In spite of the already set-up categories, more considerations were given to the blending of qualitative methods with quantitative methods in a mixed methods setting.

Quality criteria of mixed methods share a few features with quality standards of qualitative methods discussed at the beginning of this chapter. First, there are quality concerns of mixed methods research and critical needs to promote validity (Tashakkori & Teddlie, 2003, 2010), and second, there are controversies on what is "good" mixed methods research and what constitutes methodological criteria for mixed methods research, or how to assess and improve quality of mixed methods studies (Creswell & Plano Clark, 2011; Johnson & On 2006; O'Cathain, 2010). As a result, quality criteria and choice of strategies to enhance quality of mixed methods vary from different philosophical views, definitions of validity, and typologies of mixed methods design (Creswell & Plano Clark, 2011; Dellinger & Leech, 2007; Greene, 2007). The following

criteria are important and recent developments on the issue of validity in mixed methods studies.

Teddlie and Tashakkori (2003) suggested using the term inference quality to replace validity in mixed methods studies, in order to highlight the features of mixed methods and be more specific in connotation. They put forward a framework of inference quality that is constituted of design quality and interpretive rigor. Later, Teddlie and Tashakkori (2009) further developed their framework of inference quality in mixed methods research with more details. The aspects related to design quality include design suitability (whether the design can appropriately address the research question), fidelity of study procedures (whether the study is implemented appropriately), within-design consistency (consistency of different aspects of the study), and analytic adequacy (the appropriateness of analytic procedures and techniques). Interpretive rigor of the meta-inferences consists of interpretive and theoretical consistency (the consistency of study inferences with theory, study findings, and with inferences of different method strands), interpretive agreement (consistency of interpretations across the researcher, participants, and other scholars), interpretive distinctiveness (how different the interpretations are from other possible explanations), and integrative efficacy (the adequacy of the meta-inference).

Alternatively, Onwuegbuzie and Johnson (2006) used the term legitimation to describe quality criteria in mixed methods research. They suggested a framework of legitimation to include the following nine types of legitimation: sample integration legitimation (the degree to which the quantitative and qualitative sampling were

incorporated to yield the meta-inference), inside–outside legitimation (the utilization of insider and outsider views), weakness minimization legitimation (how the weakness of one approach is minimized by the other approach), sequential legitimation (how threats to validity are minimized if the sequence of qualitative and quantitative phases are reversed), conversion legitimation (how quantitizing or qualitizing can yield quality meta-inferences), paradigmatic mixing legitimation (relevant paradigmatic beliefs of the researcher are combined or blended), commensurability legitimation (the commensurability of qualitative and quantitative world views), multiple validities legitimation (addressing the quantitative, qualitative, and meta inference validities), and political legitimation (challenges on how consumers of mixed methods research value the meta-inferences).

Both of Teddlie and Tashakkori's (2003, 2009) integrative framework and Onwuegbuzie and Johnson's (2006) legitimation framework emphasized assessing quality/validity of qualitative and quantitative strands with respective quality criteria, as well as assessing the degree to which the overall inferences can be trusted. Teddlie and Tashakkori put focus on minimizing inconsistencies of each strand, while Onwuegbuzie and Johnson pointed out integrating both methodological strands by addressing multiple validity legitimations. In addition, Onwuegbuzie and Johnson also connected the legitimation model with different stages of research process, and with different typologies of mixed methods design, such as in concurrent, sequential, conversion, parallel, or fully mixed designs.

Dellinger and Leech (2007) also stressed validity in research process in a similar but different way. Their conceptualization of validity in mixed methods studies is defined as a discourse of data meaning. Dellinger and Leech reconstructed the notion of construct validity as a way to perceive quality issues in the mixed methods setting, and proposed a validation framework to promote quality by the means of continuous negotiation of data meaning. The validation framework includes foundational element (researchers' prior understanding of the phenomenon), inferential consistency audit (consistency and appropriateness of study inferences with prior understanding, theory, and literature), utilization (appropriateness of the use of measures and inferences), and consequential element (social acceptability of consequences of study findings and inferences). These elements are organized to follow the different phases of the research process, and therefore can be a useful guide for research implementation.

Last but not least, Pluye, Gagnon, Griffiths, and Johnson-Lafleur (2009) developed the "minimum set of criteria" (p.533) for mixed methods synthesis reviews. Their criteria of quality appraisal include "justification of the mixed methods design, combination of qualitative and quantitative data collection-analysis techniques or procedures, and integration of qualitative and quantitative data or results" (p.540).

As recommended in the various frameworks and criteria of mixed methods research, validity checks were applied in collecting and analyzing data for both quantitative and qualitative strands in this synthesis study. Strategies used for credibility checks were discussed in Chapter 4. However, in analyses of selected evaluation journal articles, when both qualitative and quantitative approach are combined, qualitative part of

the study remains to be the focus, and qualitative criteria are applied in coding but with consideration of the role qualitative methods in the overall design.

Summary

This chapter set the stage by introducing the background knowledge and presenting the general context of this synthesis study. Essential concepts of program evaluation standards, qualitative and mixed methods quality criteria, credibility and credibility techniques, and mixed methods research synthesis are described and discussed to familiarize the reader with the different areas covered in this study, and to point out the different characteristics and current development status of each area. Then, to take a step further than simply present the literature, this chapter also reveals the interconnections between the seemingly distinct areas of program evaluation and qualitative approach. The imperative needs and advantages of combining the two areas are emphasized, especially the connections between the appraisal standards of evaluations and credibility techniques.

Apart from laying a foundation for the context of this study, this chapter also highlights the fundamental elements in evaluation standards and credibility techniques, which not only formed the basis for the construction of the codebook to be used in this study, but also became guiding principles for conducting the synthesis.

When qualitative inquiry is used in evaluation studies, credibility techniques are essential in promoting the quality of research implementation and findings, and to meet the evaluation standards. However, there is no empirical understanding of the use of credibility techniques in evaluation studies. So this synthesis study aimed to address this gap by starting with evaluating the practice of top educational evaluation journal articles.

Findings of this study could be informative to qualitative researchers, evaluators, journal reviewers, and consumers of program evaluation findings about the current state of practice of qualitative inquiry in the field of educational evaluation.

Chapter 3: Research Design and Methodology

Introduction

This study is a synthesis that examines reports of credibility techniques that evaluators have used to enhance the rigor of their qualitative inquiry in the following journals: *Educational Evaluation and Policy Analysis; Research Evaluation; Assessment and Evaluation in Higher Education; Educational Assessment, Evaluation and Accountability; American Journal of Evaluation; and Evaluation Review,* from 2003 to 2012. This synthesis effort specifically addresses the questions: (a) To what extent are credibility techniques present in the selected journals? (b) What are the features that can be observed in the reporting of credibility techniques? This Chapter describes the methodological framework, the data collection procedures, and the development of the instrument of this study: The codebook.

This study applies a mixed method research synthesis design to describe and interpret evaluation studies that use qualitative methods. A mixed methods research synthesis design was chosen because it best meets the objectives of this study. First, synthesis research generates new knowledge, and this study posed research questions that have not yet been answered. As it is stressed in the previous chapter, synthesis research is a piece of research on its own. Many people equate *research syntheses* with literature reviews or standard *meta-analyses,* but this study is more than a literature review or a pooled quantitative results. Synthesis research serves as an efficient way to achieve the goal of describing and understanding the "present state of the science" (Whittemore, 2005, p. 546). The comprehensive retrieval and rigorous procedures of a review of the

relevant reports in the target domain can allow depiction of a general picture of what has been done and how it has been reported.

Second, a mixed methods approach was adopted because neither qualitative nor quantitative approach alone can fully capture the goals of this study. The first research question (and related sub-questions) focuses on the presence and frequencies of credibility techniques: What are the credibility techniques used in practice, and how commonly used are they? These questions could be answered by a quantitative summary of frequencies, as many other similar studies have done before. However, this synthesis study went farther and addressed the supporting details of credibility techniques used: Are the credibility techniques vaguely mentioned or thoroughly described? How are these techniques reported? Answering these questions requires critical interpretation, and cannot be answered by quantitative approach alone. Similarly, a qualitative approach alone could only partially answer the research questions, and common qualitative review approach such as meta-synthesis does not fit this study, mainly because this synthesis aims to critique credibility methods used in evaluation reports by using a pre-defined codebook, instead of combining qualitative findings. Thus, to fully answer the research questions and related sub questions, the researcher adopted a mixed methods approach to quantitatively describe the use of credibility techniques in selected journal articles, and synthesize *how* the techniques were used. Furthermore, different sources of information were combined to make larger meta-inferences and commentary about the reported use of credibility techniques in education program evaluation studies.

Of course, despite the general features of mixed method synthesis research discussed in Chapter 2, the purpose and focus of this particular study render some methodological features of its own. The objective of this study is to examine the use of credibility techniques in evaluation journal articles, to address an empirical research question in practice. It therefore needs to be clarified that, unlike the majority of research syntheses that aim at aggregation of research findings, the central task of this study is to understand the use of credibility techniques. The goal is to highlight the strengths and weakness of the techniques employed, and how the use of these techniques has contributed to the credibility of research findings. Moreover, although the composition of journal articles included had both qualitative and mixed method studies, this study focused on examining the qualitative component in the mixed method works.

As discussed in Chapter 2, one of the distinct features of research synthesis is to develop a synthesis protocol, and stress transparent methods to explicitly define the review question, search strategy, selection criteria, data extraction, and synthesis methods (EPPI-Centre, 2006; Harden & Thomas, 2010; Suri & Clarke, 2009; Whittemore & Knafl, 2005). The protocol of this synthesis has five stages of specifying the topic, data collection, developing a coding scheme, analysis, and interpreting and reporting. In the following sections, steps taken within each stage are specified and described. A simplified list of major steps can be summarized as follows:

1. Specify the topic area, including the problem to be addressed, specific research questions, and specify the types of studies to be reviewed (addressed in Chapter 1);

2. Develop inclusion and exclusion criteria for studies in the synthesis;

3. Specify the search strategy;

4. Develop a management strategy;

5. Develop a coding protocol for coding studies;

6. Develop an analysis strategy.

Data Collection

The stage of data collection includes the procedures of searching, screening, extracting, and management of evaluation studies. This section presents details of procedures for searches of articles, inclusion and exclusion criteria, and how data were extracted, stored and organized.

Journal selection. A preliminary search of the evaluation literature identified 23 relevant journals. As this synthesis study focuses on education evaluation reports, the identified number of journals is narrowed down by selecting only journals that focus on education settings. Other factors are also considered, including a journal's ranking in SCImago Journal Rank indicator (SJR) scores published on the Scopus website in 2012, the Journal Impact Factor (JIF) 2011 Journal Citation Reports, and recommendations of evaluation experts. JIF is the oldest and perhaps the most well-known ranking system that ranks journals based on frequencies of articles in a journal cited within a given period of time (Falagas, Kouranos, Arencibia-Jorge, & Karageorgopoulos, 2008). A journal with a relatively high JIF can be considered to have received more citations than a journal with a lower JIF. Another important alternative journal evaluation tool, the SJR indicator is also used in order to have a balanced account of journal rankings. SJR scores use size-

independent metrics that aim at measuring the current average prestige of journals,[6] which means SJR indicator uses a weighted metric that takes into account both citation frequency and the prestige of the journals making the citations (Guz & Rushchitsky, 2009; SCImago Journal and Country Rank, 2013). Therefore, journals that rank high in both JIF and SJR scores should enjoy relatively high prestige and citations, and therefore have relatively stronger impact.

After comprehensively considering different parameters of this synthesis study and different features of identified evaluation journals, six core educational evaluation journals were selected: *Educational Evaluation and Policy Analysis; Research Evaluation; Assessment and Evaluation in Higher Education; Educational Assessment, Evaluation and Accountability; American Journal of Evaluation; and Evaluation Review* (also see a list of the six journals with impact factor and SJR scores in the Appendix C). All six journals have boards of external reviewers and utilize a blind review process. It was decided to use peer-reviewed journals as the sampling frame as they are presumed to provide examples of evaluation practice that is among the best practice. Comparatively, the quality and academic integrity of articles published in these journals are checked through rigorous review processes. Overall, the selected six journals can be considered as leading evaluation journals in the education field, and if credibility techniques represent meritorious practice as indicated by AEA standards and methodology literature, then it is reasonable and necessary to understand their reporting in leading evaluation journals.

[6] Computation of the SJR indicator uses an iterative algorism that distributes prestige values among journals. The rationale of the metric is to use a function that takes into consideration both the quantity and quality of the citations.

Article selection. This synthesis applied an exhaustive search strategy to identify all relevant articles published in these six journals. Three basic inclusion principles are applied as follows:

1. *The content criterion*: The study includes only empirical program evaluation studies that address education related topics. By empirical it means that evaluation studies that use primarily collected data are included, and studies that consisted only of secondary analysis of data collected in another study are excluded.
2. *The methodological criterion*: Program evaluation studies must use qualitative methods to be included. Both qualitative stand-alone studies and mixed method studies are included. As mixed methods have been applied in evaluation works more than two decades ago (Greene, Caracelli, & Graham, 1989), one of the advantages of including mixed methods evaluation studies, compared to including only qualitative stand-alone evaluations, is that a more comprehensive and complete understanding of the use of qualitative research methods in different paradigms and designs can be reached.
3. *The temporal criterion*: This parameter is set to the recent decade between 2003 and 2012. This time period reflected an increased use of qualitative methods in program evaluation. As one of the significant methodological trends emerged since the new century, expanded use of qualitative methods, especially combined with quantitative methods in mixed method design are seen in evaluation studies. According to Hogan (2007), There was a recent trend of

methodological shift toward combining quantitative and qualitative methods, and multiple-method evaluations received more acceptance and preference in recent years. A ten-year span was chosen because it has been suggested that methodological approaches tend to be stable within a 5-year span (Goodwin & Goodwin, 1985; Hutchinson & Lovell, 2004), so examination of articles published during a ten-year period allows an adequate observation of current methodological practice (especially practice of credibility techniques) and possible observation of changing trends in methodological approaches.

Searching procedures and data extraction. Journal articles are retrieved from each journal's online website. The searching of qualitative evaluation reports combined keyword searching and hand searching to ensure complete recall within the selected journals (Sandelowski & Barroso, 2007). The first round of searching used keywords in online search engines of each journal's official website. There were 21 keywords including concepts related to qualitative research, credibility, validity, and specific credibility techniques reported in literature (see Appendix B for a complete list of keywords). Searching key words were developed to keep a balance between sensitivity and specificity, to include both broad terms like "qualitative" and "evaluation" to find all articles in the topic area, and terms like "member check" and "triangulation" to locate articles that are relevant. The second round of searching was hand searching to scan through each article in each of the six journals to make sure an exhaustive search is performed.

A research synthesis requires a diverse array of resources (Major & Savin-Baden, 2010; Sandelowski & Barroso, 2007), and at the same time, it requires the management of massive resources and data. An administrative system was established for data extraction, storage, index, and the overall data management. First, identified articles in the initial search were located and downloaded from official websites of selected journals, and electronic records of references, including an electronic copy of the article saved in the reference management software EndNote X4. Then, decisions about the relevance of each study and reasons for its inclusion or exclusion were recorded in an excel sheet. Primary review of the identified articles and primary summary and description of the sample were organized and recorded in a separate Excel sheet.

The Codebook

The primary impetus for developing a codebook was to use it as a template to identify, describe, and evaluate credibility strategies of selected empirical program evaluations using qualitative methods in the six leading evaluation journals, and it also provided a basis for reliability checks. Given the increased use of qualitative methods in mixed methods and used alone in evaluation studies (Sirriyeh, Lawton, Gardner, & Armitage, 2012), there is great value in having a relatively broad form of assessment that can assess a diversity of sources of evidence using qualitative approach. Therefore, it is not the purpose of this codebook to be tailored to evaluate individual qualitative studies based on a particular epistemological approach, but rather to map the reporting of credibility techniques in a diverse range of designs within the evaluation world. In addition, although this codebook aims for a broad form of quality assessment, it still

focuses on credibility techniques. The codebook was used to comprehensively describe and assess credibility techniques in detail, instead of offering an overall judgment of every aspect of the selected studies.

Construction of the codebook. In terms of format, the codebook is essentially more of an appraisal template than a list of labels assigned to data in coding interview transcripts. The structure of the codebook combined a checklist with an open-ended comment section. As stated earlier, the codebook was not developed with the intent to promote universal criteria, and it has the focus set on credibility techniques instead of every aspect of quality assessment, so it is not a checklist for the overall quality, but for the reporting of credibility techniques. Rather, the codebook was used to understand how credibility techniques were used in each of the selected studies. The researcher sought a balance between appraising the overall methodological soundness and the use of credibility techniques. A checklist format can help provide clear guidelines for assessing the general soundness of reporting the use credibility techniques in evaluation studies, and the pre-specified classification and categories can help entail a systematic review of studies and allow the researcher to focus on selected aspects. Meanwhile, the open-ended comment section can document idiosyncratic ways in which evaluations are conducted and offer flexibility in appraising different qualitative approaches (Dixon-Woods, Shaw, Agarwal, & Smith, 2004).

The basic component of the codebook may seem clear-cut and stable, but the process of developing the codebook is iterative and dynamic. Building on the procedures recommended by MacQueen, McLellan, Kay and Milstein (2009), and Boyatzis (1998),

the development process includes developing an initial code list, circulating proposed codebook categories for review among experts of the field, refining details of the codebook, and depending on the acceptability of the consistency, to either continue with another round of refining of the codebook, or proceed to conduct a pilot study to try out the codebook, and finally assessing consistency of the coding application before starting the main analysis.

The coding list. The coding list defines the appraisal domains of the codebook, such as program information, type of evaluation, general methodological components, and optional credibility techniques. The list is constituted of a number of key categories that function as frames to systematically spot and map the information in the text that reflect the synthesis review questions. The development of the coding categories was an iterative process that included constant visit and revisit and comparison of different concepts mentioned by different author, and repeated examination of the data (DeCuir-Gunby, Marshall, & McCulloch, 2011).

Categories in this codebook were built upon three sources. First, it mainly consisted of theory-driven categories developed from the existing concepts of credibility criteria and credibility techniques discussed in the literature, such as triangulation, member checks, and negative case analysis. Then, a range of available tools for quality assessment of qualitative and mixed method studies used in practice are reviewed, including eight checklist tools identified and reviewed by Hannes and Macaitis (2012) for assessment of qualitative rigor, the on-line critical appraisal instruments for qualitative research developed by Joanna Briggs Institute (JBI), Critical Appraisal Skills Program

(CASP), the Evaluation Tool for Qualitative Studies (ETQS) (Hannes et al., 2010), appraisal tools for mixed method research (Atkins, Launiala, Kagahn, & Smith, 2012; Pace et al., 2012), codebooks (DeCuir-Gunby et al., 2011; MacQueen et al., 2009), and frameworks for assessing qualitative research and evaluations (Barusch et al., 2011; Spencer et al., 2003; Walsh & Downe, 2006). Many common features that indicate general credibility shared by these tools were drawn and incorporated into the codebook, such as explicit statement about the sampling procedures, data collection and analysis process, which are consistently present and attached with importance in most of the existing tools. Apart from the two sources for a priori categories, the coding list also included a number of categories that emerged from the raw data in the process of reviewing a subsample of evaluation studies during calibration and pilot study.

Structure of the codebook. The codebook consists of three sections: General study characteristics, design related techniques, and credibility techniques (or methodologically driven techniques) (Liao & Hitchcock, 2012). The first section was created for the purpose of capturing the basic characteristics of the selected journal articles, such as the types of evaluation, the general type of the program and primary methods adopted, so the evaluation work reported in the article can later be categorized to different groups accordingly. The second section covers the fundamental elements of doing qualitative research in different stages of design, implementation and presentation of findings. In other words, the items in the second section are considered as the most basic techniques to satisfy the primary quality criteria of qualitative evaluation research. The last section includes the major credibility techniques drawn from the techniques

defined and discussed in literature (a complete list of credibility techniques can be found in Table 10 and Appendix E). In addition to spotting the presence of these techniques and assessing the grade of evidence provided to support the use of them, the codebook was also used to code the language choices of authors in using credibility terms, whether they explicitly mentioned the terms, or practiced the technique without clear indication of the terms.

Apart from a small number of open-ended items designed to capture information about the program or intervention, the codebook predominately consists of low inference, categorical items. A "0" represents absence of the technique or characteristic, "1" represents its presence, and "2" represents presence of the technique unclear given the evidence provided. Each coding category (item) in the codebook is provided with a brief name and descriptions of each category. Coding examples, ruling decision details, and guiding questions are also captured. A number of items were designed to be multiple selection items, and combination of codes were considered in analysis.

The above phrase "low inference items" means that these items require little judgment; coders simply need to indicate whether certain elements are present in an evaluation report. However, when a technique is present, there is a follow-up check to be documented in the comment section on whether associated claims seem to be supported given the information presented in the report, the relevant original text, and then an overall assessment on the use of the technique. Although such judgment was guided by detailed decision rules, it still requires a comprehensive summary of the information and a greater degree of inference from the coder. The additional "coder's note" column

allows the coder to make comments, document specific information about the evaluation work, and provide interpretations of the meaning or analysis.

Compared with templates used by Shek et al. (2005) and Barusch et al. (2011), who conducted similar studies that examined methodological quality of evaluation studies, this codebook offers a more comprehensive list of coding categories focusing on qualitative credibility techniques in a more systematic structure. Firstly, both Shek et al. and Barusch et al. studies, as well as this synthesis research combined general quality criteria with specific credibility techniques as coding categories. The current study, by contrast, separated these categories into two sections in this codebook. By doing so, the codebook distinguishes the very basic methodological techniques required for credible research from optional techniques that might be used to enhance rigor. For example, describing sampling strategy and sampling procedures are basic techniques, but member checking could be an optional technique depending on the approached used in a particular study. Secondly, instead of using credibility techniques put forward by one scholar (Barusch et al., 2011), or using selected credibility techniques based on different philosophical assumptions (Shek et al., 2005), coding categories of this codebook synthesized a considerable number of credibility techniques described in the literature. The comprehensive coverage of the credibility techniques can help avoid selective bias in assessment, and promote a more holistic mapping of the field. In addition, a number of details were added to the codebook to improve its utility and make it more transparent. For example, in this codebook, the techniques are organized according to the progressive research stages of research design, sampling, data collection, analysis, and presentation of

findings. This was dome to align with the usual order in an evaluation report for the codebook to be more easily followed by coders. Apart from providing descriptions for each coding category, the codebook also contains examples and indicative questions to help coders understand and recognize the listed techniques in the text.

A pilot study. Before coding the entire data set, some systematic procedures were used to evaluate the utility of the codes and checking reliability of the codebook. In this case, the researcher worked independently on a small subsample of articles as part of the calibration procedure, and based on initial test of the codebook, modifications were made on descriptions of the coding categories, details of item options, and a few elements that are unique to evaluation were added to the codebook.

Prior to the codebook being put in use for coding the sample articles, a pilot study was conducted. The pilot study contained a training section, an independent coding section, and a post-pilot discussion section. A coding sheet and a tutorial was presented and explained in the training of pilot coders to introduce information about the study and the codebook, and provided instructions on how to do the coding (see Appendix D for information sheet for coders). For the purpose of testing the utility and inter-coder reliability of the codebook, the pilot study sought support from graduate assistants (GA) of the Educational Research and Evaluation program. Three GAs (including the researcher) have all received systematic training in educational research design and methodologies, and they have consented to participate in the coding and refining of the codebook and also serve the role of peer debriefing. The selection of the pilot sample included six studies that cover different research designs, qualitative methods, evaluation

types, and credibility techniques. Cohen's kappa was calculated to check reliability (Crocker & Algina, 1986; DeCuir-Gunby et al., 2011).

Furthermore, in order to establish inter-coder reliability, and at the same time to enhance interpretive validity of the study, the codebook was piloted with three educational researchers. Each coder was asked to evaluate the same six articles drawn from a sub-sample of selected articles, including articles from different journals and qualitative methods. Two sessions of coder tutorials were provided before the pilot coding, and two discussion sessions were organized after the coding for feedback. The coders discussed their coding experience, pointed out advantages and difficulties of using the codebook. Based on the feedback, the codebook was revised in the following aspects: (a) codebook categories were simplified, and sub-categories of each technique were merged; (b) more elements such as selected text column, page number, explanations and instructions for techniques were added to the codebook. The pilot coding achieved 92% agreement (with a Cohen's kappa of 84%, N=40), which indicated the codebook was a reliable tool, and remaining difference was discussed and resolved in discussion sessions.

Data Analysis

Two types of coded data were included in analysis: (a) the numerical results constituted of binary or categorical scores that indicate the frequencies of credibility techniques, and (b) excerpts of articles related to codebook categories coded with open-ended notes that are designed to record details in appraisal process and to stimulate critical judgment (See Appendix E). Both quantitative and qualitative data were analyzed

in tandem to reach the overall conclusion, which is to be presented in Chapter 4 and Chapter 5.

In the main analysis, descriptive statistical methods were used to identify the number of credibility techniques and the most frequently used techniques. The research took an exploratory approach to describe the data and look for patterns among variables. Quantitative procedures, such as correlation, *t*-tests, and multiple regression were used to examine the relationships between the number of credibility techniques reported and variables of the articles, such as the methodology used (qualitative, quantitative, and mixed methods), length of articles, and year of publication.

Qualitative content analysis was used to code selected texts that describe credibility techniques in articles, and the coding process involved searching for and selection of relevant text, systematic classification of the content, identifying and interpreting themes and patterns (Hsieh & Shannon, 2005; Schreier, 2012). As described in the codebook section of this Chapter, coding categories credibility techniques were developed based on existing theories, practical appraisal tools, and prior research. Using the codebook as an initial framework, articles were carefully reviewed, and all text that appeared to describe credibility techniques were highlighted and classified into coding categories. Text that could not be classified into one of these categories was coded with new, emergent codes that captured the essence of the technique, and the coding list was adjusted to include new coding categories and sub-categories under existing technique categories. Then a thematic analysis was performed to examine the content of the selected text (Boyatzis, 1998; Schreier, 2012). Though largely an iterative process, the

thematic analysis process has mainly three interconnected stages. The first stage was open coding for characteristics of credibility technique reporting. In this analysis stage, all selected text under each coding category was carefully reviewed, characteristics of technique reporting connecting to text were identified in comments, and frequencies of codes were also calculated. Then all identified characteristics were re-organized by coding categories, such as design related technique categories like sampling, data analysis, and methodologically driven technique categories, or credibility techniques like member checks and triangulation. Sub-categories were created when necessary. The second analysis stage involved selective coding that focusing on major coding categories to identify themes in reporting characteristics. Three types of codes with the following characteristics were selected as *key codes*: Codes indicating theoretically important features of reported techniques, prevalent features that occurred a significant number of times, and interesting and unique cases. Key codes were further examined and brought back to the context in which they were used in the articles. The final analysis stage entails description and interpretation. Codes of all analysis stages and researcher notes were comprehensively considered to connect major themes, identify trends, and looking for reasons and implications of the findings. In addition, although the coding strategies and results were discussed with peers, the coding process was carried out by the researcher alone.

In the whole analysis process, both qualitative and quantitative approaches were used in tandem, and results generated from each approach were triangulated to not only promote credibility, but also provide evidence and inspirations for further analyses.

Analytic methods and the order of analysis were data driven, and qualitative approach help develop a more nuanced understanding of quantitative results, and statistical analyses were also used to confirm or provide alternative evidence of qualitatively found themes.

For example, during the qualitative coding process, factors such as length of the article, qualitative methods used, and author's disciplines were found to show interesting interactions with characteristics of credibility technique reporting. As these factors were not originally included the codebook, they were added as new categories (emergent design). Another round of data collection was conducted to obtain numeral results on these new categories, and statistical tests were performed to see their impact on the overall data. More examples and triangulation results are presented in the following Results chapter.

Researcher's Beliefs

Even though the research synthesis design gives particular emphasis to systematic procedures for the sake of bias minimization, it still upholds several assumptions that are intertwined with the personal beliefs of the researcher. In this subsection, some of these beliefs are explicitly discussed under two categories: First, the pragmatic stand in conducting this research, and second, the researcher's understanding of bias and methodological rigor.

It is undeniable that the methodological orientation of the researcher will inevitably influence her perceptions on qualitative concepts and the way the research synthesis is conducted. The landscape of educational research has showed growing

diversity and complexity in methodologies. The researcher was trained with quantitative, qualitative, and mixed methods, and based on the assumption of co-existence of different paradigms, and the complementary nature of qualitative and quantitative research, the researcher is open to the variety of approaches and methods on doing primary and synthesis research.

Another important issue about bias is that bias is introduced to research by the methodological choices, philosophical beliefs of the researchers, political beliefs and context of the project (Suri & Clarke, 2009). It is the researcher's belief that all the procedures and strategies adopted in this research synthesis, such as transparent design and implementation procedures, and constant self-reflection, are not to simply avoid bias overly influencing findings, but to illuminate bias and the source of its formation.

Finally, although there are established procedures in conducting mixed methods synthesis research, it is a relatively new field in methodology, and compared to the majority of synthesis studies that integrating findings, there are few studies that review methodological techniques and characteristics. Therefore, it is a learning process for the researcher to conduct this synthesis study, to explore, learn, and adjust the existing methods for this particular study.

Summary

This chapter has described the research methodology used in conducting this research enquiry. The main method is a mixed method research synthesis of empirical evaluation studies that use qualitative methods published in leading educational evaluation journals during the period of 2003-2012. Going through a process similar to

the one in primary studies, the research process contains the stages of synthesis questions formulation, development of a protocol that delineates search strategies, inclusion and exclusion criteria, data extraction and management procedures, construction and calibration of a codebook, data analysis and interpretation. By highlighting the reporting of credibility checks, the synthesis method can help summarize the accumulated state of practice in evaluation studies. Particularly, this synthesis yielded a first look at the frequencies of credibility and credibility techniques being reported, the prevalence of these techniques, characteristics of their use across time and methodological approaches. The empirical knowledge generated from this synthesis can not only inform current practice and reporting of credibility, but also direct future research to help it yield a maximum amount of new information.

Chapter 4: Results

This research synthesis study examined one of the aspects of quality criteria in qualitative methods, to examine the reporting of credibility techniques in published evaluation reports that use qualitative methods. More specifically, this study selected empirical educational evaluation studies that use qualitative methods published in six leading evaluation journals (from 2003 to 2012). The research questions are: (a) To what extent are credibility techniques reported in published educational program evaluation work? (b) What are the features that can be observed in the reporting of credibility techniques? The six sub-questions listed in Chapter 1 can also be categorized into the following three groups. The first group of sub-questions focus on the presence and frequencies of credibility techniques: What are the credibility techniques used in practice, and how common are they? What are the relationships of the presence of these techniques with background factors such as time of publication, or methodology of the study? The second group of sub-questions is related to supporting details of credibility techniques used: Are the credibility techniques vaguely mentioned or thoroughly described? How are these techniques reported? The last group of sub-questions deals with the reporting language: What are the characteristics of language and terms used to describe credibility techniques?

This study adopted a mixed methods synthesis approach. As delineated in Chapter 2 and Chapter 3, synthesis research is an umbrella concept that includes different types of synthesis efforts. Some of the popular synthesis types include meta-analysis, which emphasizes statistical integration of quantitative results, and meta-synthesis that

integrates qualitative knowledge to generate new understanding (Smaling, 2003; Suri & Clarke, 2009; Whittemore, 2005). The mixed methods synthesis approach adopted in this study aggregates data on the methodological use of credibility techniques, and systematically examines the data using both qualitative and quantitative methods. It follows a pre-defined synthesis protocol with specified steps and methods to help reduce researcher bias, and synthesis methods and procedures are explicitly described to promote transparency. Therefore, results to be presented in this chapter are outcomes of a systematic and rigorous methodology.

In this results chapter, the measures the researcher has taken to enhance reliability and credibility of this study, the article search and review process, and most importantly, the synthesis review results are presented. The results are discussed in three sections, following the three groups of research sub-questions. Considering that quantitative and qualitative approach are closely inter-related in the design and coding process, the order of result presentation follows major themes found in data, rather than strictly separate results into quantitative and qualitative sections. In addition, it is necessary to stress that the results do not represent all qualitative evaluation studies, but are only delimited to empirical higher educational program evaluation studies (The selected study was narrowed down to higher education. Details are discussed in Chapter 4) that use qualitative methods published in six selected leading peer reviewed evaluation journals.

Reliability and Credibility Checks

The data analysis process was carried out by the researcher alone, and 50% of the articles were double coded by the researcher. The double coding process included both

quantitative and qualitative coding, and there was a six-month interval between the first and second coding as an effort to minimize bias. Reliability analyses showed 99.2% of agreement rate in quantitative coding, and new comments generated from the second round of qualitative coding were added with a different label.

In order to enhance the rigor of a research synthesis and for it to have practical significance, the researcher has made efforts to maintain a primary commitment to produce credibility. Basic design related techniques the researcher used in this study were reported in the methodology section (see Chapter 3), and techniques such as reflexivity and thick description were also applied, but more as general principles throughout the research process. For example, the researcher made efforts to support each claim with examples from the articles with original texts, so that the readers can have an idea of the "actual data." Critical self-reflection in this study is best reflected in the efforts made to achieve transparency. The researcher constantly reflected on each choice made in the synthesis process, and provided justifications to make each selection an informed decision. Selecting one choice against another certainly introduces bias or subjectivity, but it is important to have purposeful decisions rather than un-reflected selectivity. Transparency is greatly emphasized in this research synthesis, and also in many different kinds of syntheses, largely because transparency of the research design and procedures gives readers the opportunities to judge, verify, and critically evaluate the quality of the findings (EPPI-Centre, 2006; Suri & Clarke, 2009). Transparency also allows the audience to compare the similarities of the synthesis context and facilitate transferability

of the finding (Brantlinger et al., 2005; Johnson & Christensen, 2012; Suri & Clarke, 2009).

Apart from the above credibility techniques that are considered primary quality standards, the following credibility techniques were also used in this synthesis study. First, methods triangulation was used and reported in the *data analysis* section. On the one hand, when some reporting features were spotted in qualitative analysis, it is meaningful to examine its prevalence by statistical analysis. On the other hand, frequencies can only indicate the general state, and answer the yes or no question, but further analysis has to rely on qualitative content analysis to understand the detail and context of the presence or absence of the technique. Statistics and text analysis often provide converging evidence, such as findings regarding presence of credibility techniques, and the relationship between credibility techniques and author background and methodological approach, which will be presented later in this chapter. In short, the roads of these two approaches were crossed from the beginning of this research, and become deeply intertwined in the coding process, like two sides of a mirror. Qualitative content analysis was like the magnifier to examine subtle details, while statistical analyses were like panoramic telescope that help the researcher with a bird-eye view, and both approaches inspire creative thinking using the other approach.

Second, since the construction of the codebook, the research plans and progress were shared with peers of the research community via peer debriefing. The codebook design, synthesis research ideas and results interpretations were shared and discussed with evaluators and graduate students in regional educational conferences, with fellow

doctoral students of educational research, visiting scholars of the university, and scholars outside of education discipline in formal and informal academic gatherings. As suggested as a main function of the peer debriefing technique, explaining the research ideas, sharing the research progress, and answering questions raised by peers helped the researcher to think clearly and to improve the research plans. Comments and suggestions from peer debriefing provide fresh perspectives, and sometimes even served as important warnings to bring the research back on track when the researcher has worked on the topic for a relatively long time and was sometimes deviant from the original research goals.

Some other credibility techniques used in this study include audit trail, comparison and contrast, and expert check. First, audit trail was used to document all strategies used in each phase of the study, rationales behind the selection, use, development, and abandonment of those strategies. Audit trail of this study is composed of two major types of records: (a) data related records, such as searching notes and searching result lists of each round, synthesis procedures, and retrieved articles; (b) idea related records, such as reflection notes, analysis summaries, research plans, and decision making rationales. Moreover, constant comparisons were made between evaluations that use qualitative and mixed methods approaches, between different qualitative methods, and between studies with different kinds of similarities. Last but not least, reference librarians were consulted as a means of "expert check" to help locate and narrow down leading journals in the field, reviewing reference lists of target articles, and introducing resources for the researcher to collect different statistics on the journals and articles. The design related and methodologically driven techniques were used together to form a key

mechanism throughout the whole research process to optimize the credibility of the study (Brantlinger et al., 2005; Johnson & Christensen, 2012; Lincoln et al., 2011).

Strategies taken to promote credibility of this study can also be perceived with the lens of criteria for mixed methods studies. With all discussions on quality standards of mixed methods in Chapter 2, the validity issue remains a controversial topic and is still under ongoing debates. Although scholars recommended apply credibility/validity checks when mixed methods are used, there are no generally agreed standards; and compared to discussions on conceptualization of the standards, there are few practical guidelines to follow. Thus, based on existing criteria of mixed methods research, strategies used to promote the overall quality of this study can also be examined in three major domains: design, implementation, and interpretation. First, design suitability and sampling legitimation of the study are checked to enhance design quality. At the beginning of Chapter 3, justification of using a mixed methods synthesis research to address the research question was explained, that is, mixed methods synthesis has great advantages in serving the purpose of describing the current state of the field, and neither qualitative nor quantitative approach alone can fully capture the goals of this study. Furthermore, sampling of this study, collecting both numerical and textual data from the same sample, largely enhanced the rigor in constructing meta-inferences by pulling together both qualitative and quantitative findings (Onwuegbuzie & Johnson, 2006). Second, implementation criteria of mixed methods research share many aspects in common with principles of synthesis research. Components of the synthesis protocol, such as article search (sampling) and data collection procedures, were implemented, described, and

documented to achieve both transparency and fidelity of study procedures. Finally, this study adopted a fully mixed design, and the integration of results occurred at all stages of the study, which promoted the interpretive rigor of findings and inferences. For example, the qualitative coding of the selected articles revealed the possible impacts of factors like article lengths, type of qualitative methods, and author's discipline on the use of credibility techniques and overall quality of the studies, resulted in additional collection of quantitative data. More statistical tests were also performed to provide triangulation of both quantitative and qualitative findings on the relationship between above factors and credibility techniques. It was found that although the lengths, types of qualitative methods, and author's discipline may have a strong relationship with the use of credibility in a small number of individual articles, they are not common features across all articles. Thus, statistical results in turn, helped better understand the diversity of individual articles, and draw a more complete picture. To summarize, findings and inferences of this study were formed and cross-checked within the study and with previous findings in an interactive manner throughout data collection, analysis, and reporting process to promote rigor of inferences.

Article Search and Review Results

After three major rounds of article search and review process, 118 articles were found eligible for analysis. (See Figure 1 for an overview of the article search and review process.)

The first round of article search combined both keyword search in electronic databases and hand searching. In a typical keyword search, one of the varied forms of a

keyword (e.g. qual* and qualitative) was entered, and in the search results, title, keywords, and abstract of each article was reviewed to retrieve the articles that met the criteria. Then process is repeated for all forms of a keyword and all listed keywords. Three hundred fifty eight articles were retrieved in the first round.

In the second round, the retrieved articles were more closely reviewed, including reading part of the article. In this round 53 articles were excluded including duplicates, leaving 205 articles eligible. The most common reasons for exclusion at this stage were: Not empirical evaluation studies that involved first hand data collection, not targeted at educational topics, and did not involve qualitative approach.

In the third round, given the relatively large number of articles retrieved (205), in order to make the current study more manageable and also to be more focused on a field, the selection criteria were narrowed down to studies on the field of higher education. The topic of higher education included all tertiary education and further education. Thus in the last round 118 articles were included in the analyses of present report.

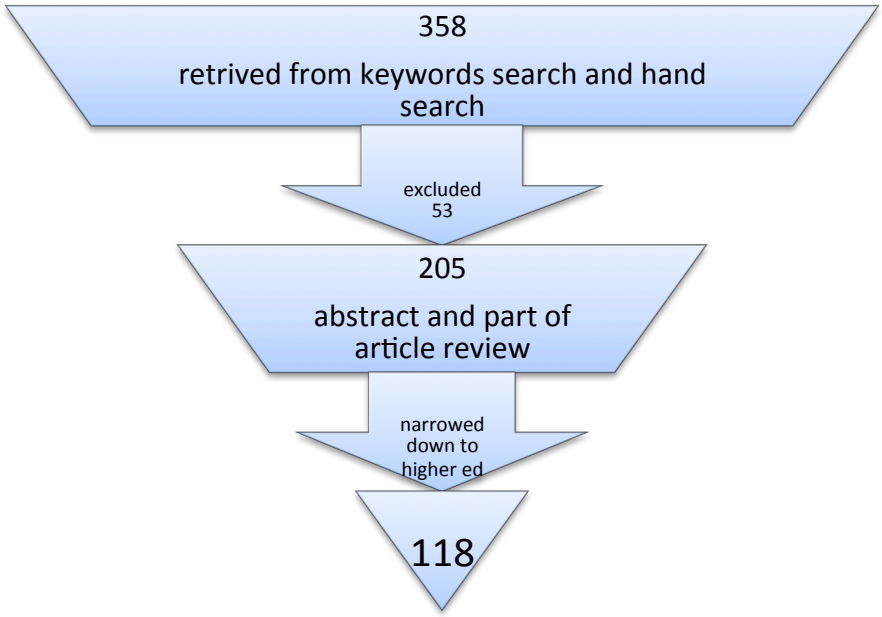


Figure 1. Flow chart of article search and review process of the present study.

Findings are presented in the following three sections, and research questions answered. The first section "frequencies of credibility technique" provides a general description of the selected articles and credibility techniques, to answer the first research question "to what extent are credibility techniques reported in published educational program evaluation work?" The rest of this chapter provides answers to the second research question "What are the features that can be observed in the reporting of credibility techniques?" from the perspectives of reporting language and vocabulary, reporting diversity, and new credibility techniques emerged from practice.

Frequencies of Credibility Techniques

The number of articles and the percentage of total published articles from each journal can be found in Table 1. The journal *Assessment & Evaluation in Higher Education* has the most selected articles: 98 articles that takes 83.1% of the total, and

when the selection criteria are narrowed down to evaluation studies in higher education, there was no article found eligible from *Educational Evaluation and Policy Analysis.* With regard to this result, it is to be noted that *Educational Evaluation and Policy Analysis* has been considered as a journal receptive to publishing qualitative research for at least ten years (Preissle, 1996; Price et al., 2005; Wark, 1992). In addition, when compared to the total number of articles published in each journal, the number of evaluation studies that use qualitative methods takes only 6.7% on average. The journal *Assessment and Evaluation in Higher Education* has 18.4% of selected articles, which is the largest percentage of all journals, and despite the journal *Educational Evaluation and Policy Analysis* that has no selected journals, Evaluation Review has 0.8% of selected articles as the second lowest percentage.

Table 1

Published and Selected Articles by Journal

Journal	Published N	Selected N	%
American Journal of Evaluation	321	6	1.9
Research Evaluation	296	8	2.7
Assessment and Evaluation in Higher Education	532	98	18.4
Evaluation Review	260	2	0.8
Educational Assessment, Evaluation and Accountability	146	4	2.7
Educational Evaluation and Policy Analysis	201	0	0
Total	1756	118	6.7

Over the decade, the number of articles that report empirical evaluation studies that use qualitative methods in the field of higher education increased 3.75 times, from 8 articles in 2003, and 6 articles in 2004, to 27 articles in 2012. There is only slight fluctuation in terms of the number of such articles per year, but there is an apparent increase of published articles since the year 2009 (see Table 2).

Table 2

Published and Selected Articles by Year

Year	Published N	Selected N	%
2012	222	27	12.2
2011	187	16	8.6
2010	171	17	9.9
2009	186	12	6.5
2008	146	6	4.1
2007	167	7	4.2
2006	145	9	6.2
2005	189	10	5.3
2004	165	6	3.6
2003	178	8	4.5

Of the 118 articles, 58 (49.2%) articles primarily used qualitative methods, 54 articles (45.7%) used both quantitative and qualitative approaches, and 6 articles (5.1%) used primarily quantitative methods but with some qualitative features.

Although the number of articles that use qualitative methods increased with time, the change in the proportion of qualitative stand-alone and mixed methods studies is minor. Before 2009, 50% of identified articles are qualitative stand-alone studies, and after 2009, despite the rise of mixed methods as the third paradigm (Johnson & Onwuegbuzie, 2004), the proportion of articles that used mixed methods only increased by 2%.

Five major qualitative methods were used in the articles: Interview, focus group, document analysis, observation, and questionnaire (see Table 3). The most commonly used qualitative method is interview, which is used in 59 articles (50%); document analysis is used in 29.7% of the articles, followed by focus group (22.9%). The least used methods are observation (in 11 articles, 9.3%), and questionnaire (5.9%). There are 21 articles (17.8%) that combined more than one qualitative methods, and methods commonly seen used in tandem include questionnaire and interview, or interview and focus group.

Table 3

Use of Qualitative Methods in Selected Articles

Methods	N	%
Interview	59	50
Focus group	27	22.9
Observation	11	9.3
Document analysis	35	29.7

To give a general picture of credibility reporting, using the categories of credibility techniques listed in the codebook, the researcher examined the frequencies of reported credibility techniques to provide an overall indication of the relative usage of credibility techniques. There are also credibility techniques reported in the articles that are not listed in the codebook, which are described in the "new credibility techniques" section.

Table 4 describes the general statistics of the credibility techniques listed in the codebook. Of the listed 20 credibility techniques, including seven design related techniques, and 13 methodologically driven techniques. Design related techniques are the most basic techniques to satisfy the primary quality criteria of qualitative evaluation research, and methodologically driven techniques are major credibility techniques drawn from literature. On average, about six (6.23) techniques are used in an article, five (4.94) design related techniques and one (1.29) methodologically driven technique. There are articles where all seven design related techniques are present, and the maximum number

of methodologically driven techniques used is five. At most, 12 techniques are used in a single article.

Table 4

Descriptive Statistics of Design Related, Methodologically Driven and Overall Credibility Techniques

Techniques	Mean	Median	Mode	Std.	Min	Max
Design Related	4.94	5	5	1.215	1	7
Methodologically Driven	1.29	1	1	1.199	0	5
Overall	6.23	5	5	1.901	1	12

As it is shown in Table 4, the mean, median, and mode of the design related techniques and overall credibility techniques are close to each other, showing the data are clustered around the mean and no extreme case is found. Distribution of the frequencies of the techniques is better shown in Figure 2. The majority of selected studies used five to seven techniques per article. Methodologically driven credibility techniques are not often used in selected articles, which is indicated in the frequencies of methodologically driven techniques, and the skewness to the left of its histogram (See Figure 2). Although at most six methodologically driven techniques were used in an article, the majority of articles used just one or no methodologically driven technique.

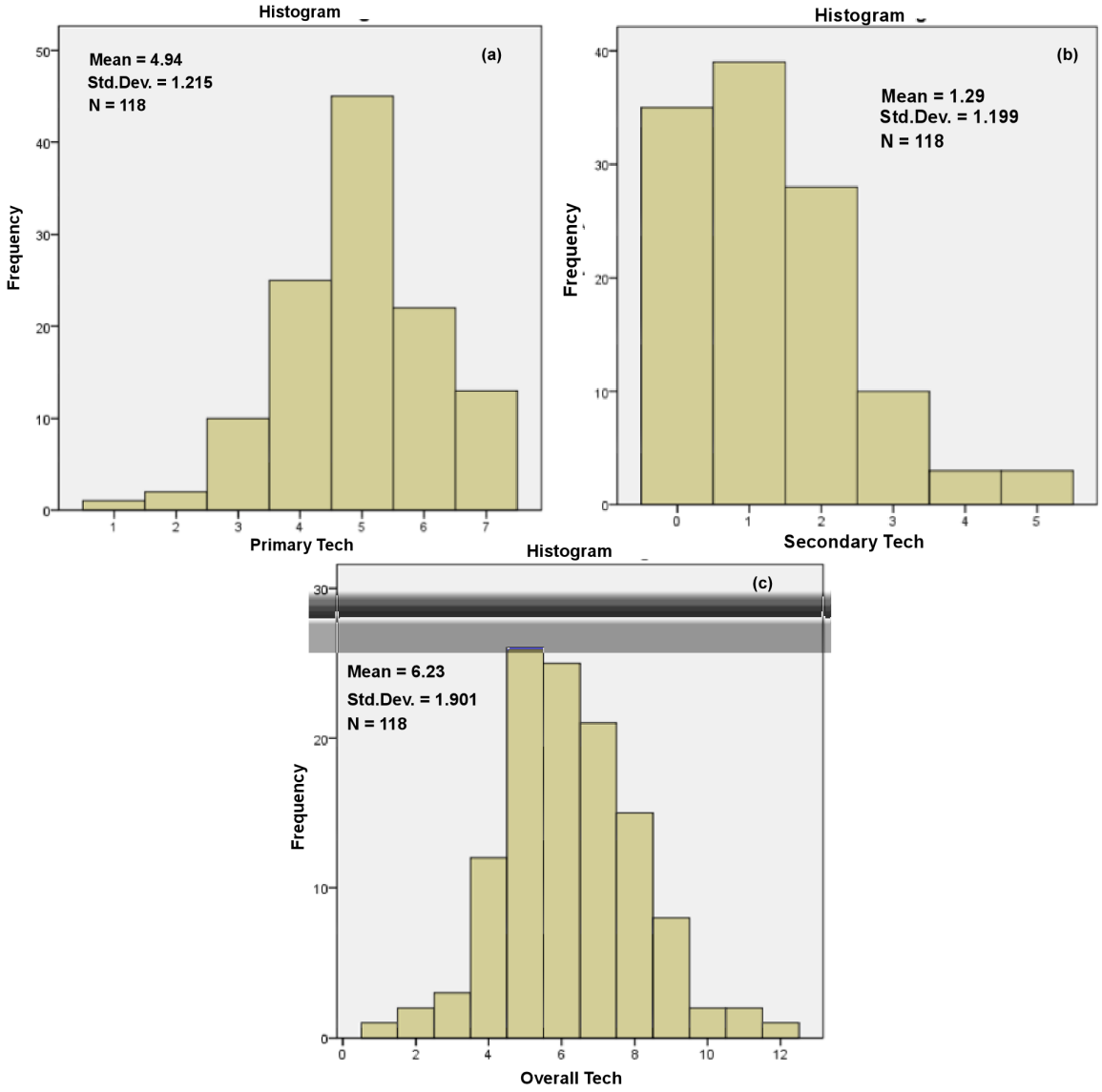


Figure 2. Histograms of credibility techniques. (a) Histogram of design related techniques. (b) Histogram of methodologically driven techniques. (c) Histogram of the overall techniques.

The average number of credibility techniques by demographic factors, including year, journal, and author's current department, and methodology are also calculated and presented in Table 5. Interestingly, the average numbers of credibility techniques used are all around 6.

Table 5

Average Credibility Techniques by Year, Journal, Author's Background

	Year	Journal	Author (Edu)	Author (Non-Edu)
Average N	6.33	6.23	6.47	6

As suggested by earlier results, the use of qualitative methods in both qualitative and mixed methods studies showed very balanced development over the past decade. Descriptive results of credibility technique used in the two approaches show that the mean and the range of credibility techniques used in qualitative stand-alone studies and mixed methods studies are also close. The average number of credibility techniques used in qualitative stand-alone studies and in mixed methods studies are 6.05, and 6.64, and their respective ranges are 2 to 11, and 2 to 12 (see Table 6).

Table 6

Credibility Techniques by methodology

	N	Min	Max	Mean	Std.
Qualitative	57	2	11	6.05	1.563
Mixed methods	55	2	12	6.64	2.067

Similarly, the average use of credibility techniques also shows to be quite steady. The average number of techniques used per year is 6.33, and there are very minor

variations over the ten years. The overall and average number of credibility techniques used in each year are summarized in table 7.

The average number of credibility techniques used per journal is 6.23, and there is not much difference in the means of credibility technique use in each journal (See Table 8).

Table 7

Number of Articles and Credibility Techniques by Year

Year	Published N	Selected N	%	Technique N	Tech Average
2012	222	27	12.2	166	6.1
2011	187	16	8.6	89	5.6
2010	171	17	9.9	112	6.6
2009	186	12	6.5	110	9.2
2008	146	6	4.1	38	6.3
2007	167	7	4.2	46	6.6
2006	145	9	6.2	56	6.2
2005	189	10	5.3	72	7.2
2004	165	6	3.6	31	5.2
2003	178	8	4.5	53	6.6

Table 8

Credibility Techniques by Journal

Journal	Technique N	Article N	Average
American Journal of Evaluation	42	6	7
Research Evaluation	54	8	6.75
Assessment and Evaluation in Higher Education	589	98	6.01
Evaluation Review	19	2	9.5
Educational Assessment, Evaluation and Accountability	31	4	7.75
Total	735	118	6.23

The length of an article has shown to play a role in frequencies of credibility techniques. As it is shown in Table 9, the length of the 118 selected articles ranges from 8 to 29 pages, and the average length is 15 pages. A Pearson correlation coefficient was computed to assess the relationship between the number of credibility techniques and length of an article. There was a positive correlation between the two variables, $r(118) = 0.281$, $p < 0.01$, $r^2 = 0.079$. Overall, length of an article is positively and significantly correlated with the number of credibility techniques reported, indicating that the longer the article, the more credibility techniques are likely to be reported. Meanwhile, r^2 shows that less than 8% of variance of the number of credibility techniques is accounted for by page length, in other words, only a small proportion of the use of credibility techniques is predictable from length of the article. One of the implications of these results is that

quality issues or lack of reporting details should not be solely attributed to special constraints.

Adopting an exploratory approach in describing the data, multiple regressions were also performed to search for patterns of the data. Descriptive statistics, histograms, and correlations were examined to test key assumptions such as normality, linearity, and variable reliability. With "number of credibility techniques used" as the dependent variable and predictors included length of the article, methodology (qualitative, quantitative, or mixed methods), qualitative methods (interview, observation, etc.), author's department, year of publication, and journal, a multiple regression model produced R square 0.157, $F(6,111) = 3.442$, $p < 0.01$. The results of the regression indicated the predictors explained 15.7% of the variance. No multicollinearity was diagnosed. Yet no individual coefficient in the model was statistically significant. When backward method was used, it was found that the number of credibility techniques used could be significantly predicted by the length of the article ($\beta = .115$, $p < .01$) and methodology ($\beta = .714$, $p < .05$), when all other predictors were removed from the model. Results of post hoc tests also confirmed that there is no significant difference in the use of credibility techniques between qualitative and mixed methods studies, but quantitative studies that used qualitative methods (mean = 4) is shown to be significantly different from qualitative or mixed methods research. However, it is to be kept in mind the basic defects of stepwise regression strategies, such as problem of multiple hypothesis testing, bias of estimation, and an inappropriate reliance on a single best model (Thompson, 1984; Whittingham, Stephens, Bradbury, & Freckleton, 2006).

Table 9

Lengths of Selected Articles

	Min	Max	Range	Mean	Std.
Length	8	29	21	15.16	3.989

Frequencies of each individual credibility technique are also calculated. Descriptive statistics of individual credibility techniques are summarized in Table 10. On the whole, all articles reported on data collection, and more than 93% reported on sampling techniques. Other frequently reported credibility techniques include design and analytic details, which are reported in more 80% of selected articles. Thick description is reported in 69.5% of articles. The rest of credibility techniques, such as triangulation, member checks, and negative case analysis, are reported in less than half of the articles. The least commonly reported technique is persistent observation, with only one article reported using such a technique.

Table 10

Frequencies, Details, and Terms of Individual Credibility Techniques

Credibility techniques	Frequency (%)	Details (%)	Terms (%)
Data collection	118(100)	112(94.9)	104(88.1)
Sampling	110(93.2)	95(86.4)	71(69.6)
Design	102(86.4)	86(84.3)	64(66.7)
Analytic details	96(81.4)	84(87.5)	36(32.7)
Thick description	82(69.5)	73(89)	26(31.7)
Triangulation/Crystallization	52(44.1)	38(95)	22(42.3)
Limitation & delimitation	40(33.9)	37(71.1)	19(70.4)
Reflexivity	30(25.4)	28(93.3)	8(57.1)
Comparison and contrast	27(22.9)	22(81.5)	6(20)
Member checking	14(11.9)	14(100)	5(6.1)
Prolonged engagement	11(9.3)	8(100)	5(71.4)
Multivocality/multiple perspectives	9(7.6)	6(75)	4(50)
Peer debriefing	8(6.8)	6(66.7)	2(40)
Exploring rival explanations	8(6.8)	5(71.4)	2(66.7)
External auditor/expert checking	7(5.9)	5(45.5)	2(22.2)
Audit trail	5(4.2)	3(60)	1(12.5)
Negative case analysis	5(4.2)	3(100)	1(20)
Weighting the evidence	3(2.5)	2(40)	1(9.1)
Pattern match	2(1.7)	2(100)	0(0)
Persistent observation	1(0.8)	1(100)	0(0)

Overall, design related techniques are more frequenctly reported than methodologically driven techniques. The most commonly used design related technique is data collection (data collection methods described), and the least common design related technique is reflexivity.

The most common design related technique is triangulation, and the least common is persistent observation.

Terms Used to Describe Credibility Techniques

The following section presents the results of terms and variation of vocabulary used to describe credibility techniques.

As it is mentioned in Chapter 2, a variety of different terms are developed to describe qualitative concepts and methods, and the use of these terms largely reflected the understanding and reporting style of the evaluator. So as an integral part of design of the codebook, language choices of credibility techniques were documented. The results are presented in the following two categories: (a) presence and absence of credibility technique terms; (b) variation and connotation of terms.

Presence and absence of the terms. Similar to the general use of credibility techniques, design related techniques are more often reported with a specific term than methodologically driven techniques. Frequencies and percentage of term reporting are summarized in Table 10. According to the results, authors of selected articles are more familiar with basic design related techniques terms like data collection methods (104, 88%) and designs (71, 69%), and terms such as questionnaires, focus groups, think aloud approach, and phenology are relatively more frequently specified. Comparatively, some

other design related technique terms, though theoretically considered as equally critical, are less commonly mentioned in these articles by name, such as “reflexivity” (6, 20%), and “thick description” (5, 6%).

Among methodologically driven techniques, triangulation (22, 42%), comparison and contrast (19, 70%), and member checks (8, 57%) are the most frequently referenced methodologically driven techniques, and they are also reported with varied names. Peer debriefing (1, 12.5%), negative case analysis (1, 20%), and prolonged engagement (1, 9.1%) are only used once in selected articles, and persistent observation and pattern match are not used with any term at all.

With presence of the techniques highlighted, it is necessary to point out that there are many examples of reporting a technique with great details but without using any term of the technique, and I call this reporting feature “detail with no term.”

Thick description is a typical technique that very few terms are used. Of the 82 (69.5%) articles that used the technique, only 5 (4.2%) articles used the term “thick description.” However, in most cases, the technique of thick description is clearly present. Low-reference language is used in describing the phenomena, and direct quotes or other forms of original data are presented to illustrate experience of participants. For example, Hay, Engstrom, Green, Friis, Dickens, and Mac Donald Hay et al. (2012) conducted an evaluation on online clinical assessment of a university medical program. In presenting interview results of assessment of expert online teaching, the authors used quotes from three students. The first and third quotes illustrated the importance and usefulness of expert demonstration in clinical learning in students’ own voice, and the second quote

directly pointed out the consolidation effect of expert teaching, and the three quotes together supported the authors' claim that online expert teaching is found to be advantageous.

In the above example, presenting student quotes from interview transcripts helped readers to see the actual data, which is an important characteristic of thick description. Later in this article, the authors continued to present the results on different approaches adopted by students in videoing their own clinical practice. Apart from providing more quotes, information on the student's performance with different approaches (individually view their practice and peer feedback) is also provided, so the quotes can be viewed and understood in a broader context. In a nutshell, the study is described in thick details, but the term "thick description" or any related term is not used.

Similar examples can also be found in the use of triangulation. Generally speaking, the term "triangulation" is more often referenced in mixed methods studies than in qualitative studies, but in both types of studies, 57.69% of the articles did not use the term even when the technique is clearly applied. For example, in De Filippo, Casado, and Gómez (2009), the primary method is document analysis, and the researchers decided to "conduct(ed) interviews to obtain more information about the motivations to engage in research stays as well as personal opinions about the importance of mobility" (p.8). Finding from document analysis was validated with interviews. Moreover, "different sources were used" (p.191, p.8), including first-hand personal data collected from participants, and documents and records from institutional databases. On top of that, it is explicitly stated in the title and in the article that both quantitative and qualitative

approaches were used. Therefore, through descriptions of the article, the methods, purpose, and process of triangulation are described, and the authors even clearly explained the features of methods triangulation and data triangulation, but the term is never mentioned.

Similar examples can be found in many other techniques, and usually the descriptions of the technique are highly integrated in the results, but the methodological terms are not brought up. Therefore, using terms of the credibility technique in reporting can certainly help make the article more identifiable in this aspect, and enable more efficient communication, but terms alone are far from enough in assessing the quality of reporting of finding. Articles with absence of the terms can still have detailed description of a technique, and articles with the terms present may only provide vague information. In contrast to "detail with no term," (detail = yes, term = no) such feature of "term with no detail" (detail = no, term = yes) is also present in the selected articles.

There are a number of cases that have only the term reported without any supporting details, despite the circumstances when many articles have no details for a term in the main body of the text due to limited space, but provide further information in an appendix or with a web-link (e.g. (Andrade & Du, 2007; Borg, 2009; Bradley, Oterholt, Nordheim, & Bjorndal, 2005). Compared with other techniques, using a term with no details is more commonly seen in the reporting of analysis. For example, Poulos and Mahony (2008) claimed "thematic analysis" in describing their analysis strategy, but no further information is given and the description jumped directly to results. There are different kinds of ways of conducting a thematic analysis, and it is difficult for readers to

specify the analysis method and make judgment on the quality with limited information. The same happened for Huxham, Campbell, and Westwood (2012). Other examples involve using very broad terms. Prins, Sluijsmans, Kirschner, and Strijbos (2005) in data analysis section stated twice that results were "analyzed qualitatively" (p. 428), but no information is provided on what was qualitatively done, or defining details or procedures of such qualitative analysis.

Thus, using the terms indeed shows some understanding of credibility techniques, and can point out a direction for reader's assessment of the quality, but lack of details in reported term makes it difficult for the reader to assess what has been done.

Variation and connotation of technique terms. The terms listed in the codebook and in discussion of this synthesis are terms used in theoretical literature of qualitative research. The results show that there is a larger variety of terms used to describe credibility techniques in practice.

Member check is not the most commonly used credibility techniques, but this technique seems to have the most variation of forms of all, especially among methodologically driven techniques. McCormack (2005) provided "member checking" for participants to "indicate whether the reconstructions of the inquirer are recognizable;" (p.467). Frank and Barzilai (2004) and Tian and Lowe (2012) shared with participants their interpretation to promote "respondent validity;" Taut and Alkin (2003) also referred to the first level member check process, interviewee reviewed transcripts, changes and additions were requested and handled, as "validation." Asghar (2010) referred to the technique as "respondent validation" to confirm interview transcripts. Craddock and

Mathias (2009) combined level one and level two member checks and called the process "participant validation."

The term "credibility" and its related terms are also mentioned in selected articles. Authors directly addressed credibility issues. Varied terms are used to address credibility. Trustworthiness (e.g. Thompson, Brooks, & Lizarraga, 2003; Weurlander, Söderberg, Scheja, Hult, & Wernerson, 2012), validity (e.g. Bradley et al., 2005; Frank & Barzilai, 2004; Smith, Brandon, Lawton, & Krohn-Ching, 2010), intersubjective certifiability (Winchester & Winchester, 2012), credibility (McCormack, 2005; Prades & Espinar, 2010)

It is also to note that in qualitative coding of the terms, although in some articles, authors with an educational background tended to use more credibility technique terms, in most of the selected articles, terms are seen used by authors who are from both educational backgrounds and other disciplines. Statistical analysis supported this finding. An independent-samples t test was conducted to compare the use of credibility techniques for authors with educational and other backgrounds. Results of the t test confirmed that there is no statistically significant difference in the number of credibility techniques for authors with educational backgrounds (M= 6.47, SD= 1.784), and authors with other backgrounds (M= 6, SD= 1.992); t (116)= 1.357, p = .177. Key assumptions of *t test* such as normality, independence of variables, and equal variance were met. The small effect size r = 0.125 also confirmed that authors from different backgrounds are similar in using credibility techniques.

Features in Reporting of Credibility Techniques

Frequencies and percentage of supporting details reported in selected articles are summarized in Table 10. As indicated by the table, more design related techniques are supported with evidence and explanations when the technique is applied than methodologically driven analysis. Generally speaking, all design related techniques are well supported with details. Design related techniques with high percentage of supporting details include data collection (94.9%), reflexivity (93.3%), and thick description (89%), and even the percentage of the least supported design related technique limitation reaches 71.1%. For methodologically driven techniques, Triangulation (95%), member checks (100%) and prolonged engagement (100%) rank on the top in terms of percentages of techniques with supporting details. An interesting feature here is that technique such as reflexivity, prolonged engagement, and member checks, though not commonly used in selected articles, once present, they are often supported with details of how the technique is used.

Compared with the presence and term use of credibility techniques, the majority of articles that used credibility techniques supported the techniques with some details, but the degree of evidence provided vary greatly with different techniques and authors. Reporting features are summarized based on themes of qualitative coding, and because the reporting of design related and methodologically driven techniques show quite different features in reporting details, the following section presents the results separately.

Design related technique reporting features. Compared with methodologically driven techniques, design related techniques have relatively broader connotations. What

should be included in each of these techniques are often specified in all kinds of literature of methodology, but in practice, authors presented their own understanding on the connotation of these techniques. In the following section, the researcher describes what were reported when design related techniques were used, and characteristics of reported details.

Content of reported design related techniques. Design related techniques in the codebook are basic methodology considerations when qualitative methods are used. A large variety of things are reported when describing design related techniques, especially the first four elements: Design, sampling, data collection, and data analysis.

Take the technique design for example, 55 articles provided specification of the type of design adopted, including names of the design, theoretical definitions and characteristics of the design, and 18 articles specified their methodological approaches, be it qualitative, quantitative, or mixed methods, and their respective characteristics. The theoretical perspective of the study is another major aspect often reported in design. There are 21 articles that described design explicitly presented information on their theoretical perspectives, which included theoretical frameworks, theoretical background information, and assumptions. Apart from the above two major aspects, rationales for choosing such designs were reported (13 articles), framework, procedures or logic models of their own studies (16 articles).

Data collection of qualitative methods is another typical technique that has a broad range of content being reported. General reporting categories on this technique include rationales for using selected data collection methods, data sources, procedures of

conducting an interview, ethnics like consent and permissions, and small details can be as specific as the durations of interviews, data collection protocols, questions asked and preparations for interviews and focus groups.

The design related technique that has the least variation in content was "reflexivity." Reporting on this technique is more concentrated on bias related issues (8 articles), information about the researchers (self-reflection) (10 articles), and other reflections on the study (10 articles), such as advantages and disadvantages of an approach or a design.

Varied but scattered reporting in an individual article. After giving an overview of what are the general categories covered, the next section presents the variety of content reported in an individual article. Although there may be much variation of information provided under each category of technique, there is an even bigger variety when each individual article is given a close look.

First, there are big variations in terms of how many aspect or kinds of content an article covered applying a technique. Some covers only one thing, others cover multiple aspects of it. For example, in terms of sampling, Lamont, Mallard, and Guetzkow (2006) provided sample size, reason for sample size, settings of sampling, sampling rationale, sample composition and a discussion on saturation. Another example in analysis is the descriptions of Cheng, Rogers, and Wang (2008). It includes the types of data included in the analysis, detailed steps the data were processed, analysis tools, researchers doing the analysis, theoretical groundings of data analysis, procedures, coder reliability and formats of results coming out of the analysis.

As basic design and methods requirements for credibility, design related techniques were more often reported with some details than methodologically driven credibility techniques. However, judging from results of qualitative analysis, important aspects of design related techniques were discussed when all selected articles were considered as a whole, but it is rare to see these important aspects of a design related technique covered in a single article. For example, when reporting sampling information, out of the many aspects that are critical to promote rigorous methods, such as sampling strategy, rationale, procedures, sample size, settings, and participant demographics or backgrounds, only few articles comprehensively reported these elements, most of the articles writers were selective in reporting one or two aspects. Conversely, no aspect in design related techniques was found commonly reported in all selected articles. Take design for example, no single aspect of design exceeds 20% in prevalence.

Comprehensive and brief reporting on an individual technique. To go even further, when describe one type of information in a technique, there is a big difference in terms of the space as well as magnitude or volume devoted to each technique. Some article described a technique by one or two sentences, some contain much more information for the readers.

Take analysis for example, more than one article reported content analysis as the approach for analysis, and Van den Berg, Admiraal, and Pilot (2006) mentioned content analysis and stopped there. Tang and Harrison (2011) made it clear about different kinds of content analysis, specified their study as “relational content analysis.” Comparatively,

Prades and Espinar (2010) stated the steps taken for the content analysis, so what were actually been done were even clearer.

Similarly, several articles adopted phenomenography as the design. Without judging better or worse, here I present the different reporting examples found in the articles. For example, both Tan and Prosser (2004) and Asghar (2010) adopted aphenomenography approach, but reported it in different ways. Asghar spent a paragraph of two sentences on reporting this approach. The first sentence stated the term and purpose of the approach, and the second sentence explained the connotation in this study's context, seeing the assessment process through participants' eyes. In general, this is brief but efficient. Tan introduced the concept and term in introduction section and background section of the article, and the paragraph of design contained six sentences. Importance of phenomenography in educational evaluation, its definition, purpose and goals of such approach, and its characteristics are reported. In summary, more information on the connotations displayed and emphasized in this study.

Apart from what is covered and how it is covered, it is also important to point out the absence of necessary information in reporting primary techniques. The reporting of sampling, looking from the frequencies and details reporting of this technique, most articles concentrate on "who" constitute the sample. This is critical for a study, but what should not be ignored is why these people become the sample and how they are recruited. The "why" question and the recruitment of the sample can largely affect the application of the findings.

All the articles reported qualitative sampling provided some information about their participants, though to different extent, but only 30 articles provided rationales, and only 25 had recruiting details. Even so, within the small number of 25 articles, sampling information is usually very brief, with phrases like participants were "invited to participate" (Cotton & Gresty, 2007, p. 586); "original list of interviewees was slightly altered and expanded as the study proceeded" (Taut & Alkin, 2003, p. 216).

In summary, although quantitative analyses suggested common presence of most design related techniques, qualitative analyses revealed the difference among studies with different degree of supporting details. Triangulated data analyses suggest that the methodology was more rigorous when more aspects of a design related technique were covered in reporting. However, although quantitative results show that page length was positively correlated with credibility techniques, qualitative analyses suggest that the space devoted to each technique was not necessarily an accurate indicator of methodological rigor. Many articles are concise in essential information to support a technique, highlight the special features, give a few examples of specifics or leave the full details in figures, tables, or even web-links (Bradley et al., 2005; Burden, 2010; Micari, Light, Calkins, & Streitwieser, 2007; Schmoch, Schubert, Jansen, Heidler, & von Görtz, 2010).

Methodologically driven technique reporting features. Methodologically driven techniques are less frequently used and less described than design related techniques. Given the limited number of descriptions of methodologically driven techniques, the reporting patterns are not as distinctively observable as design related

techniques. Most supporting details of methodologically driven techniques, though present, are often vague or only provide enough information for the researcher to recognize the technique. However, among the articles that do provide detailed description of the methodologically driven techniques used, there are two inspiring observations worth noting.

First, although the supporting details of methodologically driven techniques are often brief, the purpose of using the techniques and what has been done are usually put forward clearly. For example, experts were invited to "validate the analysis" (Hammouri, 2003, p. 575), or "disconfirming cases were sought" (Borg, 2009, p. 418) because "as the study evolved and theories developed, participants whose experiences might contradict these theories were sought" (p. 418).

Second, it is inspiring when supporting details not only include how or procedures of applying a technique, but also gave results of using the technique. For example, Hammouri (2003) applied expert checks on their analysis of classification of protocols. There were "few disagreements between the three experts and between them and us" (p. 575), and such disagreements were resolved in conference. Taut and Alkin (2003) used member checks for their interview transcripts. They reported "approximately half of the interviewees replied to the validation request. When respondents asked for changes or additions, they were very minor" (p. 218). It is helpful for the readers to know "what happened" after the techniques are applied, and like the above two examples, the majority of articles that used methodologically driven techniques reported agreement or

disagreement of the participants after the techniques were used, or the extent to which the finding was validated.

What is not commonly seen is what is learned with these results. Triangulation reporting can be a good example. Convergence of triangulation was more frequently used in studies that used mixed methods, but there are only a small number of articles that explained the meaning of convergence or contradictions, and how the converging results helped the analysis process. Examples are provided in the following section to demonstrate different levels of presentation of triangulation results.

In the first example, Gijbels, van de Watering, and Dochy (2005) explored the perceptions of students and teachers on written assessment tasks and their impact on student performance. A survey with both closed and open-ended questions and interviews were used, and findings of the survey and interviews were triangulated in searching for perceptions of written assessment tasks. The authors reported the results of their methods triangulation: “The open-ended questions and the interviews were found to have many features in common” (p. 82), and presented five emerged themes from the two sets of data. However, what is not reported in the article on triangulation is (a) what was learned from the converged results, were the authors more confidence on the shared results to be credible findings, or only converged results were reported? This question could not be answered because of the missing information. The first question on the credibility of the finding also leads to the second question, (b) what was different found by the triangulation, and how the difference affected the findings and conclusions? Therefore in this case, the author indicated the use of triangulation, but did not report its use

thoroughly. In reporting, it put the emphasis on triangulation results, instead of triangulation's impact of credibility.

Another example of triangulation is Meagher, Lyall, and Nutley (2008). They adopted a primarily qualitative approach to find out about the impact of research in policy and practice. Given the characteristics of the study with a component of methodological trial, the authors provided more methodological reflections in the technique of triangulation. Two things were emphasized in reporting triangulation in this study: First, it described how different methods intertwined with each other, or in other words, how quantitative and qualitative approaches were triangulated, such as how findings from survey and documents (reports) enriched and added to interview findings. Second, the authors provided their own assessment on triangulated findings. Through triangulation, they found that "there were no evident contradictions between results obtained by different methods" (p. 169), and "different channels for impacts were captured by different methods" (p. 170) not seems to be the case. Compared with the first example, this study reported observations of using the technique of triangulation apart from the mere results of triangulation.

In a third example of triangulation, Schmoch et al. (2010) took another approach of reporting triangulation. In this study that investigated the measurement of scientific performance, a qualitative content analysis approach was used to seek evidence to complement a quantitative study on the same issue. Different data sources, and in this case questionnaire and interview data, were explicitly stated, and findings were compared. Apart from corroborated findings that were confirmed by qualitative analysis, including

the similar perceptions of graduate students and teachers on publication-oriented profiles and the similar demand structure; the authors also reported the largest divergence: The networkers, and provided possible explanations for such difference.

It is to be stressed that the purpose of giving these three examples is to present different levels and styles of reporting on methodologically driven credibility techniques, instead of saying one of the ways of reporting is wrong or definitely better than the others. A uniform standard of reporting is not, and should not be encouraged. However, it is necessary to provide sufficient information for the readers to make informed decisions about the credibility of findings.

"New" credibility techniques found in practice. The results showed some interesting cases with useful techniques that are not listed in the codebook or even in literature. In this section, "new" credibility techniques found in practice are introduced. Of course, these techniques are not necessarily new, in the sense that some of them are commonly used by evaluators, just not often considered as credibility techniques; and some of them can also be regarded as variations of existing credibility techniques.

Member checks before interview: Member checks, or member checking is a systematic procedure to share with participants one's data, analysis, interpretations and sometimes conclusions, and obtain their feedback. Therefore member checks is conducted after the data collection. In practice, a technique similar to member checks was used before data collection. For example, Taut and Alkin (2003) had a "pre-interview" before the actual interview, to ask "about what they perceived to be others' attitudes toward evaluation, followed by a question about what were their own attitudes, and a

question about how they thought attitudes toward evaluation develop" (p. 214-215), and they "make a clear distinction between these introductory questions and the main part of the interview" (p. 215). The authors also pointed out the purposes of checking with participants before the interview, to first help "frame the interview," and then to "further helped put into context and appropriately interpret the study's main findings" (p. 215).

Participant debriefing: In evaluations, it is common practice for evaluators to debrief participants on the evaluation purpose, procedures, and progress. However, this approach is often considered as one of the steps of data collection, and reflections on and impact of such technique is usually not reported in evaluation studies. In the study of Bettany-Saltikov, Kilinc, and Stow (2009), using and reporting on participant debriefing can help enhance the credibility of evaluation findings. Bettany et al. conducted an evaluation of the reliability of the University's Masters' level generic assessment criteria. In the data collection stage, lecturers are asked to mark dissertations following the university procedure of blind double marking, then all participants are gathered in a focus group to discuss the marking process and provide feedback. In this case, the authors debriefed the participants on the study procedures so they understand the purpose of the study, and "discussed the criteria together before the study" (p. 637). Participant debriefing was treated more than a procedure in this study, as the authors explained how such a technique may have affected the results, "The group had also previously discussed the criteria together before the study. This may have helped participants develop a common understanding of the criteria before actually marking the dissertations." At the same time, the authors also clarified its possible impact on the transferability of the

finding: “In everyday practice, however, when managing large dissertation modules with large numbers of students and supervisors from different disciplines, this discussion is not always possible” (p. 637).

Peer checks: In the list of methodologically driven credibility techniques of the codebook, expert checking means inviting external evaluator or expert to assess quality of analyses or findings. In the study of Taut and Alkin (2003), the role of external experts was replaced by peers. The authors asked “an evaluation colleague unfamiliar with the study” (p. 216) to review the interview transcripts, “derive main themes from the data,” and “compared his results with our own analyses” (p. 216). The authors found that their categories were “more detailed, that is, closer to the interviewees’ statements than the themes derived by the evaluation colleague, but all of the latter’s themes closely corresponded to the authors’ categories” (p. 217). As illustrated in this example, peer checks can be considered as an extension of expert checks, and not only to perform a general assessment of the finding, but also serve as reliability checks to review the whole analysis process.

Change of plans: Emergent design is one of the characteristics of qualitative study, so that the approach taken by the research may change in respond to changes in the field. In the selected body of evaluation studies, some authors also reported their change of plans during the research process. For example, the changes of research plans were explicitly described in Bradley et al. (2005):

> More focus groups were originally planned, but lack of student time made them impossible. Instead, we held further one-to-one interviews during lunch times nearer

> the students' exams with different students from the cohort. Data were reanalyzed, and a further round of interviews was planned to investigate issues that had not been properly covered, for example, students' experiences of using EBM in clinical placements, as well as to cross-check results. A new interview guide was again developed (available from the authors). The second student cohort was interviewed just before the final teaching block. (p. 162).

The information provided by the authors is comprehensive. They first described the change of plans due to time restrictions, and series of changes to be made to cope with the change, such as reanalysis of the data. The rationales for coping strategies and new plans were also specified. The fact that qualitative approach often involves an emergent and evolving design makes it even more important for the researcher to report the evolving process or changes of original evaluation plans. Knowing what have been altered or modified will help readers to learn about the context and the researcher, to make their own decisions on whether the researcher has engaged in the best practice to obtain information from participants.

Chapter 5: Discussion and Conclusion

To the researcher's best knowledge, this is the first synthesis of the rigor of reported methodology in evaluation studies that use qualitative methods in the field of higher education. This study provides an evidence base about how credibility techniques are reported in evaluation studies that use qualitative methods. The results provide a snapshot of the current landscape of the reporting practice of credibility techniques by describing the types and frequencies of credibility techniques reported, and how the authors choose to report them. In this discussion chapter, an evaluation of the reporting as well as the synthesis methodology is presented, and suggestions are provided for future practice and reporting of credibility techniques.

Qualitative Features

Throughout this empirical synthesis, qualitative method and its unique features have been the essential thread in the fabric of the synthesis efforts. However, findings of the synthesis show that this qualitative thread also happens to be the one that most needed to be strengthened in the use of credibility techniques, and more specifically in number, frequency, and details of description.

First, the number of articles that used qualitative methods is still relatively small. Educational evaluation studies that use qualitative methods take only a small proportion in the number of articles published in top evaluation journals each year. Before narrowing down the inclusion criteria to the higher education field, the *Journal of Assessment, Evaluation, and Accountability* has the highest percentage of evaluation studies that use

qualitative methods published over the ten years: 24.66%, yet the largest proportion is still less than a quarter of the total.

It is undeniable that in the past decade, educational qualitative methodologies thrived and flourished with established authority, expanded qualitative perspectives, vocabularies and approaches, and diversified ways of collaboration with other methods (Lincoln et al., 2011; Sweeney, 2006), yet despite the growing interest in qualitative approach and the increased number of published evaluation studies that use qualitative methods, the overall proportion of such studies remain small in top evaluation journals. Furthermore, in spite of the relatively small percentage of evaluation studies that use qualitative methods in the total published articles, the percentage, as well as the absolute number of qualitative evaluation studies has been increasing with time, and has showed an apparent growth since 2009. In contrast, the average number of credibility techniques reported in these studies does not seem to increase, but rather remains more or less the same over the decade (about 6).

Second, qualitative features as the essential thread are also illustrated in the coding categories of credibility techniques. As explained in the codebook section, the credibility technique categories in the codebook included two parts. The first part contains basic requirements to retain methodological rigor, and such requirements are usually shared in qualitative, quantitative, and mixed methods approaches, such as design, sampling, data collection, and data analysis. The second part, and the majority of coding categories are credibility techniques with distinct qualitative features. These credibility techniques function as a procedure to enhance credibility of the findings, they are, at the

same time, a natural demonstration of characteristics of qualitative research. For example, the emphasis of qualitative research on rich and in-depth description is reflected in the technique of thick description; reflexivity becomes an indispensable feature in qualitative research because the researcher is the primary data collection and analysis instrument. Many credibility techniques, such as searching for rival explanation, compare and contrast, and negative case analysis, all demonstrated the dynamic inductive data analysis process to build abstractions and concepts for theory generating, and the concern with nuance and process (Borbin & Strauss, 2008; Creswell, 2007; Merriam, 2009).

The first research question of this synthesis was that to what extent was credibility techniques reported in selected journals, and despite the seemingly satisfactory use of basic techniques of methodology, the synthesis findings show that qualitative features are not sufficiently expressed through credibility techniques in the selected articles. One possible reason for the lack of reporting credibility techniques with qualitative features could be that compared with the basic techniques of design, techniques with qualitative features are much less applied in practice of qualitative methods. Over 80% of the articles reported using basic techniques like design, sampling, data collection, and data analysis, and provided a reasonable amount of information on these four techniques. Comparatively, credibility techniques with more distinct qualitative features are much less used. On the whole, credibility credibility techniques were infrequently and incompletely reported. Well-known credibility techniques are reported in less than one third of the articles, and the majority of methodologically driven techniques are used in less than 10% of articles. Among those studies where some credibility techniques are

mentioned, there were often not enough specifics to reliably code details of the practice of such techniques.

Behind the absence of credibility techniques with qualitative features, there is a large potential for more of these techniques to be used. To take the simplest example, the technique, member checks, is usually used when particular qualitative methods such as the interview or focus group are applied, but could also be used with observation or document analysis. In the selected body of articles, 72.88% of the articles used either interview or focus group, but only 11.9% explicitly reported using member checks to share with participants their data and/or analyses. For another example, triangulation is the most commonly used credibility technique, but only 32 out of 55 studies that used mixed methods reported triangulation (20 qualitative studies used triangulation). As it is pointed out by Tashakkori and Teddlie (2003) and Plano Clark (2008), triangulation is the fundamental principle of mixed methods study and a basic design, still quite a number of articles used both quantitative and qualitative methods (many claimed to be mixed methods studies) have strictly separated section for each approach.

In fact, there could be an interaction effect between the above findings: The low publication rate of evaluation studies that use qualitative methods, the low presence rate of credibility techniques, and very likely spacial constraints of journals and possibly limited understanding of editors on qualitative methods. The interacting factors could form a pernicious cycle: Sometimes, editors might be concerned about quality of the qualitative evaluation findings if no credibility techniques were used; quality concerns yield low publication rate of such studies, discouraging authors sending their work to

these journals; and as a result, reviewers and editors get to read less studies that use qualitative methods and become less experienced, and their understanding of the field was undermined.

Thus, the above findings become convergence evidence that suggests the need for more attention to credibility techniques, and qualitative methods in the field of educational evaluation. It is true that qualitative approach is time consuming, and requires highly of the researcher especially in terms of commitment and engagement, nonetheless, it is a rigorous methodology when applied properly, it is required by evaluation standards, and credibility techniques are useful and powerful procedures to promote methodological rigor and validity of the finding.

Details

The second group of research questions and sub-questions focused on characteristics of supporting details when credibility techniques are reported. There is a large variation in the narrative reporting of design related techniques. The descriptions show colorful methodological diversity of studies that are published over the past ten years. Seeing all selected articles as a whole, there is high degree of consistency between the concepts of credibility techniques in the theoretical literature and the way they are understood and used in these empirical studies. Almost all theoretically defined aspects of the design related techniques in literature are covered in the selected body of articles, and the organization of supporting details for these basic methodological techniques is often efficient, and even creative. What especially needs to be highlighted is the flexible style of reporting. Report writers usually need more space to provide details, as it is

confirmed by quantitative analyses that page length was positively correlated with credibility techniques. Yet qualitative content analyses show there are many examples where supporting details do not necessarily need to be lengthy to be comprehensive. Those studies are either skillful in presenting rich information in a concise style, or combined writing with on-line resources. Such a finding is an addition to previous findings that suggesting space limitations imposed by journal editors undermined authors' capacity of fully describe their methodology (Dixon-Woods et al., 2007; Shek et al., 2005). The practice found in selected articles could be examples of possible solutions to space constraints.

It is encouraging to see the diversity in reporting of credibility techniques, although only in the reporting of several design related techniques. There are numerous principles and methods in qualitative research that evaluators can use, and credibility techniques, together with qualitative methods, are still a growing field being defined and enriched with new techniques and approaches. Therefore it is only natural to see diversified way of practice and reporting. Just as evaluation standards do not have rigid procedures on how to do an evaluation, to be in accordance with the vibrant ways of qualitative methods, there should be only generic principles on how to report the use of credibility techniques.

Comparatively, there is a much bigger and more apparent gap between research and practice in methodologically driven techniques. Qualitative research requires researchers to know these credibility techniques well enough so that they can draw on the techniques to not only apply them to context, but also to blend them, justify the choices,

and describe them in sufficient details. In the selected body of articles, it appears that there is not sufficient awareness of methodologically driven credibility techniques, and only one article has a combined use of two methodologically driven techniques. In addition to the lack of addressing qualitative credibility with credibility techniques, in articles that used mixed methods, more information was provided for the quantitative component, and there is no address of the overall mixed methods validity in any of the selected articles.

Language and Terminology

The last group of sub research questions is with regard to reporting language and term use. A good understanding of common terminology is prerequisite for qualitative evaluation efforts, and language with qualitative characteristics should be encouraged for describing the research process to enhance both the practice and communication of research. Reporting qualitative methods with common terminologies can help readers and editors of the field to efficiently define and assess the qualitative approaches adopted, especially when the methods used are established framework or approaches. According to findings of this synthesis research, basic methodological terms can be considered widely used, but qualitative credibility terms are much less popular. The selected body of articles shows a number of typical examples of presenting a credibility technique without using any term but with great detail, and contrary cases of claiming using a technique but reported not much detail. When terminologies are used, the same credibility technique is often presented with different terms, indicating slightly different connotations. What to note is credibility technique terms could be encouraged in practice, but for real

convenience of collaborative input, assessment, and communication, terminologies cannot replace details in reporting, but should rather be combined with specifics of the particular study context to truly promote reliability and credibility of the study.

The Methodology: Implications for Methods and Practice

By using the mixed methods synthesis methodology, this study produced an empirical knowledge base with evidence that has not been collected or presented before. Therefore, using both narrative and statistical methods, this study demonstrated the viability of a research synthesis method in capturing a snapshot of the field of educational evaluation studies. Furthermore, coding such knowledge using mixed methods becomes feasible via the codebook, which has proven to be a useful tool. Coding of empirical evaluation studies helped the researcher to connect theoretical credibility categories with practical use of credibility techniques, it not only provides an empirical basis for understanding the current state of the field, but also indicated the need for additions to the codebook, including possible new credibility techniques and important factors relating to the use of credibility techniques in the empirical literature. Through identifying and documenting the reporting of credibility techniques, the systematic coding of both frequencies and text created an operationalization of credibility concepts.

In addition, methodological rigor of this synthesis study was enhanced through suitable design and planning, transparent and explicit implementation, and rigorous interpretation. Following quality criteria of quantitative, qualitative, mixed method study, and quality standards of synthesis study, the reliability, credibility, and overall validity of this study were constantly checked by means of design related and credibility techniques

throughout the research process. The researcher's reflections on the iterative process of checking and promoting credibility of this study were also documented.

Conclusion

Conclusions of this study can be drawn from both the synthesis methodology and the research findings. Mixed methods synthesis could be considered as a feasible and useful method to describe the general state of certain field, and to understand subtle details. Synthesis researchers should outline the synthesis with a synthesis protocol, and construct a codebook, details of which are described in Chapter 3. Using both qualitative and quantitative methods, the design of the codebook is recommended to be flexible enough to accommodate both quantitative variables and qualitative descriptions and interpretations. More importantly, synthesis researchers need to be reflexive and transparent, to be explicit about their methods and to document their practice carefully.

With regards to research findings, the reviewed body of articles shows two major characteristics. First, evaluations that use qualitative methods in higher education field take only a small portion of published articles, but their reporting of credibility techniques has been relatively steady over the past decade. Despite the increased use of qualitative methods in program evaluation in recent years, it is not commonly applied in published educational evaluations. However, selected articles in top evaluation journals demonstrate strong stability in their practice of addressing credibility. Credibility techniques are not frequently used, but based on the researcher's read of the article, the reported use of credibility techniques made sense and helped promote rigor of the overall design. In this sense, the average use of six credibility techniques seems acceptable.

Second, the majority of selected articles are able to maintain basic methodological rigor by using most of design related techniques, but it appears that most authors are not sensitive to qualitative features in reporting credibility. Authors are more familiar with design related techniques, and many showed skillful use of techniques such as design and data collection methods, but sometimes lack of explicitness of certain details in techniques like sampling and analysis. In contrast, credibility techniques with strong qualitative features (e.g. reflexivity, and methodologically driven techniques) are much less commonly used and elaborated in reporting. Furthermore, because of the low presence of these credibility techniques, there is hardly any pattern that could be discerned from quantitative data, and descriptions of qualitative text analyses were also impoverished. Triangulated results of both data sources suggest that credibility techniques with qualitative features are underreported in selected journals. Such a finding meets a priori expectation of the researcher, and reflected the current practice of the evaluation field. With that said, the qualitative analyses still captured the limited details describing reported credibility techniques, and it is found that methodology and findings of the article are more rigorous when more credibility techniques were reported, and when different aspects of a technique were reported.

With qualitative methods becoming a widely used methodology, it is important for authors to address the credibility of their data, and use credibility techniques, especially methodologically driven techniques to enhance credibility of their findings. It could be concluded that more awareness, informative use and improved reporting of credibility techniques are needed to promote methodological progress and better quality

of educational evaluation studies, and in turn, increased quality evidence may possibly yield more publications of evaluations that use qualitative methods. Measures such as adequate training of qualitative evaluators and more explicit guidelines provided by journals could also be put forward to narrow the gap between theoretical development and practice of credibility techniques.

Limitation and Future Research

The strength of this study is its ability to summarize existing knowledge, which can help foster an evidence-based practice for the use of qualitative methods. However, its inherent dependence on what is reported in the literature is also a limitation of this and all syntheses. As data for syntheses in general, the data from this systematic synthesis are only as comprehensive as what are reported for each case in the selected articles. It seems reasonable to assume that the major methodological elements would have been reported by the authors. Nevertheless, it remains possible that there was some underreporting of credibility practice, techniques or relevant details. In addition, there are undoubtedly more elements to be added to the codebook, more analyses to be conducted to answer additional questions, and qualitative evaluation efforts of a broader range to be reviewed.

In summary, this research synthesis, following each step specified in the synthesis protocol, aimed to achieve the goal of painting a reliable description of the current practice of credibility techniques, and to provide a preliminary assessment of factors related to such practice. The published evaluation studies that use qualitative methods in the field of higher education have presented valuable information on the practice of credibility techniques, but it always has much room for growth and refinement. This

synthesis highlights the credibility technique frequency, reporting style, and terminologies, with some preliminary results of relationships between credibility techniques and background factors. Future research for the field may include more complete description of the credibility technique reporting, perceptions of article authors, and evaluation of the actual process of using credibility techniques. This would contribute to achieving long-terms goals of quality assessment of the qualitative evaluation field, and provide explanations for the current qualitative practice.

References

AACTE. (2002). *Governmental relations update*. American Association of College for Teacher Education.

Agen, M. T. (2000). Evaluating interpretive inquiry: Reviewing the validity debate and opening the dialogue. *Qualitative Health Research, 10*, 378-395.

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82*(4), 436-476. doi: 10.3102/0034654312458162

Albainese, M., & Norcini, J. (2002). Systematic reviews: What are they and why should we care? *Advances in Health Science Education, 7*, 147-151.

Andrade, Heidi, & Du, Ying. (2007). Student responses to criteria‐referenced self‐assessment. *Assessment & Evaluation in Higher Education, 32*(2), 159-181. doi: 10.1080/02602930600801928

Asghar, Amanda. (2010). Reciprocal peer coaching and its use as a formative assessment strategy for first‐year students. *Assessment & Evaluation in Higher Education, 35*(4), 403-417. doi: 10.1080/02602930902862834

Atkins, Salla, Launiala, Annika, Kagahn, Alexander, & Smith, Helen. (2012). Including mixed methods research in systematic reviews: Examples from qualitative syntheses in TB and malaria control. *BMC medical research methodology, 12*(62), 1-7.

Barber, J. P. , & Walczak, K. K. . (2009, April). *Conscience and critic: Peer debriefing strategies in grounded theory research*. Paper presented at the the annual meeting of the American Educational Research Association, San Diego, CA.

Barusch, Amanda, Gringeri, Christina, & George, Molly. (2011). Rigor in qualitative social work research: A review of strategies used in published articles. *Social Work Research, 35*(1), 11-19.

Beck, C. T. (1993). Qualitative research: The evaluation of its credibility, fittingness, and auditability. *Western Journal of Nursing Research, 15*(2), 263-266. doi: 10.1177/019394599301500212

Berlin, J. A., & Ghersi, D. (2005). Preventing publication bias: Registries and prospective meta-analysis. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 35-48). Chichester, UK: Wiley.

Bettany‐Saltikov, Josette, Kilinc, Stephanie, & Stow, Karen. (2009). Bones, boys, bombs and booze: an exploratory study of the reliability of marking dissertations across disciplines. *Assessment & Evaluation in Higher Education, 34*(6), 621-639. doi: 10.1080/02602930802302196

Bochner, A.P. (2000). criteria against ourselves. *Qualitative Inquiry, 6*(2), 266-272.

Borg, Erik. (2009). Local plagiarisms. *Assessment & Evaluation in Higher Education, 34*(4), 415-426. doi: 10.1080/02602930802075115

Boruch, R. F., & Petrosino, A. (2010). Meta-analyses, systematic reviews, and evaluation syntheses. In J. S. Wholey, H. P. Hatry & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 531-556). San Francisco, CA: Jossey-Bass.

Bourque, L. B. (2004). Coding frame. In M. S. Lewis-Beck, A. E. Bryman & T. F. Liao (Eds.), *The Sage encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.

Boyatzis, R. E. . (1998). *Transforming qualitative information: Thematic analysis and code development.* Thousand Oaks, CA: Sage.

Bradley, P., Oterholt, C., Nordheim, L., & Bjorndal, A. (2005). Medical students' and tutors' experiences of directed and self-directed learning programs in evidence-based medicine: a qualitative evaluation accompanying a randomized controlled trial. *Evaluation Review, 29*(2), 149-177. doi: 10.1177/0193841X04269085

Brandon, P. R., & Singh, J. M. (2009). The strength of the methodological warrants for the findings of research on program evaluation use. *American Journal of Evaluation, 30*(2), 123-157. doi: 10.1177/1098214009334507

Brantlinger, Ellen, Klingner, Janette, Richardson, Virginia, & Taylor, Steven J. (2005). Importance of experimental as well as empirical qualitative studies in special education. *Mental Retardation, 43*(2), 92-119.

Breuer, F., Mruck, K., & Roth, W. (2002). Subjectivity and reflexivity: An introduction. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 3*(3). http://www.qualitative-research.net/index.php/fqs/article/view/822/1785

Bryman, A. (Ed.). (2004). *Social research methods* (2nd ed.). Oxford: Oxford University Press.

Buchanan, D. A., & Bryman, A. (2007). Contextualizing methods choice in organizational research. *Organizational Research Methods, 10*(3), 483-501.

Burden, Peter. (2010). Creating confusion or creative evaluation? The use of student evaluation of teaching surveys in Japanese tertiary education. *Educational Assessment, Evaluation and Accountability, 22*(2), 97-117. doi: 10.1007/s11092-010-9093-z

Burns, N., & Grove, S. (2001). *The practice of nursing research: Conduct, critique and utilization* (4th ed.). Philadelphia, PA: W. B. Saunders.

Campbell, D. T. (1988). *Methodology and epistemology for social science: Selected paper*. Chicago: University of Chicago Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health Professions, 25*(1), 12-37. doi: 10.1177/0163278702025001003

Chatterji, M. (2008). Comments on Slavin: Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher, 37*(1), 23-26. doi: 10.3102/0013189x08314287

Chen, H. . (2005). *Practical program evaluation: Assessing and improve planning, implementation, and effectiveness.* Thousand Oaks: CA: Sage.

Cheng, Liying, Rogers, W. Todd, & Wang, Xiaoying. (2008). Assessment purposes and procedures in ESL/EFL classrooms. *Assessment & Evaluation in Higher Education, 33*(1), 9-32. doi: 10.1080/02602930601122555

Cho, J., & Trent. (2006). Validity in qualitative research revisited. *Qualitative Research, 6*(3), 319-340. doi: 10.1177/1468794106065006

Cooper, H. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52*(2), 291-302. doi: 10.3102/00346543052002291

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-analsysis* (2nd ed., pp. 4-16). New York: Russell Sage Foundation.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analsysis* (2nd ed.). New York: Russell Sage Foundation.

Cotton, D. R. E., & Gresty, K. A. (2007). The rhetoric and reality of e‐learning: using the think‐aloud method to evaluate an online resource. *Assessment & Evaluation in Higher Education, 32*(5), 583-600. doi: 10.1080/02602930601116920

Craddock, Deborah, & Mathias, Haydn. (2009). Assessment options in higher education. *Assessment & Evaluation in Higher Education, 34*(2), 127-140. doi: 10.1080/02602930801956026

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River: Pearson.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2012a). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston: Pearson.

Creswell, J. W. (2012b). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage

Creswell, J. W., & Miller, D. . (2000). Determining validity in qualitative inquiry. *Theory Into Practice, 39*(3), 124-130.

Creswell, J. W., & Tashakkori, A. (2007). Editorial: Differing perspectives on mixed methods research. *Journal of Mixed Methods Research, 1*(4), 303-308. doi: 10.1177/1558689807306132

Crocker, L., & Algina, J. . (1986). *Introduction to classical and modern test theory.* Philadelphia: Harcourt Brace Jovanovich College Publishers.

Curtin, Michael, & Fossey, Ellie. (2007). Appraising the trustworthiness of qualitative studies: Guidelines for occupational therapists. *Australian Occupational Therapy Journal, 54*(2), 88-94. doi: 10.1111/j.1440-1630.2007.00661.x

Cutcliffe, J. R. (2001). Establishing the credibility of qualitative research findings: the plot thickens. *Journal of Advanced Nursing, 30*(2), 374-380.

De Filippo, Daniela, Casado, Elías Sanz, & Gómez, Isabel. (2009). Quantitative and qualitative approaches to the study of mobility and scientific performance: a case study of a Spanish university. *Research Evaluation, 18*(3), 191-200. doi: 10.3152/095820209x451032

De Witt, L., & Ploeg, J. (2006). Critical appraisal of rigour in interpretive phenomenological nursing research. *Journal of Advanced Nursing, 55*(2), 215-229. doi: 10.1111/j.1365-2648.2006.03898.x

DeCuir-Gunby, Jessica T., Marshall, Patricia L., & McCulloch, Allison W. (2011). Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field Methods, 23*(2), 136-155. doi: 10.1177/1525822x10388468

Denk, Nikola, Kaufmann, Lutz, & Carter, Craig R. (2012). Increasing the rigor of grounded theory research – A review of the SCM literature. *International Journal of Physical Distribution & Logistics Management, 42*(89), 742-763. doi: 10.1108/09600031211269730

Denzin, N. K. (1989). *Interpretive interactionism*. Newbury Park, CA: Sage.

Denzin, N. K. (2002). Social work in the seventh moment. *Qualitative Social Work, 12*, 25-23.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2011). *The Sage handbook of qualitative research* (4th ed.). Thousand Oaks, CA: Sage.

Dixon-Woods, M., Booth, A. , & Sutton, A. J. (2007). Synthesizing qualitative research: A review of published reports. *Qualitative Research, 7*(3), 375-422.

Dixon-Woods, M., Shaw, R. L., Agarwal, S., & Smith, J.A. . (2004). The problem of appraising qualitative research. *Quality & Safety in Health Care, 13*, 223–225.

Doyle, S. . (2007). Member checking with older women: A framework for negotiating meaning. *Health Care for Women International, 8*(10), 888-908.

Easterby-Smith, M, Golden-Biddle, K, & Locke, K. (2008). Working with pluralism: Determining quality in qualitative research. *Organizational Research Methods, 11*, 419-429.

Eisner, E. W. (1979). *The educational imagination : On the design and evaluation of school programs*. New York: Macmillan.

Eisner, E. W. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.

Ellington, L. L. . (2008). *Engaging crystallization in qualitative research*. Thousand Oaks, CA: Sage.

Ely, M., Anzul, M., Friedman, T., Garner, D., & Steinmetz, A. M. (1991). *Doing qualitative research: Circles within circles.* London: Falmer Press.

EPPI-Centre. (2006). EPPI-Centre methods for conducting systematic reviews. London: EPPI-Centre, Social Science Research Unit, Institute of Educaiton, University of London.

European Evaluation Society. (2012). European Evaluation Society Library. Retrieved Feb, 23, 2013, from http://www.europeanevaluation.org/library/evaluation-standards/national-and-regional-evaluation-societies/1261848882-europe.htm

Fade, S .A. . (2003). Communicating and judging the quality of qualitative research: The need for a new language. *Journal of Human Nutrition and Dietetics, 16*, 139—149.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R. , & Karageorgopoulos, D. E. . (2008). Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal, 22*(8), 2623-2628.

Feldman, K. A. . (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education, 4*(1), 86–102.

Frank, M., & Barzilai, A. (2004). Integrating alternative assessment in a project-based learning course for pre-service science and technology teachers. *Assessment & Evaluation in Higher Education, 29*(1), 41-61.

Gephart, R. (2004). From the editors: Qualitative research and the academy of management. *Academy of Management Journal, 47*(4), 454-462.

Gersten, R., & Hitchcock, J. (2009). What is credible evidence in education? The role of the What Works Clearinghouse in informing the process. In S. I. Donaldson, C. A. Christie & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice* (pp. 78-95). Thousand Oaks, Calif: Sage.

Giddens, A. (1984). *The constitution of society.* Cambridge, UK: Polity.

Gijbels, D., van de Watering, G., & Dochy, F. (2005). Integrating assessment tasks in a problem-based learning envrionment. *Assessment & Evaluation in Higher Education, 30*(1), 73-86.

Glensne, C. (2011). *Becoming qualitative researchers: An introduction* (4th ed.). Boston: Pearson Education.

Glesne, C., & Peshkin, A. (1992). *Becoming qualitative researchers: An introduction.* White Plains, NY: Longman.

Golafshani, N. . (2003). Understanding reliability and validity in qualitative research. *Qualitative Report, 8*, 597-606.

Goodwin, L. D., & Goodwin, W. L. . (1985). Statistical techniques in AERJ articles, 1979–1983: The preparation of graduate students to read the educational research literature. *Educational Researcher, 14*(2), 5-11.

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.

Guba, E. G., & Lincoln, Y. S. . (1981). *Effective evaluation*. San Francisco: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. . (2005). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 191-216). Thousand Oaks, CA: Sage.

Guz, A.N. , & Rushchitsky, J. J. . (2009). Scopus: A system for the evaluation of scientific journals. *International Applied Mechanics, 45*(4), 351-362.

Halpern, E. S. (1983). *Auditing naturalistic inquiries: The development and application of a model.* (Unpublished doctoral dissertation), Indiana University.

Hammouri, Hind A. M. (2003). An investigation of undergraduates' transformational problem solving strategies: cognitive/metacognitive processes as predictors of holistic/analytic strategies. *Assessment & Evaluation in Higher Education, 28*(6), 571-586. doi: 10.1080/0260293032000130225

Hannes, K., Lockwood, C., & Pearson, A. (2010). A comparative analysis of three online appraisal instruments' ability to assess validity in qualitative research. *Qualitative Health Research, 20*(12), 1736-1743. doi: 10.1177/1049732310378656

Hannes, K., & Macaitis, K. (2012). A move to more systematic and transparent approaches in qualitative evidence synthesis: Update on a review of published papers. *Qualitative Research, 12*(4), 402-442. doi: 10.1177/1468794111432992

Hardcastle, M. A., Usher, K., & Holmes, C. (2006). Carspecken's five-stage critical qualitative research method: an application to nursing research. *Qualitative Health Research, 16*(1), 151-161. doi: 10.1177/1049732305283998

Harden, A. (2010). Mixed-Methods systematic reviews: Integrating quantitative and qualitative findings. *Focus Technical Brief, 25*, 1-8.

Harden, A., & Thomas, J. (2010). Mixed methods and systematic reviews. In A. Tashakkori & C. Teddlie (Eds.), *Sage Handbook of Mixed Methods in Social & Behavioral Research* (pp. 749–774). Thousand Oaks: Sage.

Hay, Peter J., Engstrom, Craig, Green, Anita, Friis, Peter, Dickens, Sue, & Macdonald, Doune. (2012). Promoting assessment efficacy through an integrated system for online clinical assessment of practical skills. *Assessment & Evaluation in Higher Education, 38*(5), 520-535. doi: 10.1080/02602938.2012.658019

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. New Directions for Evaluation, No. 97 (pp. 69–80). San Francisco, CA: Jossey-Bass.

Hesse-Biber, S.N. . (2010). *Mixed method research: Merging theory with practice*. New York: Guilford Press.

Heyvaert, M., Maes, B., & Onghena, P. (2011). Mixed methods research synthesis: definition, framework, and potential. *Quality & Quantity, 47*, 659-676.

Higgins, JP.T., & Green, S (2011). Cochrane handbook for systematic reviews of interventions version 5.1.0. Retrieved from The Cochrane Collaboration website: http://www.cochrane-handbook.org

Hitchcock, J, & Newman, I. (2012). Applying an interactive quantitative-qualitative framework: How identifying common intent can enhance inquiry. *Human Resources Development Review, 12*(1), 36-52.

Hogan, R. Lance. (2007). The historical development of program evaluation: Exploring the past and present. *Online Journal of Workforce Educaiton and Development, 2*(4), 1-14.

Holloway, I., & Wheeler, S. (1996). *Qualitative research for nurses*: Blackwell Science, Oxford.

Hopewell, S., Clarke, M., & Mallett, S. (2005). Grey literature and systematic reviews. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 49-72). Chichester, UK: Wiley.

Hoskins, K. . (1968). The examination, disciplinary power and rational schooling. *History of Education, 8*(1), 135-146.

House, E.R. . (2005). Qualitative evaluation and changing social policy. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 1069-1082). Thousand Oaks: Sage.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288. doi: 10.1177/1049732305276687

Huberman, A. M., & Miles, M. B. (1994). *Qualitative data analysis: An expanded sourcebook.* Thousand Oaks, CA: Sage.

Hutchinson, Susan R., & Lovell, Cheryl D. (2004). A review of methodological characteristics of research published in key journals in higher education: Implications for graduate research training. *Research in Higher Education, 45*(4), 383-403.

Huxham, Mark, Campbell, Fiona, & Westwood, Jenny. (2012). Oral versus written assessments: a test of student performance and attitudes. *Assessment & Evaluation in Higher Education, 37*(1), 125-136. doi: 10.1080/02602938.2010.515012

Johnson, R. B., & Christensen, L. (2012). *Educational research: Quantitative, qualitative, and mixed approaches* (4th ed.). Los Angeles: Sage Publications.

Johnson, R. B., & Onwuegbuzie, A. J. . (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133. doi: 10.1177/1558689806298224

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*. Newbury Park, CA: Sage.

Kitchenham, B. . (2004). Procedures for performing systematic reviews. Keele, Eversleigh : Keele University and NICTA, Technical Report. http://www.idi.ntnu.no/emner/empse/papers/kitchenham_2004.pdf

Koro-Ljungberg, Mirkai, & Douglas, Ellip P. (2008). State of qualitative research in engineering education: Meta-analysis ofJEE articles, 2005-2006. *Journal of Engineering Education*, 163-175.

Kushner, S. (2005). Qualitative control: A review of the framework for assessing qualitative evaluation. *Evaluation, 11*(1), 111-122. doi: 10.1177/1356389005053194

Kvale, S. . (1996). *Interviews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.

Lamont, M., Mallard, G., & Guetzkow, J. (2006). Beyond blind faith: Overcoming the obstacles to interdisciplinary evaluation. *Research Evaluation, 15*(1), 43-55.

Lather, P. . (1991). *Getting smart: Feminist research and pedagogy with/in the postmodern*. New York: Routledge.

Lather, P. . (1993). Fertile obsession: Validity after poststructuralism. *Sociological Quarterly, 34*(4), 673–693.

Lewis, J. (2009). Redefining qualitative methods: Believability in the fifth moment. *international Journal of Qualitatvie Methods, 8*(2), 2-14.

Liao, H., & Hitchcock, J. (2012). *Development of a codebook for examining the application of credibility techniques in program evaluations that use qualitative methods*. Paper presented at the Annual Meeting of the Mid- Western Educational Research Association (MWERA), Evanston, Illinois.

Lichtman, M. . (2010). *Qualitative research in education: A user's guide.* Thousands Oaks, CA: Sage.

Lietz, C. A., & Zayas, L. E. (2010). Evaluating qualitative research for social work practitioners. *Advances in Social Work, 11*(2), 188-202.

Lincoln, Y. S. . (1995). Emerging criteria for quality in qualitative and interpretive research. *Qualitative Inquiry, 12*, 275—289.

Lincoln, Y. S. , & Guba, E. A. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Lincoln, Y. S. , Lynham, S. A., & Guba, E. G. (2011). Pragmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (4th ed., pp. 97-128). Thousand Oaks, CA: Sage.

MacQueen, Kathleen M., McLellan, Eleanor, Kay, Kelly, & Milstein, Bobby. (2009). Codebook development for team-based qualitative analysis. In K. Krippendorff & M. A. Bock (Eds.), *The Content Analysis Reader* (pp. 211-219). Thousand Oaks, CA: Sage.

Major, C., & Savin-Baden, M. . (2010). *An introduction to qualitative research synthesis: Managing the information explosion in social science research*. London: Routledge.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey Bass.

Marshall, C., & Rossman, G. B. (2011). *Designing Qualitative Research* (5th ed.). Thousand Oaks, CA: Sage.

Mathison, S. (1988). Why triangulate? *Educational Researcher, 17*(2), 13-17.

Maxwell, J. A. . (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*(3), 279-299.

Maxwell, J. A. . (1996). *Qualitative research design.* Newbury Park, CA: Sage.

Maxwell, J. A. . (2004). Casual explanation, qualitative research, and scientific Inquiry in education. *Educational Researcher, 33*(2), 3-11.

Maxwell, J. A. . (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.

McCormack, Coralie. (2005). Reconceptualizing student evaluation of teaching: an ethical framework for changing times. *Assessment & Evaluation in Higher Education, 30*(5), 463-476. doi: 10.1080/02602930500186925

McDavid, J. C., & Hawthorn, L.R.L. (2006). *Program evaluation & performance measurement: An introduction to practice.* Thousand Oaks, CA: Sage Publications.

Meagher, Laura, Lyall, Catherine, & Nutley, Sandra. (2008). Flows of knowledge, expertise and influence: a method for assessing policy and practice impacts from

social science research. *Research Evaluation, 17*(3), 163-173. doi: 10.3152/095820208x331720

Micari, M., Light, G., Calkins, S., & Streitwieser, B. (2007). Assessment Beyond Performance: Phenomenography in Educational Evaluation. *American Journal of Evaluation, 28*(4), 458-476. doi: 10.1177/1098214007308024

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation, 27*, 296-319. doi: 10.1177/1098214006291015

Morgan, D. . (2008). Emergent design. In L. Given (Ed.), *The SAGE encyclopedia of qualitative research methods* (pp. 246-249). Thousand Oaks, CA: SAGE Publications, Inc.

Nastasi, B. K. , & Schensul, S. L. . (2005). Contributions of qualitative research to the validity of intervention research. *Journal of School Psychology, 43*(3), 177-195.

National Research Council. (2002). *Scientific Research in Education*. Washington, DC: National Academy Press.

Newman, I. , & Benz, C. R. (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum.* Carbondale, Illinois: Southern Illinois University Press.

Newman, I. , & Hitchcock, J.H. . (2011). Underlying agreements between quantitative and qualitative research: The short and tall of it all. *Human Resources Development Review, 10*(4), 381-398.

Onwuegbuzie, A. J., & Leech, N. L. (2006). Validity and qualitative research: An oxymoron? *Quality & Quantity, 41*(2), 233-249. doi: 10.1007/s11135-006-9000-3

Pace, Romina, Pluye, Pierre, Bartlett, Gillian, Macaulay, Ann C., Salsberg, Jon, Jagosh, Justin, & Seller, Robbyn. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies, 49*, 47-53. doi: 10.1016/j.ijnurstu.2011.07.002

Parker, Ian. (2004). Criteria for qualitative research in psychology. *Qualitative Research in Psychology, 1*(2), 95-106. doi: 10.1191/1478088704qp010oa

Patton, M. Q. . (1980). *Qualitative evaluation methods*. Beverly Hills: Sage.

Patton, M. Q. . (1999). Enhancing the quality and credibility of qualitative analysis. *HSR: Health Services Research, 34*(5 Pt 2), 1189-1208.

Patton, M. Q. . (2003). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks: Sage.

Patton, M. Q. . (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Perakyla, A. (1997). Reliability and validity in research based on transcripts. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice*. London: Sage.

Pluye, P., Gagnon, M. P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews.

International Journal of Nursing Studies, 46(4), 529-546. doi: 10.1016/j.ijnurstu.2009.01.009

Polkinghorne, D. E. (1989). Phenomenological research methods. In R. S. Valle & S. Halling (Eds.), *Existential-phenomenological perspectvies in psychology* (pp. 41-60). New York: Plenum Press.

Poulos, Ann, & Mahony, Mary Jane. (2008). Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Education, 33*(2), 143-154. doi: 10.1080/02602930601127869

Prades, Anna, & Espinar, Sebastían Rodríguez. (2010). Laboratory assessment in chemistry: an analysis of the adequacy of the assessment process. *Assessment & Evaluation in Higher Education, 35*(4), 449-461. doi: 10.1080/02602930902862867

Preissle, J. (1996). List of journals friendly to qualitative work. *QUALRS-L, Sep 22*.

Price, E. G., Beach, M. C., Gary, T. L., Robinson, K. A. , Gozu, A., Palacio, Ana, . . . Cooper, Lisa A. (2005). A systematic review of the methodological rigor of studies evaluating cultural competence training of health professionals. *Academic Medicine, 80*(6), 578-586.

Prins, Frans J., Sluijsmans, Dominique M. A., Kirschner, Paul A., & Strijbos, Jan - Willem. (2005). Formative peer assessment in a CSCL environment: a case study. *Assessment & Evaluation in Higher Education, 30*(4), 417-444. doi: 10.1080/02602930500099219

Putnam, H. (1990). *Realism with a human face.* Cambridge, MA: Harvard University Press.

Richardson, L. . (1997). *Fields of play: Constructing an academic Life.* New Brunswick, NJ: Rutgers University Press.

Richardson, L. . (2000). Evaluating ethnography. *Qualitative Inquiry, 6*, 253-256.

Rolfe, G. (2000). Postmodernism: The challenge to empirical research. In G. Rolfe (Ed.), *Research, truth, authority: Postmodern perspectives on nursing* (pp. 28–45). London: Macmillan.

Rossi, Peter H., Lipsey, Mark W., & Freeman, Howard E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 1-8). Chichester, UK: Wiley.

Russ-Eft, D., Bober, M. J., de la Teja, I., Foxon, M. J., & Koszalka, T. A. (2008). *Evaluator competencies standards for the practice of evaluation in organizations.* San Francisco: Jossey-Bass.

Sandelowski, M, & Barroso, J. (2007). *Handbook for synthesizing qualitative research*. New York: Springer Publishing Company.

Sandelowski, M, Voils, Corrine I., & Barroso, Julie (2006). Defining and designing mixed research synthesis studies. *Research in the Schools, 13*(1), 29-40.

Sandelowski, M, Voils, Corrine I., Leeman, Jennifer , & Crandell, Jamie L. . (2012). Mapping the mixed methods–mixed research synthesis terrain. *Journal of Mixed Methods Research, 6*(4), 317–331. doi: 10.1177/1558689811427913

Schmoch, Ulrich, Schubert, Torben, Jansen, Dorothea, Heidler, Richard, & von Görtz, Regina. (2010). How to use indicators to measure scientific performance: a balanced approach. *Research Evaluation, 19*(1), 2-18. doi: 10.3152/095820210x492477

Schofield, J. W. (1990). Increasing the generalizability of qualitative research. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 201-232). New York: Teachers College Press.

Schreier, M. (2012). *Qualitative content analysis in practice.* Thousand Oaks, CA: Sage.

Schwandt, T. A. (1996). Farewell to criteriology. *Qualitative Inquiry, 2*(1), 58-72.

SCImago Journal and Country Rank. (2013). SCImago Research Group. Retrieved Mar 18, 2013, from http://www.scimagojr.com

Scott, G., & Garner, R. (2013). *Doing qualitative research: Design, methods, and techniques*. Upper Saddle River, N.J.: Pearson Education.

Scriven, M. . (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & E. C. Phillips (Eds.), *Evaluation and Education: A Quarter Century*. Chicago: University of Chicago Press.

Scriven, M. . (1996). The theory behind practical evaluation. *Evaluation, 2*(4), 393-404.

Seale, C. . (2002). Quality issues in qualitative inquiry. *Qualitative Social Work, 12*, 97-110.

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

Shek, D. T. L., Tang, V. M. Y., & Han, X. Y. (2005). Evaluation of evaluation studies using qualitative research methods in the social work literature (1990-2003): Evidence that constitutes a wake-up call. *Research on Social Work Practice, 15*(3), 180-194. doi: 10.1177/1049731504271603

Silverman, D. (1993). *Interpreting qualitative data: Methods for analyzing talk, text and interaction*. London: Sage.

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: the development and evaluation of a new tool. *Journal of evaluation in clinical practice, 18*(4), 746-752. doi: 10.1111/j.1365-2753.2011.01662.x

Smaling, A. (2003). Inductive, analogical, and communicative generalization. *international Journal of Qualitatvie Methods, 2*(1), Article 5.

Spencer, L., Ritchie, J., Lewis, J. , & Dillon, L. . (2003). *Quality in qualitative evaluation: A framework for assessing research evidence*. London: Cabinet Office.

Stufflebeam, D. L. (2004). A note on the purpose, development, and applicability of the Joint Committee Evaluation Standards. *American Journal of Evaluation, 25*(1), 99-102.

Stufflebeam, D. L., Madaus, G.F., & Kellaghan, T. . (2000). *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.). Boston: Kluwer Academic Publishers.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). CIPP model for evaluation: An improvement/accountability approach. In D. L. Stufflebeam (Ed.), *Evaluation theory, models, and applications* (pp. 325-365). San Francisco: Jossey-Bass.

Suri, H., & Clarke, D. (2009). Advancements in Research Synthesis Methods: From a Methodologically Inclusive Perspective. *Review of Educational Research, 79*(1), 395-430. doi: 10.3102/0034654308326349

Sweeney, Kieran. (2006). *Complexity in primary care: Understanding its value*. Abingdon: Radcliffe.

Tan, Kelvin H. K., & Prosser, Michael. (2004). Qualitatively different ways of differentiating student achievement: a phenomenographic study of academics' conceptions of grade descriptors. *Assessment & Evaluation in Higher Education, 29*(3), 267-282. doi: 10.1080/0260293042000188230

Tang, Jinlan, & Harrison, Colin. (2011). Investigating university tutor perceptions of assessment feedback: three types of tutor beliefs. *Assessment & Evaluation in Higher Education, 36*(5), 583-604. doi: 10.1080/02602931003632340

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage.

Taut, S. M., & Alkin, M. C. (2003). Program staff perceptions of barriers to evaluation implementation. *American Journal of Evaluation, 24*(2), 213-226. doi: 10.1016/S1098-2140(03)00028-6

10.1177/109821400302400205

The Review of Evaluation Research. (2013). Aims and scope. Retrieved April 1, 2013, 2013, from http://www.sagepub.com/journals/Journal201854/manuscriptSubmission - tabview=aimsAndScope

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Beverly Hills: Sage.

Thorne, S., Jensen, L., Kearney, M. H., Noblit, G., & Sandelowski, M. (2004). Qualitative metasynthesis: Reflections on methodological orientation and ideological agenda. *Qualitative Health Research, 14*(10), 1342-1365. doi: 10.1177/1049732304269888

Thurmond, V. (2001). The point of triangulation. *Journal of Nursing Scholarship, 33*(3), 254–256.

Tian, Mei, & Lowe, John. (2012). The role of feedback in cross-cultural learning: a case study of Chinese taught postgraduate students in a UK university. *Assessment & Evaluation in Higher Education, 38*(5), 580-598. doi: 10.1080/02602938.2012.670196

Tobin, G. A., & Begley, C. M. . (2004). Methodological rigour within a qualitative framework. *Journal of Advanced Nursing, 4*, 388-396.

Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry, 16*(10), 837-851. doi: 10.1177/1077800410383121

Van den Berg, Ineke, Admiraal, Wilfried, & Pilot, Albert. (2006). Peer assessment in university teaching: evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31*(1), 19-36. doi: 10.1080/02602930500262346

Van Maanen, J. . (1988). *Tales of the field: On writing ethnography*. Chicago: University of Chicago Press.

Walsh, Denis, & Downe, Soo. (2006). Appraising the quality of qualitative research. *Midwifery, 22*(2), 108-119. doi: 10.1016/j.midw.2005.05.004

Wark, L. (1992). Qualitative research journals. *The Qualitative Report, 1*(4).

Whittemore, R. (2005). Combining the evidence in nursing research: Methods and implications. *Nursing Research, 54*, 56–62.

Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research, 11*(4), 522-537. doi: 10.1177/104973201129119299

Whittemore, R., & Knafl, Kathleen. (2005). The integrative review: Updated methodology. *Methodological Issues in Nursing Research, 52*(5), 546–553.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology, 75*(5), 1182-1189. doi: 10.1111/j.1365-2656.2006.01141.x

Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (Eds.). (2010). *Handbook of practical program evaluation* (3rd ed.). San Francisco, CA: Jossey-Bass.

Winchester, Maxwell K., & Winchester, Tiffany M. (2012). If you build it will they come?; Exploring the student perspective of weekly student evaluations of

teaching. *Assessment & Evaluation in Higher Education, 37*(6), 671-682. doi: 10.1080/02602938.2011.563278

Worthen, B.R. , Sanders, J.R. , & Fitzpatrick, J.L. (2004). *Educational evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston: Allyn & Bacon.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.): Thousand Oaks, CA: Sage.

Appendix A: Examples of How Evaluation Standards Can Be Operationalized Using Qualitative Credibility Techniques

Evaluation Standards	Credibility Techniques	Connections
U4 Explicit Values	Multivocality	Including multiple voices, and acknowledge diverse backgrounds and experiences.
	Reflexivity	Establish a clear sense of one's own orientations and assumptions, and values of the evaluations.
	Cultural / contextual information	Clarify and specify the individual and cultural values.
U5 Relevant Information	Triangulation (data/methods)	Collect diverse forms of information through multiple methods.
	Prolonged engagement Persistent observation	Accessibility and relevance of information (the most accessible information may not be the most relevant).
	Weighing evidence	Weighing the relevance, scope and accuracy of evidence.
	Member checking	Relevance of information

	Peer debriefing	should be negotiated and discussed with evaluators, stakeholders, and participants.
U6 Meaningful Process and Products	Explicit design limitation and delimitation	Allow stakeholders and evaluation users to better understand the limitation and boundary of the evaluation to enhance utility of the findings.
	Explicit data collection and analysis processes	Help stakeholders to understand the strength and potential of the evaluations.
P5 Transparency and Disclosure	Explicit design, sampling, data collection and analysis	Explicit methodology leads to greater acceptance, credibility and accuracy.
	Thick description	Provide details about research procedures.
	Member checking	Candid and honest relationship with participants can help promote transparency.
A1 Justified Conclusions and Decisions	Triangulation Multivocality	Include multiple perspectives to justify the findings and conclusions within cultural contexts.

	Reflexivity	Identify and describe important assumptions.
	Prolonged engagement Persistent observation Weighing evidence Expert checking	Help check and enhance quality of information.
	Peer debriefing Member checking	Enhance soundness of logic from information to findings
	Negative case analysis Explore rival explanations	Enhance plausibility of alternative interpretations
A2 Valid Information	Explicit context information	Validity of interpretations in specific context
	Thick description	Provide sufficient details and evidence
	Audit Trail	Documenting the program's theories and components.
	Triangulation	Provide the scope of evidence, to have different data sources and methods grounded in different perspectives.

A3 Reliable information	Inter-coder reliability	Help estimate the consistency of information
	Investigator triangulation	
	Audit trial	Documented procedures for replicability.
A4 Explicit Program and Context Descriptions	Thick description Cultural/contextual info	Describe the program with sufficient detail to provide users with a shared context and perspective.
	Sample demographics described	Provide key information about the participants in compiling program and context descriptions.
A5 Information management	Explicit sampling strategy and data collection	Make explicit decisions about what kind of information is collected, how much is collected.
	Prolonged engagement	
	Persistent observation	
	Explicit coding approach and coding progression	Information management in analysis
	Statistics in qualitative analysis	Managing different types of data to maintain accuracy of analysis
A6 Sound Designs	Appropriate design and	Describe and demonstrate

and Analyses	theoretical framework	understanding and appropriateness of the overall design and theoretical framework that are responsive to the evaluation purposes.
	Explicit analytic procedures	Demonstrate logic and reasoning of the analyzing process.
A7 Explicit Evaluation Reasoning	Exploring rival explanations	Describe the evaluation reasoning from data to findings and interpretations
	Explicit analytic procedures	
	Reflexivity	Describe assumptions of the research team to constantly reflect, discuss, and examine the quality of the reasoning
	Peer debriefing	
	Audit trial	Document the chain of reasoning for further reflection or later investigation.
A8 Communication and reporting	Peer debriefing	Facilitate conversations and communications among stakeholders, participants, evaluators and audience.
	Expert review	
	External auditor	
	Member check	
E1	Audit trail	Document the purpose, design,

Evaluation Documentation	Thick description	implementation and findings of the evaluations.
	Cultural/context information	Document evaluation contexts and data sources.
	Explicit data sources	

Appendix B: The Synthesis Protocol

I. Synthesis Objectives

The research synthesis examines the practice of credibility checks presented in empirical educational program evaluation studies that use qualitative methods (including both qualitative stand-alone studies and qualitative components in mixed method studies) published in six selected leading peer reviewed evaluation journals. A mixed method research synthesis is conducted to identify, describe and evaluate key techniques used by evaluators to enhance credibility of their qualitative work.

II. Synthesis Questions

Two main questions are asked in this synthesis: (a) To what extent are credibility techniques reported in published educational program evaluation work? (b) What are the features that can be observed in the reporting of credibility techniques?

III. Assumptions

1. Researcher bias. Since qualitative research promotes a level of self-revealing and reflexivity, researcher-bias inevitably permeates the qualitative inquiries when conducting program evaluations. However, the author believes that accounting for research bias and dealing with the likelihood of bias undermining the capacity to draw conclusions is something the program evaluator should pursue if attempting to meet AEA standards. Therefore the goal is never to eliminate bias, because bias is an ever-present concept as long as the researcher serve as the instrument in qualitative research. So the presence

of bias should rather to be accounted for when assessing the credibility of evaluation findings.

2. A combination of priori methods and an element of emergent design. The conceptual principle applied in this synthesis is a combination of predetermined research decisions and an element of flexibility throughout the design, implementation, and analysis stage. In other words, although this protocol has specified methodology decisions such as data collection procedures, inclusion and exclusion criteria of the primary studies, the researcher remains open to possible changes in order to be able to response to what is learned in the research process. All changes made over the course of research are noted in the final report.

IV. Search Strategy

1. Journal searching

The searching of evaluation journals is carried out in three steps. First, a manual search is conducted in the databases of ISI Journal Citation Report and Scopus for all evaluation journals. Then the number of journals are narrowed down taking into consideration the aims and scope of the journal, subject areas (focus on education), and publication history (to cover the time span of 2003-2012). Finally, six journals are selected based on their rankings in both the SCImago Journal Rank indicator (SJR) scores published in Scopus website in 2012, and in the Impact Factor 2011 Journal Citation Reports (Thomson Reuters, 2012).

2. Article searching

The searching of qualitative evaluation reports combines keyword searching and hand searching to ensure complete recall within the journal selected. The first round of searching uses keywords in online search engines of each journal's official website. The 21 key words include "qualitative," "evaluation," "credibility," "validity," "trustworthiness," "member check," "triangulation," "peer debriefing," "reflexivity," "participant feedback," "persistent observation," "prolonged engagement," "negative case analysis," "collaborative work," "thick description," "field work," "audit trail," "pattern match," "weight evidence," "rival explanation," and "multivocality." The same set of key words is used for searching of all six journals. Then, the second round of search is hand search to scam through each article in every issue of each of the six journals to make sure an exhaustive search.

V. Inclusion and Exclusion Criteria

This synthesis applies an exhaustive search strategy to identify all relevant articles published in these six journals. Three basic inclusion principles are applied as follows:

(1) *The content criterion*: the study includes only empirical program evaluation studies that address education related topics. By empirical it means that evaluation studies that use primarily collected data are included, and studies consisted of secondary analysis of data collected in another study are excluded.

(2) *The methodological criterion*: program evaluations studies must use qualitative methods. Both qualitative stand-alone studies and mixed method studies are included.

(3) *The temporal criterion*: This parameter is set to the recent decade

between 2003 and 2012. This time period reflected the increased use of qualitative methods in program evaluation. As one of the significant methodological trends emerged since the new century, expanded use of qualitative methods, especially combined with quantitative methods in mixed method design are seen in evaluation studies. A ten-year span is chosen to allow an adequate observation of current methodological practice (especially practice of credibility techniques) and possible observation of changing trends in methodological approaches.

VI. Data Extraction

Identified articles can be located and downloaded from official websites of selected journals, and within the campus of Ohio University, all downloads from these journal websites are free of charge. Also, access to journal articles can be supported by Ohio University library and the Ohio library and information network OhioLink.

A systematic coding protocol, or a codebook is developed as an essential instrument for the synthesis. The codebook is followed by the coder as a set of standards and guidelines for coding the presentation and application of credibility techniques. The codebook is constituted of components to record general characteristics of the evaluation work, primary technique options for design, implementation and presentation of findings in qualitative research and major credibility techniques discussed in literature. The codebook has been reviewed by experts in the evaluation field and tested against a subsample of evaluation articles and refined accordingly, and a pilot study is conducted to further test the reliability of the codebook and to improve its usefulness.

VII. Synthesis Methods

Both quantitative and quantitative methods are used to analyze the data. In the main analysis, descriptive statistical methods are used to identify the number of credibility techniques and the most frequently used techniques, and make comparisons of the use of credibility techniques between qualitative stand-alone and mixed method studies. A thematic analysis is conducted on all descriptions and comments about identified studies. All coder's notes are coded and sorted into categories to reflect different aspect of the evaluation. These categories are refined, compared or combined to produce a final series of patterns of the use of credibility techniques. The thematic analysis is iterative and cyclic that is flexible for emergent themes driven by the data.

Appendix C: A List of Educational Evaluation Journals Selected for the Synthesis

	Journal	SJR (2011)	Impact factor (2011)	Focus
1	Educational evaluation and policy analysis	2.07	1.378	Theoretical, methodological, evaluation of education policy
2	Research evaluation	0.939	0.845	Evaluations of research output and impact; methods for appraising and evaluating research
3	Assessment and evaluation in higher education	0.869	0.841	Assessment and evaluation practices and processes within higher education
4	Educational assessment, evaluation and accountability	0.632	0.694	Functions, theories, values and practices of assessment, evaluation and accountability as they impact schools, higher education and educational systems
5	American journal of evaluation	0.438	2.022	Methods, theory, and practice of evaluation
6	Evaluation review	0.43	1.196	Methodology

Appendix D: Pilot Study of the Codebook: Information Sheet for Coders

Purpose of the codebook

The codebook is used as a template to identify, describe and evaluate credibility techniques of educational program evaluations that use qualitative methods.

Construction of the codebook

Pre-defined checklist

Open-ended comments: note idiosyncratic ways that evaluations are conducted

Coder's tasks

Identify any credibility techniques listed in the codebook

Mark the presence/absence/uncertainty of a technique

Mark the page number when a techniques is present or uncertain

Any comments about the practice of the technique (e.g. claimed using interview but provide no information about participants; jump to conclusion with no evidence provided, etc.)

Please write down approximately how long it takes you to code each study

Tips for coders

1. Scoring guidance (explanation of the categories or indicative questions) is not prescriptive. Please point out in the comment column if you understand it differently.
2. The coding process is iterative. You probably will need to go back and forth in an article or read more than once.
3. The focus is methodology

Post-coding discussion

Suggested topics for a discussion right after the pilot coding:

1. Coding experiences
2. To what extent does the coding categories reflected your own perception about the methodological quality of the studies?
3. Given the guidance (explanations, examples, and indicative questions), how can it be improved to help the coding process?
4. Difficulties experienced?
5. If possible, some of the difference in coding

Pilot Articles

1. Bisset, S., Daniel, M., & Potvin, L. (2009). Exploring the Intervention-- Context Interface: A Case From a School-Based Nutrition Intervention. *American Journal of Evaluation, 30*(4), 554-571.
2. Bradley, P., Oterholt, C., Nordheim, L., & Bjorndal, A. (2005). Medical students' and tutors' experiences of directed and self-directed learning programs in evidence-based medicine: A qualitative evaluation accompanying a randomized controlled trial. *Evaluation Review, 29*(2), 149-177.
3. Esbensen, F. A., Matsuda, K. N., Taylor, T. J., & Peterson, D. (2011). Multimethod strategy for assessing program fidelity: The national evaluation of the revised G.R.E.A.T. program. *Evaluation Review, 35*(1), 14-39.
4. Goldstein, J. (2004). Making Sense of Distributed Leadership: The Case of Peer Assistance and Review. *Educational Evaluation and Policy Analysis, 26*(2), 173-197.
5. Lipnevich, A. A., & Smith, J. K. (2009). “I really need feedback to learn": Students’

perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability, 21*(4), 347-367.

6. Orsmond, P., Merry, S., & Reiling, K. (2005). Biology students' utilization of tutors' formative feedback: A qualitative interview study. *Assessment & Evaluation in Higher Education, 30*(4), 369-386.

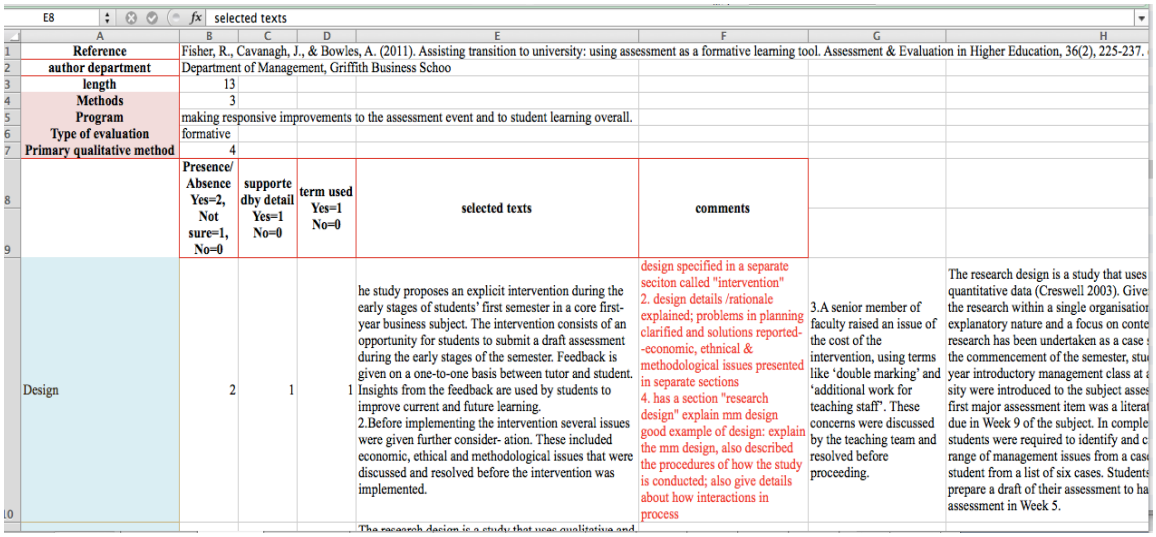
Appendix E: Format and an Example of the Coding Sheet

Format of the coding sheet

Reference	
Author Department	
Length	
Methodology	
Program	
Type of Evaluation	
Primary qualitative method	

Credibility Techniques	Presence/Absence (Y=2, N=0, NS=1)	Supported by Detail	Term used	Relevant Text	Comments
Design					
Sampling					
Data Collection					
Analytic Details					
Thick Description					
Reflexivity					
Limitation/Delimitation					
Triangulation					
Audit Trail					
Expert Checking					
Member Checks					
Peer Debriefing					
Negative Case Analysis					
Prolonged Engagement					
Persistent Observation					
Exploring Rival Explanations					
Pattern Match					
Weighing Evidence					
Comparison & Contrast					
Multiple Perspectives					
Any combined use of credibility techniques					

Snapshot of an Example



E8 selected texts

	A	B	C	D	E	F	G	H
1	**Reference**	Fisher, R., Cavanagh, J., & Bowles, A. (2011). Assisting transition to university: using assessment as a formative learning tool. Assessment & Evaluation in Higher Education, 36(2), 225-237.						
2	**author department**	Department of Management, Griffith Business Schoo						
3	**length**	13						
4	**Methods**	3						
5	**Program**	making responsive improvements to the assessment event and to student learning overall.						
6	**Type of evaluation**	formative						
7	**Primary qualitative method**	4						
8		**Presence/ Absence Yes=2, Not sure=1, No=0**	**supporte dby detail Yes=1 No=0**	**term used Yes=1 No=0**	**selected texts**	**comments**		
9								
10	Design	2	1	1	he study proposes an explicit intervention during the early stages of students' first semester in a core first-year business subject. The intervention consists of an opportunity for students to submit a draft assessment during the early stages of the semester. Feedback is given on a one-to-one basis between tutor and student. Insights from the feedback are used by students to improve current and future learning. 2.Before implementing the intervention several issues were given further consider- ation. These included economic, ethical and methodological issues that were discussed and resolved before the intervention was implemented.	design specified in a separate seciton called "intervention" 2. design details /rationale explained; problems in planning clarified and solutions reported--economic, ethnical & methodological issues presented in separate sections 4. has a section "research design" explain mm design good example of design: explain the mm design, also described the procedures of how the study is conducted; also give details about how interactions in process	3.A senior member of faculty raised an issue of the cost of the intervention, using terms like 'double marking' and 'additional work for teaching staff'. These concerns were discussed by the teaching team and resolved before proceeding.	The research design is a study that uses quantitative data (Creswell 2003). Give the research within a single organisatio explanatory nature and a focus on conte research has been undertaken as a case the commencement of the semester, stu year introductory management class at a sity were introduced to the subject asse first major assessment item was a literat due in Week 9 of the subject. In comple students were required to identify and c range of management issues from a cas student from a list of six cases. Students prepare a draft of their assessment to ha assessment in Week 5.
					The research design is a study that uses qualitative and			

Thesis and Dissertation Services