May 2019

A Validation Of Critical Constructs Of Essential Evaluator Competency And Evaluation Practice: An Application Of Structural Equation Modeling

Jie Zhang
Syracuse University

ABSTRACT

The study aims to examine the interplay of two critical constructs in evaluation: essential evaluator competency and evaluator practice. The research questions in this study, according to Smith (2008), are essentially, what he defined as "fundamental issues in evaluation." These issues fall into one or multiple of the four aspects identified in the fundamental issues in evaluation framework: theory, practice, method, and profession. The intertwined nature of these aspects implies the interactive relationships between the two constructs. The study utilizes the structural equation modeling (SEM) methodology, first to examine construct validity and psychometric properties of the measurement scales, and then explore how the two latent variables of evaluator competencies and evaluator practice interact when evaluators conduct evaluations.

A random sample of 2,000 was drawn from the American Evaluation Association membership directory (n = 7,700), and 459 evaluators from a variety of backgrounds responded. After analyses in the exploratory, confirmatory, and structural phases, the study confirmed five competency dimensions of *evaluative practice, meta-competencies, evaluation knowledge base, project management,* and *professional development.* In addition, analytical results confirmed factor structures of the eight evaluator practice subscales and also revealed four distinct practice patterns, similar to previous research results (Shadish & Epstein, 1987). Despite a small number of significant effects of covariates such as years of experience and evaluation background, multiple indicators multiple causes (MIMIC) model results concluded that the measurement models were mostly invariant across various population groups. Lastly, the structural phase analyses uncovered that the relationship between evaluator self-assessed competencies and evaluator practice patterns are interactive. The findings from the SEM model with self-assessed

competencies as predictors indicated that evaluators with higher self-assessed *evaluative practice* competencies tend to engage in the *academic* and *method-driven* practice patterns; Evaluators with higher self-assessed *meta-competencies* tend to engage in the *use-driven* practice pattern more frequently. On the other hand, when evaluator practice patterns served as predictors, the results showed that evaluators engaging in the *academic* pattern more often tended to rate higher of their *evaluative practice, meta,* and *evaluation knowledge base* competencies; and evaluators engaging in the *use-driven* practice pattern tended to rate higher of their competencies in all areas except *evaluation knowledge base*.

The study extends previous research by confirming the factor structures of two critical constructs in the evaluation field and providing empirical support for future studies. The findings contribute to a better understanding of several fundamental issues in evaluation, evaluation professionalization and the general knowledge base of the field.

Keywords: *professional competency, essential competencies for program evaluators, evaluation practice, fundamental issues in evaluation, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), structural equation modeling (SEM), multiple indicators multiple causes (MIMIC)*

A Validation of Critical Constructs of Essential Evaluator Competency and Evaluation Practice: An Application of Structural Equation Modeling

by

Jie Zhang

M.S., Syracuse University, 2013
M.S., Syracuse University, 2014

Dissertation

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Instructional Design, Development & Evaluation

Syracuse University
May 2019

Acknowledgments

This dissertation symbolizes a critical milestone in my educational pursuit. Along the journey reaching this milestone, many have inspired, mentored, guided, supported, encouraged, and believed in me, and more importantly, have never given up on me. I would like to express my sincere appreciation to all of you.

First and foremost, I would like to express my deepest gratitude to my committee members. Dr. Nick L. Smith, I am truly honored to have you as my advisor, mentor, friend and role model. The discussions we had on evaluation theory and practice have greatly shaped my understanding of the evaluation field and my dissertation study. Those thought-provoking conversations with you on research and life have challenged and enriched my reasoning competencies. Your wisdom, philosophy, optimism, sense of humor, patience, and intellectual integrity have influenced me profoundly.

Dr. Qiu Wang, thank you for tirelessly reviewing my Method and Results sections, and guiding me through the wonderful world of factor analysis and structural equation modeling. Your advice has improved the rigor of my study.

Dr. Katherine McDonald, I am indebted to your willingness to serve on my committee and bringing your expertise and unique perspectives on research to my study.

IDD&E community has played a tremendous role to help me grow academically and professionally. Drs. Tiffany Koszalka, Phil Doughty, Michael Spector, Jing Lei, Chuck Spuches, Jerry Klein, Jerry Edmonds, Rob Pusch, and Linda Tucker, thank you for introducing me to the field of instructional design and helping me develop essential professional and life skills. Linda, without your smile and warm help, I would not have been able to finish my study. Special appreciation goes to Rob and Jerry for giving me opportunities to acquire invaluable practical knowledge and experience with survey research at Project Advance. I was able to complete all my doctoral coursework and a Master's degree in Applied Statistics. I am forever grateful!

I would also like to extend my gratitude to the late Dr. William Shadish, for sharing his insight with me at the early stage of this study.

My friends, Hongyuan Dong, Ruzanna Topchyan, Deniz Eseryel, Yeliz Eseryel, Sunghye Lee, Dr. Susan Branson, your friendship has made my time in Syracuse the happiest of my life.

To my colleagues and friends at Johns Hopkins University, I thank you for your constant encouragement and support. Dr. Sylvia Long-Tolbert, I am grateful for your friendship, guidance, encouragement, and particularly your enthusiasm to discuss formative and reflective measurement models with me. Dr. Lindsay Thompson, you always light up my face with your encouraging words and thank you for believing in me. Darlene Dixon, you are a relentless cheerleader for even my smallest progress. I particularly want to thank my good friends, Veena Radhakrishnan and Patrick Dempsey. I cannot think of anyone better than the two of you to go through this journey with!

To my family, I could not have done this without you! Mom and Dad, thank you for your unconditional and unwavering love, and all the sacrifices you have made to support my pursuit. My sister, Lihua, I am forever indebted for your willingness to care for our parents while I am thousands of miles away. My niece, Runxi, I aspire to be a role model for you, and am so proud of your accomplishments thus far! Zack, I appreciate the joy, emotional support, and unconditional love that you have provided me throughout the years. You make America home for me!

I dedicate this dissertation to my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Evaluation scholars study a wide range of topics and issues, from evaluator roles, communication strategies with stakeholders, methodological choices, to theoretical justification of different evaluation designs, to advance evaluation as a field of professional practice. Smith (2008) suggests that these issues and problems should be defined as "fundamental issues in evaluation"—the "underlying concerns, problems, or choices that continually resurface" (p. 2). Four separate yet closely connected aspects of theory, method, practice, and profession, are used to characterize these fundamental issues. However, the interconnected nature of theory, method, practice, and profession make it challenging to specify any particular fundamental issue under one particular aspect. Any given issue may have characteristics across multiple aspects. The fundamental issues in evaluation framework facilitates the identification of common reoccurring problems, recognizes the interconnected relationship patterns in evaluations, and provides a systematic and holistic, rather than an isolated view of research on evaluation.

In this study, three such fundamental issues are under investigation:

- What essential professional competencies should evaluators possess to conduct evaluations efficiently and effectively?
- Are there common practice patterns in evaluation practice?
- What is the relationship between evaluators' professional competencies and their practice patterns?

These fundamental issues of interest have characteristics crossing all four aspects of theory, method, practice, and profession. The following sections first identify the gaps in the current research on evaluation and explore the importance of the research agenda set in the study.

Then, research purpose and detailed research questions are presented. Lastly, the significance and contributions of the study are discussed at the end of the chapter.

Problem Statement

Although much progress has been made on various fronts in research on evaluation, many gaps remain unbridged, such as: ongoing calls for more empirical research on evaluation to build a more robust evidence base (Schwandt, 1997; Shadish, Cook, & Leviton, 1991; Smith, 1993; Mark, 2001; Worthen, 2001); a lack of a commonly accepted set of evaluator competencies despite many years' discussion and research efforts (King, Stevahn, Ghere, & Minnema, 2001; Smith, 1999; Worthen, 1999); a limited amount of research on how evaluation theory guides practice (Christie, 2003; Shadish, 1998; Williams, 1989); and particularly scarce studies on how evaluators conduct evaluations and their practice patterns (Shadish & Epstein, 1987; Schwandt, 1997, 2002) in relation to their professional competencies.

The study takes a closer look at the essential evaluator competencies, evaluator practice, and the relationship between the two critical constructs. The following sections discuss the three gaps in detail by providing definitions and mapping out detailed inquiries.

Professionalization Requires Evaluator Competencies

A profession is defined in the Oxford Dictionary as "an occupation in which a professed knowledge of some subject, field, or science is applied." A similar definition by Carr-Saunders and Wilson (1933), states "an occupation based upon specialized intellectual study and training, the purpose of which is to apply skilled service or advice to others for a definite fee or salary (p. 5)". According to Cheetham and Chivers (2005), both definitions for profession fail to draw a clear boundary among various occupations. After comparing various approaches of defining

profession, Cheetham and Chivers (2005) provided their definition of profession as "an occupation based upon specialized study, training or experience, the purpose of which is to apply skilled service or advice to others, or to provide technical, managerial or administrative services to, or within, organizations in return for a fee or salary" (p. 11). Eraut (1994), however, viewing a profession as an ideology, agreed with Johnson (1972) to define professionalization as "the process by which occupations seek to gain status and privilege in accord with that ideology." Altschuld (1999) provided a more comprehensive view of the concept of profession:

> A profession is a vocation requiring specialized training in a field of learning, art or science. The term profession also refers to the body of persons engaged in this calling or vocation. Professions are characterized by specialized training (skills and competencies), engagement in a field as the major source of livelihood, skills beyond the level of novice or beginner (or even amateur), and commitment to the profession, for example, by involvement in professional associations. Being in a profession usually entails adherence to a code of ethics (psychology, evaluation, medicine) and performance in accord with a set of guidelines for practice (p. 483).

Since the 1960s, evaluation researchers have debated widely about whether evaluation had achieved the status of a profession. Viewing evaluation as a profession, Anderson and Ball (1978) surveyed sixty-four evaluation experts in an attempt to establish a set of essential competencies for professional training purposes focusing on areas of content knowledge and skills. While the survey provided useful information, Sechrest (1980) deemed it premature to claim evaluation as a profession, but agreed with Morell and Flaherty (1978) that evaluation had demonstrated some characters and started to emerge as a profession with an increasing number of unique training programs and the formation of a professional association, the Evaluation Research Society (ERS).

Light (1995) did not declare evaluation as a profession but recognized that the 1986 merger of the Evaluation Network (ENET) and the ERS into the American Evaluation

Association (AEA) signified a significant step towards professionalization. On the contrary, House (1994) considered evaluation as a "specialized profession" since 1965 with "its own organizations, journals, and studies conducted by those who call themselves evaluators" (p. 239).

Worthen (1994) summed up the professionalization debates by providing a checklist with nine criteria and determined that six out of the nine criteria have been met. Additionally, Worthen contended that the judgment of the professionalization of evaluation, in many ways, was still subjective, depending on how rigorously the nine criteria were executed. Michael Scriven (as cited in Worthen, 1994, p. 13) suggested a compromising view of evaluation as a "hybrid of profession and discipline."

Despite the different views, there seems to be an implicit agreement on evaluation as a profession among evaluators. It is evident in the evaluation literature that the discussion has switched from debating on whether evaluation is a profession to examine profession-specific issues, such as evaluator competencies, certification/licensure of evaluators, and development of evaluation training programs (Altschuld, 1999; Becker & Kirkhart, 1981; Jones & Worthen, 1999; Love, 1994; Smith, 1999; Worthen, 1999). Stevahn, King, Ghere, and Minnema (2005) noted the importance of having a set of commonly accepted evaluator competencies, as not only the defining characteristics for the evaluation profession but also influencing factor for evaluation training and professional development. These researchers summed up five consequences of the lack of competencies: the obstruction of certifying/licensing evaluators; the difficulty of selecting /hiring qualified evaluators for the job; the missing guidelines for future evaluators; the lack of systematic evaluation curricula and professional development; and the increasing gap between evaluation theory and practice. Stevahn et al. (2005) also laid out four

benefits of establishing evaluator competencies: improving training; enhancing reflective evaluation practice; advancing research on evaluation; and furthering professionalization.

However, Stevahn and colleagues (2005) were not the first to recognize the importance and benefits of evaluator competencies. Efforts in searching for evaluator competencies can be traced back to Worthen (1975), who synthesized the results of three previous taskforces and derived 25 general tasks crucial for educational researchers and evaluators. However, these tasks were generic, and the sub-skills for these tasks were not unique to evaluators. Over the years, various attempts (Davis, 1986; Dewey, Montrosse, Schroter, Sullins, & Mattox, 2008; King, et al., 2001; Kirkhart, 1981; Mertens, 1994; Sanders, 1986; Scriven, 1996; Stevahn, et al., 2005; Stufflebeam & Wingate, 2005) have been made to establish a set of commonly accepted evaluator competencies conceptually and empirically. Results of most of the attempts were either highly conceptual, unsystematic or narrowly focused. The recent works on the taxonomy of essential competencies for program evaluators (ECPE) (King, et al., 2001; Stevahn, et al., 2005; Ghere, King, Stevahn, & Minnema, 2006) put forth a set of 61 specific and behaviorally-based competencies, which were empirically derived using Multi-Attribute Consensus Reaching (MACR) methodology, and tested empirically in professional development training seminars. Although the advantages of establishing the ECPE taxonomy are apparent, there is a lack of rigorous and systematic research to validate the set of competencies. For example, the small sample size of existing research on the ECPE taxonomy might affect the accuracy of the findings. Given that only face and content validity has been achieved, much work has to be done in establishing construct validity. There is a need for large-scale validation studies on the ECPE in the evaluation field.

This study aims to answer such a call to build upon the research effort to establish the content validity of the ECPE competencies, confirm such construct validity within a much larger sample, and explore the interactions of evaluator competencies with evaluators' practices.

Towards a Better Understanding of Evaluation Practice

As the previous section established how important the establishment of a set of well-validated essential evaluator competencies is to the profession, this section directs attention to evaluation practice and seeks to answer two questions of why evaluator practices should be studied, and how essential evaluator competencies relate to evaluator practices.

Evaluator practice is the process of how evaluators conduct evaluations using specific knowledge and skills, such as knowledge of evaluation theory, knowledge of various evaluation designs, motivation to satisfy clients, the pursuit of professional standards and ethical conduct, skills of managing evaluation personnel, skills of communicating with stakeholders, and skills of effective reporting. These knowledge, skills, and dispositions that evaluators use in their daily practice are essential competencies (Ghere, et al., 2006). Worthen (1999) referred to these competencies as the "sine qua non" of program evaluator performance. Scriven (1996) also connected evaluator competencies with practices at the professional level. He argued that many professionals engage in some evaluation activities, but not all qualify as professional evaluators. Only those who conduct "technically challenging" evaluations "with reasonable competence" (p. 154) can claim the title. This contention reflects Scriven's view of the close relationship between evaluator competencies and their practices.

Though evaluation researchers repeatedly emphasized its importance, evaluation practice has yet received as much attention as other areas of program evaluation such as evaluation theory

and methods (Shadish et al., 1991; Smith & Brandon, 2008). Few empirical studies have been conducted, and most studies focus narrowly on specific areas of practice: evaluation use (Shulha & Cousins, 1997); data collection (Benkofske, 1996); and decision-making (Kundin, 2010; Tourmen, 2009). Several literature reviews on practice focus mainly on the methodology utilized in practice (Lynch, 1988) and needs assessment (Witkin, 1994).

Despite that William (1989) and Christie (2003) examined evaluator practices from the perspectives of how evaluation theories and theorists' practice were mapped to evaluation practitioners' practice, the study by Shadish and Epstein (1987) remains as the only one that investigated evaluation practice comprehensively. The uniqueness of the study lies in that researchers constructed a comprehensive instrument to measure evaluation practice as a latent construct, and uncovered four distinct practice patterns as a result of advanced multivariate analyses. As a result, this quantitative approach made it possible to examine the relationships among evaluation practice with other constructs in the field. Shadish and Epstein created a set of 74 questions to measure evaluation practice in eight aspects. The consequence of the approach was a loss of salient details compared with a qualitative approach. However, this weakness is inherent to any quantitative research, and the establishment of a strong content validity also mitigates such weakness. Because the Shadish and Epstein study has just partially established construct validity of evaluation practice, the ensuing examination of relationships with covariates such as training, work settings, and theoretical influences seemed premature. Furthermore, changing evaluator demographics may result in different practice patterns from those discovered in the original 1987 study.

Building upon research by Shadish and Epstein (1987), the present study intends to advance the line of research on evaluation practice in three areas: 1) validation of the factor

structure of evaluation practice scale with the current evaluator population; 2) examination of how covariates (such as work setting, educational background, and years of experience) influence the factor structure; and 3) exploration of the relationship of evaluator competencies and evaluation practice.

Study Purposes & Research Questions

The study has three main goals: to establish and confirm the construct validity of the scale of evaluator competencies adapted from the ECPE framework (King et al., 2001), to confirm the construct validity of the scale of evaluator practice adapted from Shadish and Epstein (1987), and examine the relationships between the two constructs using structural equation modeling. The research questions are addressed in three phases.

Exploratory phase. In the exploratory phase, exploratory factor analyses (EFA) were performed to explore and confirm the factor structures of two rating scales of ECPE and evaluator practice. EFA procedure is mostly used to uncover the factor structure of a construct. However, in the current case, EFA was used in a confirmatory capacity (Klein, 2016) because the factor structures of both scales have been previously established conceptually or empirically. Stevahn and colleagues (2005) conducted a preliminary content validity test and proposed a conceptual 6-factor structure. The evaluator practice scale was analyzed in the Shadish and Epstein (1987) study and yielded 22 first-order factors and 4 second-order factors. The EFA conducted in this phase imposed the factor structures established in the previous studies to verify whether these factor structures still hold. The research questions in this phase include,

- **R1.** What is the factor structure of the ECPE scale? Precisely, does the factor structure of the ECPE scale conform to the 6-factor structure conceptualized by Stevahn et al. (2005)?

This research question examines whether evaluator competency is a multidimensional construct with six dimensions as previous researchers contended.

- **R2.** What is the factor structure of the evaluation practice scale? Specifically, can the same factor structures of 22 first-order factors be derived from 8 sub-domains of evaluation practice? This research question examines whether the evaluator practice scale can reproduce the same factor structure as in Shadish and Epstein (1987) study.
- **R3.** What is the higher-order factor structure of the evaluation practice scale? Specifically, can the same four-factor structure be derived from the first-order factors in R2? This research question builds on the previous question and continues to confirm whether evaluator practice can be summarized by four practice patterns as presented in Shadish and Epstein (1987).

Confirmatory phase. Confirmatory Factor Analyses (CFA) are conducted to confirm the factor structures resulting from the previous phase. Even though the exploratory phase also had a confirmatory purpose, there is a significant difference in statistical modeling procedures applied in these two phases. To be specific, the CFA models tested in the confirmatory stage provide model fitting statistics since CFA is a more restrictive analytical technique. In EFA, items or indicators are allowed to load freely on all latent factors. Various rotation methods, orthogonal or oblique, can be used to produce a clear pattern structure with items/indicators having salient loadings on one factor. However, in CFA models each item or indicator has been pre-determined to load on only one factor. The goodness-of-fit statistics and model modification indices are provided to facilitate CFA model improvements. Furthermore, the analyses conducted in CFA are also known as testing measurement models—a crucial precursor for testing structural models. Measurement invariance was also examined in this phase using multiple indicators and multiple

causes (MIMIC) modeling (Jöeskog & Goldberger, 1975). Research questions in the confirmatory phase include,

- **R4.** Does the factor structure yielded in R1 achieve reasonably good model fit? The research question aims to confirm the factor structure of the evaluator competencies established in the exploratory phase.
- **R5.** Does the factor structure yielded in R2 achieve reasonably good model fit? The research question aims to confirm the first-order factor structure of evaluator practice established in the exploratory phase.
- **R6.** Does the factor structure yielded in R3 achieve reasonably good model fit? The research question intends to confirm whether the four higher-order factors of evaluator practice can be achieved in the exploratory phase.
- **R7.** Does the factor structure established in R4 vary by the levels of covariates? (The eight covariates are years of experience, professional identity, primary affiliation, highest degree achieved, the field of study, job settings, evaluation background, and gender).
- **R8.** Do the above eight covariates have statistically significant effects on the measurement model established in R3?

Structural Phase. Once the CFA models were tested and confirmed in the previous phrase, the study proceeded to investigate the relationship between evaluator self-assessed competencies and evaluation practice patterns. The relationship in this study refers to the statistically predictive effects of the two constructs on each other. To infer true casual relationships, Shadish, Cook, and Campbell (2015) contend that three characteristics or criteria have to be met: 1) time precedence. The cause variable preceded the effect variable. 2) correlation. Both variables have to be correlated, and 3) no plausible alternative explanations.

Other confounding variables have to be ruled out. In the current study, the causal direction of the relationship cannot be established with the two constructs, as only one of the three criteria, correlation, is present. However, the findings of the study can be informative for designing future research to investigate the causal relationship between the two constructs. The research question to be answered in this phase is,

- **R9**. How do evaluator self-assessed competencies and evaluation practice patterns relate to each other? Specifically, do evaluator self-assessed competencies have significant effects on evaluation practice patterns, or evaluators' practice patterns have significant effects on their self-assessed competencies?

Significance of the Study

The importance of empirical knowledge has been noted by many evaluation researchers (Mark, 2001; Schwandt, 1997; Scriven, 1995; Shadish et al., 1991; Smith, 1993; Worthen, 1999). Smith (1993) argued that empirical knowledge on evaluation practice had direct impacts on developing more relevant evaluation theories, facilitating better decisions on choosing alternative theories/models, and consequently guiding more competent practices. The study intends to contribute to the empirical knowledge base in several ways.

Firstly, the current study has extended previous research on evaluator competencies and develops a deeper understanding of evaluator competencies as a multidimensional construct. As evaluation moves further into the professionalization process, it becomes imperative to establish a set of rigorously-tested and widely-accepted evaluator competencies. Not only can current evaluators benefit from these competencies by critically reflecting on their knowledge and skills, but also new evaluators can be better guided and prepared. Having well-established

competencies can also propel evaluation a step closer to the certifying and licensing process towards becoming a more mature and better-regulated profession.

Secondly, the present study has examined current evaluator practices comprehensively using the instrument developed by Shadish and Epstein (1987). It is also the goal of the study to further validate the scale, and confirm the patterns discovered in the previous study. The study not only resulted in a more reliable measurement scale but also provided a comparison of evaluation practice patterns of the 1980s and now.

Lastly, the study has explored the relationships of the two critical constructs in evaluation systematically and dynamically. Smith (2008) points out that the fundamental issues in evaluation are often connected by their underlying characteristics in theory, method, practice, and profession. Previous research on these fundamental issues was often restricted to one or two aspects. The current study answered the research questions from multiple perspectives, and the results provided better evidence to support a better understanding of these fundamental issues in evaluation.

Challenges/ Limitations of the Study

Establishing sound measures is crucial for any research. The main challenges of the study were the lack of any psychometric properties of the two scales. As the ECPE competencies were only subjected to a content validity test, no psychometric property information such as reliability has been established. Furthermore, the conceptually hypothesized six dimensions have not been validated through rigorous methodologies in large samples. Similarly, the evaluator practice scale developed by Shadish and Epstein (1987) faced the limitation of lacking necessary

psychometric information. With the rapid progress in the field of evaluation since the study was conducted, there might be concerns about content currency and relevancy of the items.

However, the challenges and limitations may also confirm how crucial and timely the study can be to the field of evaluation. The present study examined the psychometric properties, yielded reliabilities, and established construct validity for measurement instruments for both constructs. More importantly, a solid foundation has been built to further this line of research involving the two critical constructs.

Chapter Summary

Fundamental issues in evaluation are often connected through the underlying themes and characteristics of theory, method, practice, and profession. Three such related issues of evaluators' competencies and their practices are under scrutiny. Research on these issues was scarce, and often conceptual. Aiming to contribute to the empirical knowledge base, the current study has the main purpose of exploring the interactive relationships among evaluators' competencies and their evaluation practices. To achieve the goal, it is an integral part of the process to develop psychometrically sound measurement scales. Once the sound measurement models were established for the two measurement scales, the study moved on to examine the structural part of the investigation.

Chapter II proceeds to a comprehensive review of related literature, concentrating on the conceptual frameworks and existing empirical studies. Since scale development is the focus of the current study, the literature on various kinds of validity (face, content, criterion, and construct) and general procedures of instrument development are also discussed.

CHAPTER II: REVIEW OF LITERATURE

The purpose of the literature review is to summarize the conceptual development and findings in empirical research. This chapter first explored briefly about the concepts of competence and competency and competency frameworks. Then, it continues with an extensive content review of research on the development of evaluator competency dimensions and systematic analyses of competency dimensions and specific competencies from existing research and professional evaluation organizations worldwide. Next, the chapter discusses the nature of professional practice and summarizes the findings from empirical research on evaluation practice. The chapter concludes that the ECPE competency framework and the evaluation practice scale by Shadish and Epstein (1981) are by far the most comprehensive available measurement instruments in this area of research.

Concepts of Competence and Competency

Competence and competency are widely used terms in education, training, and human resource management. As competence and competency are closely related, two trends of use are often observed: competence or competency has been used without clear definitions, and competence and competency tend to be used interchangeably. Researchers often make implicit assumptions about their definitions and ignore the connections and differences between the two concepts. This misuse often results in confusion at different levels. For example, researchers often cite McClelland's seminal work on "testing competence rather than 'intelligence'" (1973) for the definition of competence. However, Barrett and Depinet (1991) argued that McClelland did not provide a clear definition in this seminal work. They pointed out "a fundamental problem with McClelland's (1973) research was his failure to define his concept of competency. To obtain a definition of this term, we had to rely on subsequent papers he and his associates had

written" (p. 1019). The lack of definitions for these two key terms stir up researchers' curiosity of why it is so challenging that McClelland did not properly define competency?

The difficulty of acquiring a precise definition has been well recognized and discussed. Eraut (1994) argued that the scope of competence (general and specific) carries different meanings when used in different professions and contexts. While general and specific competences can be inferred from each other more consistently in professions with similar tasks, generic competence could be less useful or even detrimental in professions with diverse sets of tasks, and specific competence is much more desired. He also contended that competence, as a stage in the professional development of expertise, has the dual meanings of "getting the job done" or "adequate but less than excellent" (p. 166). The first meaning implied that competence was judged on a binary level as the state of being competent (Richey et al., 2001); and the later, on the other hand, was evaluated on a continuum. Eraut advocated for the view of competence on a continuum because the arbitrary judgment on the binary scale of competent or not does not explain what competence the person has, and being competent varies drastically in different professions and contexts.

Other researchers tend to agree with the view of competence on a continuum. Dreyfus and Dreyfus (1986), for instance, developed a five-stage framework of competence development from novice to expert. Consistent with the view, Cheetham and Chivers (2005) argued for the dynamic nature and defined competence as "effective overall performance within an occupation, which may range from the basic level of proficiency through to the highest levels of excellence" (p. 54). This definition is compatible with most American scholars to define competence as "a person's overall capacity" (Eraut, 1994, p. 179).

Richey and colleagues (2001) observed the diverse views on the nature of competency, and provided a definition of competency in the International Board of Standards for Training, Performance, and Instruction (IBSTPI) as "a knowledge, skill, or attitude that enables one to effectively perform the activities of a given occupation or function to the standards expected in employment" (p. 31). In human resource management, Klemp (1980) identified competency as "an underlying characteristic of a person which results in effective and superior performance on the job" (p. 21). A close definition given by Boyatzis (1982), building upon McClelland's works on competence, regarded competency as "an underlying characteristic of a person in that it may be a motive, trait, skill, aspect of one's self-image or social role, or a body of knowledge which he or she uses" (p. 21).

Parry (1998), on the other hand, argued that competencies should not be confused with personality trait and characteristics. Even though he admitted that a person's style/value influences how one uses his competencies, he advocated viewing competencies and style/values as two distinct concepts. Parry defined competency as "a cluster of related knowledge, skills, and attitudes that affects a major part of one's job (a role or responsibility), that correlates with performance on the job, that can be measured against well-accepted standards, and that can be improved via training and development."

Richey and colleagues (2001) took a similar approach as Parry, also viewed competency as a concept that is "innately behavioral and positivistic in nature" (p. 31). Lucia and Lepsinger (1999) provided two reasons why competencies should be behavioral. First of all, defining in behavioral terms makes it easy for identification and demonstration of specific competencies; Secondly, behaviors can be modified and trained easier than personality traits.

However, Parry's definition contradicted his argument because attitude is an essential aspect of personal trait and characteristics. Also, personal traits can be changed and assessed, though with some difficulty. Spencer and Spencer (1993) and Spencer, McClelland, and Spencer (1994) incorporated personality trait and characteristics into the definition of competency to a great extent. To be more specific, Spencer and Spencer's definition stated, "a competency is an underlying characteristic of an individual that is causally related to criterion-referenced effective and superior performance in a job or situation" (p. 9). This underlying competency characteristic included five components of motives, trait, self-concepts (such as attitudes and values), content knowledge, and skills. At the core are the first three more hidden components of motives, trait, and self-concepts. The last two components, knowledge and skills, were considered the outside layers and visible aspects. Spencer and Spencer (1993) acknowledged the difficulty of developing and assessing the three hidden and core competencies in training but also suggested other alternative methods to foster change, such as psychotherapy or positive developmental experiences.

How does competency relate to competence then? Richey et al. (2001) contended that competency merely was how competence was demonstrated and represented in practice. Eraut (1994) made a similar observation and contended that competency could relate to competence in two ways. Firstly, competency can be considered as a performance manifestation of a specific capability or competence in a specific context. Secondly, competency can be viewed as knowledge or skill needed for the specific capability or competence. Russ-Eft (1995) provided an analogy to describe this dual role of competency:

> Competencies may be thought of as the core elements in a periodic table for human behavior. The "atoms" in such a model are behavioral indicators. These behavioral

> indicators can be grouped into competencies, or "elements." Finally, several competencies can be combined to form other competencies, or "molecules" (p. 329).

The analogy, consistent with Eraut (1994, p. 181), describes flexibility of competency in functioning individually as well as collectively at the micro and macro levels to mean specific competencies and general/generic competencies.

In other words, competence, as an abstract construct, cannot be observed and measured directly. Competencies, however, if stated in performance terms, can be directly assessed and often used as indicators for competence. Gonzi, Hager, and Athanasou (1993) described the relationship between competence and competency regarding performance:

> The competence of professionals derives from their possessing a set of relevant attributes such as knowledge, skills and attitudes. These attributes jointly underlie competence and are often referred to as competencies. So a competency is a combination of attributes underlying some aspect of professional performance…[But] attributes of individuals do not in themselves constitute competence. Nor is competence the mere performance of a series of tasks. Rather, the notion of competence integrates attributes with performance." (p. 5).

Gonzi and colleagues further pointed out that competence is not merely an overarching term summarizing competencies. Instead, it is an integration of acquiring competencies as well as the ability to use them in performing various job-related tasks.

Approaches to Modeling Professional Competence

Because the purpose of this study is to examine how evaluator competencies were modeled, it is imperative to outline approaches and methods for modeling competencies. In deriving a typology of competence, Le Deist and Winterton (2005) identified three prominent competence-modeling approaches: the behavioral competency approach in the U.S., the functional approach in the UK, and a multi-dimensional and holistic approach in European countries. Le Deist and Winterton observed, even though the behavioral competency approach to

competence is still much relevant, “a broader conception of competence, which also emphasizes job-related functional skills and underpinning knowledge, is clearly gaining ground” (p. 33). Subsequently, Le Deist and Winterton presented their multi-dimensional competence framework (see figure 2.1). This holistic typology, they further contended, provides an integrated view of competencies by combining knowledge, skills, and social competences.

	Occupational	**Personal**
Conceptual	Cognitive competence	Meta Competence
Operational	Functional competence	Social competence

Figure 2.1 Typology of competence. Adapted from Le Deist and Winterton (2005)

To build a comprehensive professional competence model, Cheetham and Chivers (1996, 2005) compared and analyzed several competency approaches, including the technical-rational approach, Schön’s reflective practitioner approach (1983), UK’s functional competence approach, and the personal or behavioral competence approach. Two additional dimensions, as they pointed out, would broaden the perspectives: meta-competence and emotional intelligence. While each approach has its unique strengths in framing professional competence, Cheetham and Chivers (1996) argued for a need for a more holistic professional competence model. After the initial conceptualization of such a model, they empirically tested the model with extensive

interviews and survey data. The revised model is an integrated, holistic competency model that incorporates all the competence approaches. In the model depicted in Figure 2.2, meta-competencies encompass four dimensions of knowledge/cognitive competence, functional competence, personal/behavioral competence, values/ethical competence. The outcomes from the four competency dimensions then serve as definite evidence for professional competence. The model also incorporates reflection into the competency model process.

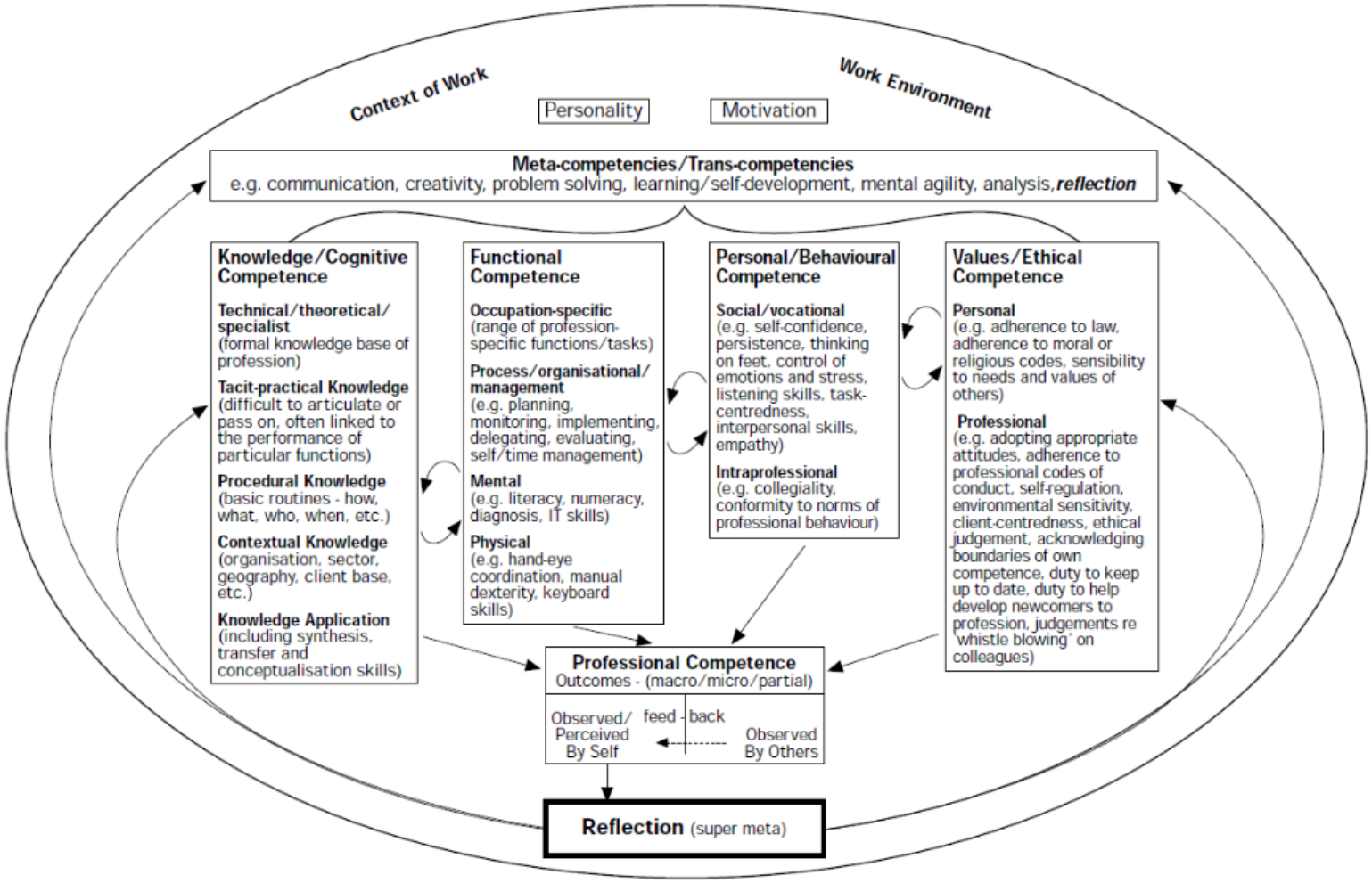


Figure 2.2 The revised holistic model of professional competence. Adapted from *Professions, Competence and Informal Learning* (p. 112), by Cheetham, G. and Chivers, G. E., 2005, Cheltenham, UK: Edward Elgar. Copyright 2005 by the Edward Elgar Publishing.

Wilcox and King (2014) reported three major approaches to professional competency modeling in the U.S., differential-psychology approach, the educational and behavioral psychology approach, and the management-sciences approach. While the first two approaches emphasize individual and developmental abilities, the third approach focuses on job analysis.

The brief review of professional competency modeling in this section is essential because it provides not only a theoretical context to define competence and competencies, but also an exemplary competency framework and its dimensional components (see Figure 2.2). Among all the approaches reviewed in this section, the framework by Cheetham and Chivers (2005) had the most impacts in facilitating the analyses and interpretation of factor dimensions in the ensuing Results and Discussion Chapters. The next logical step is turning towards the evolution of professional competencies in the field of evaluation and the comparison of various competency frameworks.

Evaluator Competencies

Evaluator competencies have long been under scrutiny as the result of discussions on the professionalization of the field (Anderson & Ball, 1978; Hauer & Slee, 1989; King, Stevahn, Ghere, & Minnema, 2001; Kirkhart, 1981; Love, 1994; Smith, 1999; Stevahn, King, Ghere, & Minnema, 2005; Worthen, 1994, 2001). Worthen (1999) argued, “evaluator competencies - skills and knowledge that enable an individual to conduct a quality evaluation study - represent the sine qua non in performance as an evaluator” (p. 546). Using behavioral terms, Stevahn et al. (2005) defined evaluator competencies as “the knowledge, skills, and dispositions program evaluators [need] to be effective as professionals” (p. 48). Singular competency is used to refer

to evaluator competency as a construct; while plural form, competencies, are used to refer to specific evaluator competencies.

The majority of the investigations on evaluator competencies have been conceptual. A few existing empirical studies were mostly descriptive (Anderson & Ball, 1978; Dewey, Montrosse, Schroter, Sullins, & Mattox, 2008; Stufflebeam & Wingate, 2005). For the purpose of examining construct validity, the literature review of this study has been focused on categorization, organizational schemes, and dimensionalities of evaluator competency frameworks, and the corresponding specific evaluator competencies.

Research on Evaluator Competencies

Despite the awareness of the crucial role of competencies to the profession of evaluation and abundant literature addressing the importance, there has been little conceptual guidance on how to systematically derive a set of sound competencies. As a result, little empirical effort has been made in this endeavor. A research review of major evaluation work and journals was conducted and 12 principal evaluation and educational journals were reviewed: American Journal of Evaluation, Canadian Journal of Program Evaluation, Contemporary Education, Educational Evaluation & Policy Analysis, Evaluation and Program Planning, Educational Researcher, Evaluation Review, Evaluation & Health Professionals, Evaluation in Education, Journal of Personnel Evaluation in Education, New Directions for Evaluation, and Studies in Learning, Evaluation, Innovation and Development. The review resulted in 41 peer-reviewed articles and one book.

To discover how evaluation researchers understand, define, develop, and categorize evaluator competencies, two selection criteria were applied for further analysis: 1) those sources

that discussed competency related issues conceptually or empirically, and 2) those that included specific competencies or categories/areas of competencies. Subsequently, 28 articles were eliminated because they did fit the two criteria. The final selected resources (n = 14) were identified for comparison in five areas of methods used to derive competencies, dimensionality/categories identified, the number of specific competencies included, sample size if derived empirically, the types of validity assessed, and the definition of competency if available. The detailed comparison is presented chronologically in Table 2.1.

Table 2.1 Summary of Research Development on Evaluator Competencies

Author (year)	Method	Categorization/Dimensionality	# of Competencies	Sample Size	Validity	Definition
Brzezinski & Ahn (1973)	Empirical Survey study	Eight dimensions/sub-scales: 1) knowledge of innovation in evaluation (46 items); 2) public relations (8 items); 3) data processing (11 items); 4) educational measurement (34 items); 5) evaluation administration (50 items); 6) relating evaluation to relevant disciplines (12 items); 7) communications (22 items); 8) research design analysis (51 items).	234 items	77 responded out of a stratified random sample of 252 (10% of 2500 of the sampling frame, since only the pilot test was conducted.)	Content validity	No definition
Worthen (1975)	Conceptual Proposition	No specific dimensions were provided.	25 general research and evaluation tasks and related competencies	NA	NA	Defined as knowledge and skills
Anderson & Ball (1978)	Empirical survey study	Two areas: 1) knowledge and content; 2) skills	33 (26 in quantitative methodology, and 7 skills)	48 out of 64 responded with a purposeful sample	Content validity	NA
Ingle & Klauss (1980)	Conceptual review	Four categories: 1) technical skills; 2) conceptual knowledge; 3) interpersonal and communication skills; 4) administrative skills.	NA	NA	NA	NA
Kirkhart (1981)	Conceptual proposition	Eight descriptive categories: 1) methodological skills; 2) evaluation knowledge (generic and specific to support evaluation skills); 3) system analysis skills; 4) political understanding; 5) professional ethics; 6) management skills; 7) communication skills; 8) interpersonal skills/ character traits.		NA	Conceptually examined against the Standards for Evaluation of Educational Programs.	Inferred as skills.
Davis (1986)	Conceptual review	No specific dimensional information was provided.	12 areas or topics	NA	NA	NA
Sanders (1986)	A conceptual review of evaluation course syllabi	Four categories: 1) history and philosophy of evaluation in education; 2) alternative approaches to evaluation in education; 3) techniques and tactics; 4) issues and special topics.	15 topical areas of knowledge and skills	NA	NA	NA

Author (year)	Methodology	Categorization/Dimensionality	# of Competencies	Sample Size	Validity	Definition
Brown & Dinnel (1992)	Empirical (survey)	Five dimensions: 1) evaluation knowledge; 2) hiring someone to do evaluation; 3) critiquing evaluations; 4) conducting evaluation in a team; 5) conduct evaluations.	15 competencies	78/78 responded with a purposive sample	Internal consistency (Cronbach Alpha = .92)	NA
Mertens (1994)	Conceptual review	Four categories: 1) knowledge/skills unique to evaluation; 2) knowledge and skills in research methodology; 3) knowledge/skills borrowed from other disciplines; 4) knowledge/skills unique to a particular discipline.	21 areas of knowledge and skills were identified.	NA	NA	Inferred as knowledge and skills
Scriven (1996)	Conceptual proposition	NA	A mix of 10 areas of knowledge and skills	NA	NA	NA
King, Stevahn, Ghere, & Minnema (2001, 2005)	Empirical study and literature review	Six dimensions: 1) professional practice; 2) systematic inquiry; 3) situational analysis; 4) project management; 5) reflective practice; 6) interpersonal competence.	61 competencies	31 participants (3 men and 28 women)	Face validity has been tested using Multi-attribute Consensus Reaching (MACR) method.	Defined as knowledge, skills, and dispositions.
Stufflebeam & Wingate (2005)	Empirical pre-post assessment	Eight areas: 1) standards/meta-evaluation; 2) evaluation approaches and models; 3) evaluation of particular areas; 4) designing evaluations; 5) evaluation methods and techniques; 6) providing evaluation training; 7) professional development; 8) developing one's own view of evaluation.	77 competency items	N/A	Face validity and content validity were tested.	NA
Dewey, Montrosse, Schroter, Sullins & Mattox (2008)	Empirical survey study	The competencies were developed and explicitly chosen for employability purposes. No dimensional information was provided.	19 competencies	Respondents included 53 job-seekers and 47 employers on two surveys.	Content validity was assessed in two focus groups with 27 employers and 17 job seekers.	NA
Russ-Eft, Bober, De la Teja, Foxon & Koszalka (2008)	Empirical survey and literature review	Four domains/dimensions: 1) professional foundations; 2) planning and designing the evaluation; 3) implementing the evaluation plan; 4) managing evaluation.	14 competencies with 86 performance or behavioral indicators.	443	Validation focused on the criticality of competencies to respondents' profession.	Defined as knowledge, skills, and attitudes.

Even though many of the 14 selected works did not reference each other extensively, particularly in conceptual developments, the advantage of chronological presentation nonetheless made it obvious to exhibit the progressive pattern of evaluator competency. The analytical results showed that, of 13 articles and one book examined, seven were empirical, and the other seven were conceptual. Also, only two resources by Stevahn et al. (2005) and Russ-Eft et al. (2008) provided formal definitions for evaluator competency. A vague definition of competency was inferred from three articles. The rest references did not define the term specifically, and no inference can be made.

Although most resources (10, 71%) presented some dimensional information or categorization schemes to organize competencies, categorizations or dimensions developed conceptually tend to be more intuitive and related to sources of competencies, such as in Mertens (1994) and Ingle and Klauss (1980); While dimensions derived from empirical studies tend to be more general and contextually-based, such as in Brzezinski and Ahn (1973), Stevahn, et al. (2005), Russ-Eft, Bober, De la Teja, Foxon, and Koszalka (2008).

Regarding validity, all seven empirical studies were validated to a certain extent, but validity assessments were mostly at the basic establishments of face validity and content validity. Regarding statistical methods, most studies did not utilize advanced statistical methods, and hence the findings had limited generalizability. For example, only descriptive analysis was used in Stufflebeam and Wingate (2005), Dewey et al. (2008), Russ-Eft et al. (2008). Stevahn et al. (2005) applied the Multi-Attribute Consensus Reaching (MACR) method to analyze the dimensionality for the ECPE competencies. However, the researchers only established face validity. Additionally, most of the seven empirical studies had small sample sizes of 100 or less except Russ-Eft et al. (2008) study with 443 respondents.

Various numbers of unevenly developed competencies ranging from 12 to 77 were included and presented in the articles. While some resources included well-written and clear structured competencies that could be transformed into measurement instruments; Other resources, however, the competencies were often ambiguously labeled, e.g., topics, categories, tasks, or areas of knowledge and skills, and hence not presented in a consistent manner, such as in Sanders (1986), Scriven (1996), and Worthen (1975).

Since four (total n = 14) articles were developed in the 2000s, the other ten were published in 1970s (3), 1980s (4), and 1990s (3), the content validity may be under question. With the fast development in the field of evaluation, evaluation researchers have gained a much better understanding of the knowledge, skills, and attitudes required for effective practices. Competencies proposed in earlier times might be obsolete.

Recent works by King, Stevahn, Ghere, and Minnema (2001, 2005) resulted in an elaborated taxonomy of evaluator competencies. This taxonomy of essential competencies for program evaluators (ECPE) was modified as a self-assessment scale and later integrated into professional development seminars with positive feedback (Ghere, et al., 2006). The initial taxonomy was developed through multiple phases of rigorous pilot tests and revisions using a Multi-Attribute Consensus Reaching (MARC) method. A number of evaluators (n = 31) participated in the validation process to determine the face validity. A subsequent revision by the same group of researchers was completed. As such, a much more user-friendly and structurally clear taxonomy was created with six distinct dimensions. Compared with other conceptual and empirical research on evaluator competencies, the ECPE instrument has apparent advantages: a) the ECPE competencies were systematically and empirically derived; b) the ECPE competencies were comprehensive and compliant to professional standards; c) the ECPE competencies have

gone through a systematic qualitative and quantitative validation process; and d) the ECPE competencies have been empirically applied and tested in professional development seminars. Nevertheless, the ECPE researchers continue to call for a more systematic and comprehensive validation using larger samples with diverse backgrounds and more advanced methodologies (King et al., 2001; Stevahn et al., 2005). Therefore, it is crucial to establish other validity of the ECPE beyond its initial face validity.

Using the ECPE framework as a benchmark against the other 13 identified resources, two levels of analyses were carried out: a) analysis at the dimensional level to discover the dimensionality of evaluator competency as a construct, and b) analysis at item level to examine the content coverage of evaluator competencies.

Mapping and Analyzing Evaluator Competency Dimensionality

The review of evaluator competencies showed that researchers had varied views on how to categorize various competencies, but all recognize that evaluator competency is a multi-dimensional construct. Netemeyer, Bearden, and Sharman (2003) emphasized the importance of establishing the construct dimensionality in developing measurement scales. The examination of competency dimensionality was carried out by comparing dimensional information provided in the ECPE (Stevahn et al., 2005) with those proposed by nine other articles/book in Table 2.2.

Stevahn et al. (2005) proposed six dimensions undergird the ECPE competencies: 1) professional practice competencies as the professional norms and values that are foundational for evaluation practice; 2) systematic inquiry competencies as the technical aspects of evaluations, e.g., design, measurement, data analysis, interpretation, and sharing results; 3) situational analysis competencies aiming to analyze and attend to the contextual and political issues related

to the evaluation; 4) project management competencies concerning the nuts and bolts of moving an evaluation from the initial stages through completion, such as negotiating contracts, budgeting, and conducting the evaluation in a timely manner; 5) reflective practice competencies as understanding one's practice and level of evaluation expertise, including an awareness of the need for professional growth; and 6) interpersonal competence competencies addressing people skills needed to conduct a program evaluation, such as written and oral communication, and cross-cultural skills.

Since reviewed frameworks organized competencies into a different number of dimensions, it would be disorienting to compare articles directly based on the number of dimensions. Instead, the six dimensions of the ECPE competencies were used as the benchmark to compare the dimensional information provided in the other nine articles. During the process, if any dimensions from the nine articles were unable to be placed in one of the six dimensions, a new dimension would be added. Additionally, if the dimensional information were unclear in the nine articles, detailed explanations or specific details would be included to facilitate the placement.

The analytical results revealed that the six dimensions in Stevahn, et al. (2005) were adequately comprehensive. While no articles/book provided competencies that were mapped onto all dimensions, all articles identified competencies that were mapped onto two dimensions of professional practice and systematic inquiry. Additionally, project management competencies were included in seven articles. Overall, the results of comparison demonstrated that the six dimensions proposed by Stevahn et al. (2005) were quite comprehensive.

Table 2.2 Analysis of Evaluator Competencies at Dimensional Level

	Articles/Book								
Dimensions:	1	2	3	4	5	6	7	8	9
Professional Practice: professional norms and values such as standards and ethics	X	X	X	X	X	X	X	X	X
Systematic Inquiry: technical aspects such as design, measurement, data analysis, interpretation, and sharing results	X	X	X	X	X	X	X	X	X
Situational Analysis: evaluability assessment, conflict, and evaluation use					X	X	X		
Project Management: negotiation on contracts, budget, resources, time management		X	X	X	X	X	X		
Reflective Practice: understanding practice and level of expertise		X	X						X
Interpersonal competence: people skills, written and oral communication, negotiation, and cross-cultural skills			X	X	X	X	X		

Articles/book:

1. Anderson and Ball (1978)
2. Brown and Dinnel (1992)
3. Brzezinski and Ahn (1973)
4. Ingle and Clauss (1980)
5. Kirkhart (1981)
6. Mertens (1994)
7. Russ-Eft, et al. (2008)
8. Sanders (1986)
9. Stufflebeam and Wingate (2005)

Mapping and Analyzing Evaluator Competencies at Item Level

Although it is critical to compare how researchers make sense of evaluator competency dimensionality, it is imperative to examine evaluator competencies at item level as dimensionality was reflected and manifested through individual competencies. The analysis was carried out by mapping competencies provided in 10 other articles in Table 1 onto the ECPE competencies (n = 61). Three specific heuristics were followed in the process: 1) competencies were reduced to a single concept of knowledge, skills, or dispositions for easier comparison; 2) comparable competencies were counted as one; 3) incomparable competencies were documented for further analysis.

The results of the comparison showed that the ECPE remained more comprehensive than any other competency framework. Comparing with the ECPE's 61 competencies, only two other frameworks have a higher number of competencies; and the rest of the 11 frameworks have fewer competencies. Also, the ECPE includes fewer competencies to represent particular dimensions efficiently, comparing with other frameworks, which have more repetitive items on a dimension. For instance, Stufflebeam & Wingate (2005) included eight items on knowledge of evaluation approaches and models (a list including utilization-focused evaluation, responsive-evaluation, CIPP evaluation model, consumer-oriented evaluation, participatory evaluation, constructivist evaluations, and theory-based evaluation); While the ECPE only has one summary item of a knowledge base of evaluation. The difference lies in how detailed to be in presenting essential competencies on theoretical knowledge. One could argue for the inclusion of more specific evaluation approaches or models such as naturalistic inquiry approach, case study evaluation, or empowerment evaluation. Other competency frameworks mostly are in agreement with the ECPE, not to include specific evaluation approaches and models.

The results also revealed some discrepancies, particularly in three areas: 1) knowledge and skills in developing evaluation instruments, 2) awareness or knowledge of legislation, regulations, or current legal issues related to evaluation, and 3) several skills in evaluation management: strategic planning, and evaluation planning. Some of the identified discrepancies, such as evaluation planning and strategic planning, are too broad and ambiguous for immediate adoption.

Table 2.3 An analysis of evaluator competencies at the item level

	Articles/Book												
Competencies	1	2	3	4	5	6	7	8	9	10	11	12	13
Professional Practice:													
Applies professional evaluation standards		x	x			x	x	x	x	x	x		x
Acts ethically and strives for integrity and honesty in conducting evaluations	x		x		x	x	x	x	x	x			x
Conveys personal evaluation approaches and skills to potential clients	x							x	x				
Respects clients, respondents, program participants, and other stakeholders	x				x			x	x				
Considers the general and public welfare in evaluation practice					x								
Contribute to the knowledge base of evaluation	x	x										x	
Systematic Inquiry:													
Understands the knowledge base of evaluation (terms, concepts, theories, assumptions)		x	x		x	x	x	x			x	x	x
Knowledgeable about quantitative methods	x	x	x		x				x	x		x	
Knowledge about qualitative methods	x								x	x		x	
Knowledge about mixed methods									x				
Conducts literature reviews		x							x			x	
Specifies program theory												x	x
Frames evaluation questions	x	x					x		x				x
Develops evaluation design		x			x	x		x		x		x	x
Identifies data sources	x	x											x
Collects data		x	x				x		x	x	x	x	
Assesses validity of data	x	x											x
Assesses reliability of data					x								x
Analyze data	x	x	x		x	x	x		x				x
Interprets data		x			x	x			x				x
Makes judgments		x						x		x			x
Develops recommendations		x											x
Provides rationales for decisions throughout the evaluation		x			x			x					
Reports evaluation procedures and results	x	x	x			x	x	x		x	x	x	x
Notes strengths and limitations of the evaluation		x											
Conducts meta-evaluation	x					x	x		x	x	x		

Situational Analysis:												
Describes the program				x								
Determines program evaluability					x			x				
Identifies the interests of relevant stakeholders		x	x							x		x
Serves the information needs of intended users		x									x	x
Addresses conflicts								x				
Examines the organizational context of the evaluation												
Analyzes the political considerations relevant to the evaluation	x			x	x	x		x				x
Attends to issues of evaluation use		x		x	x	x						x
Attends to issues of organizational change	x											x
Respects the uniqueness of the evaluation site and client							x					
Remains open to input from others		x		x								
Modifies the study as needed												
Project Management:												
Responds to requests for proposals			x							x	x	x
Negotiates with clients before the evaluation begins								x				x
Writes formal agreements												
Communicates with clients throughout the evaluation process								x				x
Budgets an evaluation				x				x		x		x
Justifies cost given information needs	x	x	x	x	x	x			x	x		
Identifies needs resources for evaluation, such as information, expertise, personnel, instruments		x		x		x		x	x		x	x
Uses appropriate technology		x		x		x		x				x
Supervises others involved in conducting the evaluation			x				x	x		x	x	x
Trains others involved in conducting the evaluation								x				
Conduct the evaluation in a nondisruptive manner		x									x	
Presents work in a timely manner				x							x	x
Reflective Practice:												
Aware of self as an evaluator (knowledge, skills, dispositions)	x			x		x	x			x		
Reflects on personal evaluation practice (competencies and areas for growth)												x
Pursues professional development in evaluation	x											x
Pursues professional development in relevant content areas	x							x				x
Builds professional relationships to enhance evaluation practice	x							x		x		X

Interpersonal Competence:										
Uses written communication skills		x	x	x			x			x
Uses verbal/listening communication skills		x	x	x			x		x	x
Uses negotiation skills				x			x		x	x
Uses conflict resolution skills	x			x			x			x
Facilitates constructive interpersonal interaction (teamwork, group facilitation, processing)	x					x	x			x
Demonstrates cross-cultural competence							x			x

Articles and Book:

1. Brzezinski & Ahn (1973)
2. Worthen (1975)
3. Anderson & Ball (1978)
4. Ingle & Clauss (1980)
5. Kirkhart (1981)
6. Davis (1986)
7. Sanders (1986)
8. Brown & Dinnel (1992)
9. Mertens (1994)
10. Scriven (1996)
11. Stufflebeam & Wingate (2005)
12. Dewey, et al. (2008)
13. Russ-Eft, et al. (2008)

Evaluator Competencies in International Evaluation Organizations

Many professional organizations and associations for program evaluation worldwide have also engaged in creating and validating professional competencies for evaluation practitioners. This section of the review takes a close look at research on evaluation competencies conducted at 11 international evaluation organizations and associations including the American Evaluation Association (AEA), the Australasia Evaluation Society (AES), the Aotearoa New Zealand Evaluation Association (ANZEA), the Canadian Evaluation Society (CES), the German Evaluation Society (DeGeval), the Department of Planning Monitoring and Evaluation in South Africa (DPME), the European Evaluation Society (EES), the International Development Evaluation Association (IDEAS), the Swiss Evaluation Society (SEVAL), the United Kingdom Evaluation Society (UKES), and the United Nations Evaluation Group (UNEG). While the evaluation associations and organizations included in the study approach competencies differently, CES and the Japanese Evaluation Society (JES) were the only professional evaluation organization that has established a set of competencies that have been applied in the credentialing process (Maicher, Kuji-Shikatani, & Buchanan, 2009; Wilcox & King, 2014). However, because the researcher was unable to locate the list of JES competencies on its website and from other sources, the competencies were not included in the analysis.

In Table 2.4, the competencies from these 11 organizations were compared in four aspects: 1) dimensionality or categories identified; 2) the number of specific competencies included; 3) whether the competencies have been empirically tested, and 4) types of validity assessed.

Table 2.4 Summary of evaluator competencies of international professional evaluation organizations

Organization	Categorization/Dimensionality	# of Competencies, items/indicators	Empirically Tested	Validity
AEA (2017)	5 domains: 1) professional practice (12 competencies) 2) methodology (16 competencies) 3) context (10 competencies) 4) planning and management (10 competencies) 5) interpersonal (10 competencies)	A total of 58 competencies	AEA member survey (2017)	Content validity
AES (2013)	7 domains: 1) evaluative attitude and professional practice (7 competencies) 2) evaluation theory (theoretical foundations, evaluative knowledge, theory, and reasoning) (10 competencies) 3) culture, stakeholders, and context (16 competencies) 4) research methods and systematic inquiry (15 competencies) 5) project management (13 competencies) 6) interpersonal skills (12 competencies) 7) evaluative activities (15 competencies	A total of 28 competencies	Forty-seven respondents in the survey study in (English, 2002).	Content validity
ANZEA (2011)	4 domains: 1) contextual analysis and engagement (4 competencies) 2) systematic evaluative inquiry (5 competencies) 3) evaluation project management and professional evaluative practice (3 competencies) 4) reflective practice and professional development (3 competencies)	A total of 15 competencies	In Wehipeihana, N., Bailey, R., Davidson, E. J., & McKegg, K. (2014)	Content validity
CES (2010)	5 domains: 1) reflective practice (7 competencies) 2) technical practice (16 competencies) 3) situational practice (9 competencies) 4) management practice (7 competencies) 5) interpersonal practice (10 competencies)	A total of 49 competencies	Yes	Content validity
DeGeval	5 fields: 1) theory and history of evaluation (4 dimensions) 2) methodological competencies (5 dimensions) 3) organizational and subject knowledge (3 dimensions) 4) social and personal competencies (5 dimensions) 5) evaluation practice (3 dimensions)	A total of 20 dimensions	No	No

DPME (2014)	5 dimensions: 1) overarching considerations (3 sub-domains) 2) leadership (1 sub-domain) 3) evaluation craft (2 sub-domains) 4) implementation of evaluations (4 sub-domains)	A total of 51 competency descriptors	A mixed method survey research (N = 42) in Goremucheche (2017)	Content validity
EES	3 domains: 1) evaluation knowledge (3 sub-categories) 2) professional practice (2 sub-categories) 3) dispositions and attitudes	A total of 30 competencies (5 competencies under each sub-category)	Survey research of EES members in 2009 and 2011	Content validity
IDEAS (2012)	For evaluators, 7 dimensions: 1) professional foundations (9 competencies) 2) monitoring systems (1 competency) 3) evaluation planning and design (4 competencies) 4) managing the evaluation (5 competencies) 5) conducting the evaluation (2 competencies) 6) communicating evaluation findings (2 competencies 7) promoting a culture of learning from evaluation (4 competencies)	Evaluators: a total of 27 competencies	Three rounds of member reviews on the framework with no sample size information specified.	Content validity
	For evaluation managers, 7 dimensions: 1) professional foundations (7 competencies) 2) monitoring systems (2 competencies) 3) evaluation planning and design (7 competencies) 4) managing the evaluation (6 competencies) 5) conducting the evaluation (6 competencies) 6) communicating evaluation findings (5 competencies) 7) promoting a culture of learning from evaluation (4 competencies)	Evaluation managers: a total of 37 competencies		
	For evaluation commissioners, 6 dimensions: 1) understand and upholds the integrity of the evaluation process (8 competencies) 2) understands and acts on the need for communication throughout the evaluation process (5 competencies) 3) supports evaluation access to people and records and the public's right to Information (5 competencies) 4) respects the terms of the agreement (2 competencies) 5) supports actions on recommendations from an evaluation (2 competencies) 6) supports monitoring and evaluation (1 competency)	Evaluation commissioners: a total of 23 competencies		

SEVAL (2014)	4 dimensions focusing on evaluation managers: 1) leadership and contextual related (5 competencies) 2) methodological competencies (9 competencies) 3) evaluation project management (7 competencies) 4) communication, social, personal (4 competencies)	A total of 25 competencies	Empirically derived from three workshops of evaluation managers (N = 17) and reviewed by members of the Federal Administration's Evaluation network.	Content validity
UKES (2013)	3 categories: 1) evaluation knowledge (3 sub-domains & 13 competencies) 2) professional practice (2 dub-domains & 13 competencies 3) qualities and attitudes or dispositions (6 competencies)	A total of 32 competencies	Reviewed by UKES members	Content validity
UNEG (2016)	5 domains: 1) professional foundations (5 competencies) 2) technical evaluation skills (5 competencies) 3) management skills (3 competencies) 4) international skills (4 competencies) 5) promoting a culture of learning for evaluation (2 competencies)	A total of 19 competencies	Empirically derived from UNEG working groups, past task force, desk review, stakeholder interviews, round-table discussions, and attendees at the UNEG 2016 Evaluation Week in Geneva.	Content validity

The results of the comparison showed that all but one organizations had examined the content validity of their competency frameworks. Additionally, nine out of 11 organizations conducted empirical studies, mostly survey research of their members, to test the content validity. Among those who reported, the majority of these empirical studies were carried out with small sample sizes ($n \leq 100$). Subsequently, analyses in these studies were mainly descriptive. Therefore, no statistically rigorous examinations were conducted to establish additional validity, e.g., construct validity. Furthermore, even though most competencies frameworks focus on evaluators, SEVAL competency framework focuses explicitly on evaluation manager competencies.

Similarly, IDEAS and UNEG provided detailed performance descriptors on each competency dependent upon years of experiences and roles in conducting evaluations. For instance, UNEG framework specified expected levels of performance for Officer, Intermediate Officer, and Senior Officer. IDEAS, on the other hand, provided two sets of competencies for evaluators and evaluation managers. Even though both sets of competencies were organized under the same seven dimensions, additional expected competencies or higher level of competencies were expected of managers. For IDEA evaluation commissioners, the competency dimensions focus on a macro or evaluation policy level of ensuring evaluation integrity, interpretation of findings, and supporting evaluations access to the public, and facilitating evaluation practices, such as: establishing evaluation recommendation tracking systems and supporting monitoring and evaluation capacity building.

An analysis of all available competency dimensions in Table 2.4 revealed 11 non-overlapping dimensions, 1) professional practice (other terms were also used, such as evaluation practice, and evaluation planning and design); 2) systematic inquiry (other terms were also used,

such as methodological competencies, technical evaluation skills, technical practice); 3) situational practice (other terms were also used, such as contextual analysis and engagement, context, culture, stakeholders, and context); 4) project management (other terms were also used, such as planning and management, evaluation project management, and managing the evaluation); 5) interpersonal practice or interpersonal competencies; 6) evaluation theory (other terms were also used, such as evaluation knowledge, and theory and history of evaluation); 7) reflective practice or reflective practice and professional development; 8) leadership; 9) competencies promoting a culture of learning from evaluation; 10) qualities and attitudes or dispositions, and 11) international skills.

Some of the dimensions, such as international skills, and competencies promoting a culture of learning, are essential to several organizations, e.g., IDEAS and UNEG frameworks, with a primarily international development scope, hence may not be represented and applicable in other organizational contexts. However, it becomes evident that a set of common core competency dimensions transcend organizational differences, such as professional practice and systematic inquiry, despite different terms were adopted.

Mapping and Analyzing Evaluator Competency Dimensionality

At the dimensional-level in Table 2.5, while CES and UNEG have competencies that were mapped onto all six ECPE dimensions, AEA, DPME, and IDEAS competencies were mapped onto five dimensions. Competencies in EES, Seval, and UKES frameworks were mapped onto four dimensions. The rest (ANZEA, AES, and DeGeval) have competencies mapped onto three dimensions. Moreover, systematic inquiry dimension was the only dimension included in all 11 frameworks, but with different terms. Professional practice, situational

analysis, project management, and interpersonal competence dimensions were well represented in the competency frameworks of 11 organizations. Among all six dimensions, reflective practice is the least-mapped competency dimension. Only three organizations (ANZEA, CES, and UNEG) have included a competency dimension that pertains to reflective practice.

Table 2.5 An Analysis of Evaluator Competencies at Dimensional Level

Dimensions:	AEA	AES	ANZEA	CES	DeGeval	DPME	EES	IDEAS	SEVAL	UKES	UNEG
Professional Practice: professional norms and values such as standards and ethics	X	X		X	X	X	X	X		X	X
Systematic Inquiry: technical aspects such as design, measurement, data analysis, interpretation, and sharing results	X	X	X	X	X	X	X	X	X	X	X
Situational Analysis: evaluability assessment, conflict, and evaluation use	X	X		X		X		X	X		X
Project Management: negotiation on contracts, budget, resources, time management	X		X	X		X	X	X	X	X	X
Reflective Practice: understanding practice and level of expertise			X	X							X
Interpersonal competence: people skills, written and oral communication, negotiation, and cross-cultural skills	X			X	X	X	X	X	X	X	X

Note. AEA – American Evaluation Association (https://www.eval.org/page/competencies);

AES – Australasia Evaluation Society (https://www.aes.asn.au/images/stories/files/Professional Learning/AES_Evaluators_Competency_Framework.pdf);

ANZEA – Aotearoa New Zealand Evaluation Association (https://www.anzea.org.nz/aotearoa-evaluations-competencies/);

CES – Canadian Evaluation Society (https://evaluationcanada.ca/txt/2_competencies_cdn_evaluation_practice.pdf);

DeGeval – German Evaluation Society (https://www.degeval.org/fileadmin/Publikationen/Publikationen_Homepage/Recom_Education_Training.pdf);

DPME –Department of Planning Monitoring and Evaluation in South Africa

(http://www.dpme.gov.za/keyfocusareas/evaluationsSite/Evaluations/Competencies_14%2007%2010.pdf);

EES – European Evaluation Society (https://www.europeanevaluation.org/sites/default/files/ees-leaflet-FINAL.pdf);

IDEAS – International Development Evaluation Association (https://ideas-global.org/the-competencies-framework/);

SEVAL – Swiss Evaluation Society (http://www.seval.ch.ranger.iway.ch/documents/Competences/Brochure_SEVAL-Kompetenzen%20Evaluationsmanag-e_final.pdf);

UKES – United Kingdom Evaluation Society (https://www.evaluation.org.uk/images/ukesdocs/UKES_Evaluation_Capabilities_Framework_January_2013.pdf);

UNEG – United Nations Evaluation Group (www.unevaluation.org/document/download/2610).

Table 2.6 Item-level analysis of evaluator competencies of international evaluation organizations

Competencies	AEA	AES	ANZEA	CES	DeGeval	DPME	EES	IDEAS	SEVAL	UKES	UNEG	Count
Professional Practice:												
Applies professional evaluation standards		X	X	X	X	X		X			X	7
Acts ethically and strives for integrity and honesty in conducting evaluations	X	X	X	X		X		X		X	X	8
Conveys personal evaluation approaches and skills to potential clients	X	X						X				3
Respects clients, respondents, program participants, and other stakeholders		X	X	X			X			X	X	6
Considers the general and public welfare in evaluation practice	X	X		X								3
Contribute to the knowledge base of evaluation		X	X				X	X	X	X		6
Systematic Inquiry:												
Understands the knowledge base of evaluation (terms, concepts, theories, assumptions)	X	X	X	X	X	X	X	X	X	X	X	11
Knowledgeable about quantitative methods	X	X	X	X	X	X	X		X	X	X	10
Knowledge about qualitative methods	X	X	X	X	X	X	X		X	X	X	10
Knowledge about mixed methods	X	X	X	X		X			X			6
Conducts literature reviews	X	X	X									3
Specifies program theory	X			X			X	X		X	X	6
Frames evaluation questions	X	X	X	X	X					X		6
Develops evaluation design	X	X	X	X	X	X	X	X		X	X	10
Identifies data sources	X	X		X		X			X		X	6
Collects data	X	X	X	X	X		X			X	X	8
Assesses validity of data		X		X	X	X			X	X		6
Assesses reliability of data		X		X	X				X	X	X	6
Analyze data	X		X	X	X	X	X	X	X	X	X	10
Interprets data	X	X	X	X	X	X	X	X	X		X	10
Makes judgments	X	X	X	X			X			X	X	7
Develops recommendations	X		X	X	X	X		X	X		X	8
Provides rationales for decisions throughout the evaluation	X											1
Reports evaluation procedures and results	X	X	X	X	X	X	X			X	X	9
Notes strengths and limitations of the evaluation	X	X			X		X					4
Conducts meta-evaluation	X											1
Situational Analysis:												
Describes the program	X	X										2
Determines program evaluability		X	X	X		X					X	5
Identifies the interests of relevant stakeholders	X	X	X	X		X	X		X		X	8
Serves the information needs of intended users				X							X	2
Addresses conflicts	X					X		X	X	X	X	6

Competencies	AEA	ANZEA	AES	CES	DeGeval	DPME	EES	IDEAS	SEVAL	UKES	UNEG	Count
Examines the organizational context of the evaluation	X	X	X	X	X	X			X		X	8
Analyzes the political considerations relevant to the evaluation	X	X	X	X	X		X			X		7
Attends to issues of evaluation use	X	X	X	X		X	X	X	X	X	X	10
Attends to issues of organizational change				X	X				X			3
Respects the uniqueness of the evaluation site and client	X	X	X	X			X		X		X	7
Remains open to input from others		X						X				2
Modifies the study as needed			X		X		X	X	X			5
Project Management:												
Responds to requests for proposals		X										1
Negotiates with clients before the evaluation begins			X		X	X						3
Writes formal agreements												0
Communicates with clients throughout the evaluation process		X	X		X	X	X	X	X			7
Budgets an evaluation					X	X					X	3
Justifies cost given information needs		X	X		X							3
Identifies needed resources for evaluation, such as information, expertise, personnel, instruments	X		X	X	X	X		X	X			7
Uses appropriate technology	X	X	X									3
Supervises others involved in conducting the evaluation	X		X	X							X	4
Trains others involved in conducting the evaluation												0
Conducts the evaluation in a nondisruptive manner												0
Presents work in a timely manner	X	X						X				3
Reflective Practice:												
Aware of self as an evaluator (knowledge, skills, dispositions)		X	X	X			X				X	5
Reflects on personal evaluation practice (competencies and areas for growth)		X		X		X	X					4
Pursues professional development in evaluation	X	X				X		X	X	X	X	7
Pursues professional development in relevant content areas		X				X		X				3
Builds professional relationships to enhance evaluation practice	X	X		X			X	X				5
Interpersonal Competence:												
Uses written communication skills	X	X	X	X	X	X	X			X	X	9
Uses verbal/listening communication skills	X	X	X	X	X	X	X				X	8
Uses negotiation skills	X		X	X	X	X	X	X	X	X	X	10
Uses conflict resolution skills	X			X		X	X	X			X	6
Facilitates constructive interpersonal interaction (teamwork, group facilitation, processing)	X	X	X	X	X	X	X	X			X	9
Demonstrates cross-cultural competence	X	X	X	X	X	X	X		X	X	X	10
Total Number of ECPE Items Mapped:	40	38	41	39	28	31	28	23	23	22	33	

Mapping and Analyzing Evaluator Competency at Item Level

At item level as shown in Table 2.6, four organizations, AEA, ANZEA, AES, and CES, shares the most number of competencies with ECPE. Of all six ECPE dimensions, the competencies in interpersonal competence dimension have the highest occurrences across all organizational competency frameworks, except the competency of "uses conflict resolution skills." In professional practice dimension, the competency of "acts ethically and strives for integrity and honesty" has the highest occurrence. Additionally, in systematic inquiry dimension, competencies about evaluation knowledge base such as evaluation theories, concepts, quantitative/qualitative methods, evaluation design, and data analysis and interpretation are among the highest occurrences across all frameworks. Furthermore, in situational analysis dimension, three competencies in identifying stakeholder interest, examining the organizational context, and attending to evaluation use have the highest occurrences. Concluding from the comparison analyses, ECPE remains as the most comprehensive evaluator competency framework.

Other than the comprehensiveness, the ECPE framework also has several other advantages. Firstly, the ECPE competencies were undergone rigorous crosswalk comparison with the *Program Evaluation Standards* endorsed by the Joint Committee on Standards for Educational Evaluation (1994), the *Guiding Principles for Evaluators*, and the CES competency framework. Secondly, the ECPE competencies were written in a behavioral approach that "tends to task-analyze competencies into discrete behaviors" (Stevahn et al., 2005, p. 48). Thirdly, the ECPE competencies have gone through iterations of improvements and empirical research to establish its validity and usability. Lastly, the ECPE framework has served as the foundation or lent its influences to many organizations such as the Canadian Evaluation Society (Wilcox &

King, 2014), to set up their evaluator competency framework. The benefit of studying ECPE will be consequential.

Empirical Research on the ECPE

Since the ECPE is by far the most comprehensive and rigorously constructed competency framework, a series of empirical studies have been conducted in an attempt to utilize the framework as a measurement instrument to answer a number of research questions.

As the most comprehensive competency framework currently available, the competencies included in the ECPE framework have not gone through rigorous validity examination until the recent study by Wilcox (2012), which adopted a unified approach to examine the validity of the ECPE. Wilcox applied six criteria of validity including, 1) content-related validity to answer the questions of to what extent the ECPE competencies measure evaluators' competence; 2) substantive-related validity to answer the question of how inclusive or comprehensive the ECPE framework is; 3) structural-related validity to address the question—to what extent the ECPE dimensions reflect the factor structure of evaluator competencies; 4) generalizability-related validity to assess the extent to which the competencies put forth in the ECPE framework are relevant to evaluators in different content areas; 5) externally-related validity to correlate evaluator competence with evaluator competency frameworks other than the ECPE; and 6) consequence-related evidence to examine the extent to which any negative consequences are existent when using and interpreting the ECPE.

Based on data collected from the surveys and interviews, Wilcox study addressed all research questions except the structure-related validity. Specifically, a survey instrument was developed to collect evaluators' perceived necessity (5-point Likert scale from *not at all*

*necessar*y to *extremely necessary*) of each of the 61 ECPE competencies. The survey results were analyzed to address content-related and substantive-related validity. Additionally, an interview protocol was also developed to solicit responses from practicing evaluators' general comments on each of the six general categories of ECPE. The interview results were analyzed to address generalizability-related, externally-related, and consequence-related validity.

On content-related validity for ECPE, the study concluded that 58 out of 61 competencies were rated strongly necessary and the other three were rated moderately necessary. The fact that there were no changes regarding adding or removing any competencies from ECPE suggested strong substantive-related validity. Meanwhile, the study also reported mixed results for generalizability-related validity, limited externally-related validity, and strong consequence-related validity. Overall, Wilcox (2012) study has extended the existing ECPE research and, to a great extent, systematically addressed important validity issues.

Kaesbauer (2012) conducted a study on evaluator competencies by examining 26 doctoral programs focusing on evaluation training across the United States. The study utilized a multi-method and multi-sample approach to answer two questions of what evaluator competencies were taught in these doctoral programs and how evaluator competencies were taught. The foundational competencies adopted in the study were based on the ECPE and CES competency frameworks. The study concluded that the ECPE has a significant influence on doctoral students and doctoral program curriculum. Of all the competencies, data collection, analysis and interpretation, and planning and design competencies were the most frequently taught competencies in the doctoral programs. Competencies of project management and ethics competencies were the least frequently taught or addressed. The study also demonstrated that the

ECPE was one of the most influential competency frameworks that can be utilized effectively in assessing the current state of the evaluation of educational programs.

Both studies were built on the ECPE framework and contributed to moving forward the research agenda in the area of evaluator competency. The study by Wilcox (2012) took a step further to extend the research effort by King and colleagues (2001, 2005) on the ECPE framework. With the face validity and content validity for ECPE well established, the next step should be to continue with the establishment of the construct validity. Specifically, the dimensionality of the evaluator competency should be examined within the context of a large sample.

The most recent study by Galport and Azzam (2017) used the ECPE framework to examine the gap between evaluator perceived importance of competencies and evaluator training needs. Researchers discovered that three competencies from the professional practice domain were viewed as the most important, and the competency of conducting meta-evaluation was rated as the least important. Additionally, the study also revealed how evaluator characteristics, such as gender, professional identity, age, experience, and work setting, related to their views on competency domains. For example, an evaluator's gender had a significant impact on how they viewed the importance of professional practice, situational analysis, and reflective practice domains. Female evaluators were more likely to identify these domains as important than male evaluators. While the majority of respondents viewed project management as unimportant, a significantly higher percentage of evaluators in a higher education setting rated project management as important compared to evaluators in other settings. Evaluators with less than two years and more than 16 years of experience rated interpersonal competence as important, evaluators with between 2 and 15 years of experiences did not view interpersonal competence as

important. Furthermore, the study also identified six competency gaps between evaluator rating on importance and need for training.

Although Galport and Azzam (2017) have advanced research on the ECPE competencies with relatively large sample size (N = 403), respondents were randomly assigned to respond to only 31 competencies in one of two conditions, importance ranking or training need. This data collection strategy was efficient but greatly diminished the statistical power. In addition, no statistical inferences could be drawn between evaluator importance rating and their identified training needs.

Section Summary

This section first explored how competence and competency have been defined and their differences; then, a formal definition for evaluator competencies was provided; lastly, the related literature on various evaluator competency frameworks was reviewed and analyzed at a dimensional level as well as an item/competency level, using ECPE as a benchmark.

Results of these comparisons revealed that the ECPE framework has distinct advantages: a) the ECPE competencies were systematically and empirically derived; b) the ECPE competencies were comprehensive and compliant to professional standards; c) the ECPE competencies have gone through rigorous qualitative and quantitative validation process; and d) empirical studies have been carried out and established face and content validity. Despite the advantages, existing research on the ECPE competencies has many limitations, e.g., the small sample sizes affecting the accuracy of the findings. As such, much work has to be done in establishing construct validity in order to use the ECPE as a measurement scale more rigorously in a large sample context. Researchers, therefore, have called for further systematic and

comprehensive validation with larger diverse samples and using more advanced methodologies (King et al., 2001; Stevahn et al., 2005; Wilcox, 2012;).

The next section reviews relevant research literature for evaluation practice, where evaluator competencies including knowledge, skills, and attitudes are crucial for practitioners.

Evaluator Practice

In a field as practical as evaluation, it is troublesome that there is still very little known about how evaluators conduct evaluations in their practice (Fitzpatrick, Christie, & Mark, 2009). Smith and Brandon (2008) further argued that evaluator practice has yet to become the central topic of evaluation research. In this section, the review focuses on two aspects of evaluation practice. First, conceptual discussions on evaluation practice from various perspectives are presented; then, empirical studies are presented and discussed; lastly, evaluator practice as a construct is examined.

Nature of Evaluation Practice

Smith and Brandon (2008) summed up evaluation practice as the process of conducting evaluations. To them, evaluation practice deals with issues of exploring feasible, practical, and cost-effective ways to conduct evaluations and making appropriate choices under various contextual limitations (p.16). Shadish et al. (1991) provided an alternative view on evaluation practice as "the tactics and strategies evaluators follow in their professional work, especially given the constraints they face" (p. 32). They contended that other than making decisions on limited resources to conduct feasible evaluations, practitioners also made decisions on what roles to assume, what evaluation questions to raise, and what methods and designs are appropriate.

Schwandt (2005) proposed that there were two viewpoints on the nature of evaluation practice in terms of its evidence base. The first viewpoint, the technical rationality, functions as methods, criteria, and goals for evaluation practice. To be more specific, evaluation practice should only be conducted on the basis of scientific methods with the goal of evaluation practice to generate scientific knowledge and guides future practice. Schwandt further suggested, under

this technical rationality view, there should be "at least an implicit skepticism regarding any practice that cannot justify itself as a worthwhile social understanding in terms of scientific rationality, technical expertise, and effectiveness" (p. 97). This view is compatible with those of Smith and Brandon (2008) and Shadish, et al. (1991), characterizing evaluation practice as an applied research activity to "use considerable methodological skills to determine whether a practice intervention 'works'" (p. 98).

On the other hand, the second viewpoint is based on an integrated outlook of evaluation practice as a complex decision-making process, which involves "simultaneous consideration of evidence, professional values, political considerations, and individualized goals" (Schwandt, 2005, p. 98). Evaluation practice, under this view, is beyond the simple application of scientific knowledge. Rather, it is a process of generating practical knowledge as well as rationalizing and interpreting complex decisions made under various contexts. In other words, evaluation practice is a pedagogy or a practical hermeneutics (Schwandt, 2002, p. 66).

These two competing views, according to Schwandt (2002), are rooted in two different philosophical foundations (modernist/naturalistic and humanistic/hermeneutic) in six aspects: 1) object of evaluation; 2) attitude towards the world; 3) the nature of educational experience; 4) the nature of knowledge; 5) conception of dialogue; 6) basis of authority or expertise; (p. 12).

In deliberating these two views on evaluation practice, Schwandt (2002) also presented his logic of emphasizing the second view of practice as practical hermeneutics:

> It is this second view of evaluation that I have been talking and writing about for many years. I do not object to the idea of generating evaluation knowledge of "what works"—that is, to conducting theory-based or experimental studies of how and why a particular social intervention or program achieves its intended effects. This kind of scientific evidence can be helpful to practitioners. What I worry about is that science-based or evidence-based approaches to practice are too readily becoming an ideology that aims to

> instill scientific rationality as authoritative for everyday practice, that threatens to eclipse practical knowledge and reasoning, and that comes dangerously close to regarding the practitioner as a judgmental dunce, who if left to his or her own way of doing things will inevitably be inefficient, ineffective, and squander precious social resources. We are at risk in believing in a false dichotomy: that the only legitimate knowledge for practice is scientific, for all else is unreliable intuition, habit, custom, or mere belief. We are in danger of accepting without reservation the myth of a scientifically guided society, a society in which science (not everyday life) occupies center stage (p. 99).

Practical Knowledge & Technical Knowledge

Central to these views of evaluation practice are two different kinds of knowledge: technical knowledge and practical knowledge. Technical knowledge, also as scientific, cognitive, and professional knowledge, is referred to as formally acquired and taught the specific subject and content knowledge from education and training (Cheetham & Chivers, 2005, p. 55). Schwandt (2008) implied that technical knowledge is "the skillful performance of technique or the competent carrying out of procedures" (p.35). Practical knowledge, however, is the tacit knowledge that can only be revealed by one's actions. He further contended "this kind of knowledge is shown or demonstrated via the kind of pre-reflective familiarity one has with ideas and concepts used to express oneself, one's ability to be present in and handle a situation, and one's capability to exercise judgment of when to apply, or not apply, a particular kind of understanding of a situation" (p. 31). In other words, practical knowledge can be viewed as implicit decision rules developed through experiences. These implicit decision rules are highly contingent upon various situations.

Schwandt (2002) believed that practical knowledge is required by, according to Aristotle, productive activity (poiesis), engaged by social science researcher and evaluator as "a maker or craftsman" (p. 45). However, it is not adequate to have just practical knowledge to engage in productive activity. A cognitive capacity, a "habitual ability" (p. 46), is another ingredient to enable evaluators and researchers to create reliable solutions to various problems. This ability,

capacity, or competence, developed by experience, makes it possible for experienced evaluators to observe nuances in various situations, and make appropriate judgments and decisions on applying strategies and approaches accordingly (Schwandt, 2008). Schwandt also noted that the development of such an ability is equally crucial as the technical aspect of evaluation knowledge.

The difference between these two kinds of knowledge, according to Schwandt (2008), lies in the defining characters of instrumental reasons for technical knowledge and judgment for practical knowledge. He observed a tendency of theorizing evaluation practice, which seeks to justify and assimilate practical knowledge into technical knowledge. The assimilation effort reflected the traditional narrow view of practice as technical rationality. Schwandt warned us of the danger of this tendency of reducing evaluation practice into a unidimensional mechanical application of tools and implementation of procedures. Furthermore, he argued, practical knowledge developed through experience is indispensable for good practice, because "no matter how well developed and sophisticated the scientific-technical knowledge base for practice, the skillful execution of that practice is ultimately a matter of practical wisdom" (p. 37).

Implications for Evaluation Research

Discussions on the nature of practice and practical/technical knowledge have profound implications for empirical research on evaluation practice. In building a practical knowledge base, researchers need to examine different aspects of decision-making including, how evaluation practitioners made decisions, what decisions were made under various circumstances, how experts/experienced evaluators made different decisions from inexperienced evaluators, and how these decisions made under various situations relate to evaluation theory building.

Another implication of studying evaluation practice and theory together is that evaluation practice needs both technical knowledge and practical knowledge. Similarly, theorists need to build evaluation theories based on these two kinds of knowledge to guide practice. Evaluation practice, in turn, would inform and empirically test various evaluation theories and models. Contingency theory in Shadish, et al. (1991), for example, functions as a heuristic device to provide practical guidance to specific scenarios.

Empirical Research on Evaluation Practice

Two major characteristics are observed in existing empirical studies on evaluation practice: 1) empirical studies often examine and discuss evaluation practice in relation to evaluation theory; 2) empirical studies examine evaluation practice through decisions evaluators made in practice. Sometimes, these two characters were reflected and addressed in the same study. The following empirical studies are discussed and presented to reflect these two characteristics.

Evaluation Practice in Relation to Theory. Shadish and Epstein (1987) were pioneers to study patterns of evaluation practice and influences of training, working-settings, and evaluation theories. Evaluation theories were defined as a collection of classic writings and concepts of theorists highlighted in Foundations of Program Evaluation: Theories of Practice (Shadish et al., 1991). Researchers examined evaluation practice with 74 questions concerning the purposes of evaluation, influences on the decision to conduct evaluations, self-perceived evaluator roles, the sources of evaluation questions, the data sources, the sources of dependent variables, methods used, and measures taken to facilitate evaluation use. Other than discovering four practice patterns, Shadish and Epstein also furthered their inquiries by investigating the relationships of

practice patterns with evaluator educational/training background, work setting, and theoretical influences.

In the earliest and most comprehensive studies to examine evaluation practice, Shadish et al. (1991) concluded that the low level of familiarity with evaluation theory exhibited “a danger of scholarly illiteracy in evaluation about its own writings and concepts (p. 586)”. They called for efforts to increase knowledge breadth and width. They also identified the gap between the academic and service-oriented practice patterns and encouraged continued efforts to integrate both academic and service-oriented practices.

Using multidimensional scaling, Williams (1989) developed a quantitative taxonomy and captured perceptions of 14 theorists on how similar evaluation theories are to each other and how theorists’ practices align with the resultant theoretical dimensions. The study discovered four distinct dimensions that evaluation theories form: 1) quantitative versus qualitative; 2) accountability versus policy orientation; 3) client participation versus nonparticipation; 4) general utilization versus decision-making utilization. The cluster analysis then classified theorists on each of the four theory dimensions into three groups of application approach, flexible approach, and formal approach. The study concluded that even though evaluation theorists tend to be more diversified in their theoretical claims and arguments, their practices demonstrated fewer differences.

Christie (2003) conducted her study using a similar rationale and method as in Williams (1989) to investigate how evaluation theories connected with practice. The study first recruited eight theorists to frame their theoretical approaches into descriptors under the use, value, and method framework by Alkin and House (1992); then a survey instrument with 38 questions

concerning evaluation practices was created based on the descriptors; finally the survey questionnaires were sent to 138 evaluators who were engaged in evaluating the Healthy Start program in California. The study concluded that only a third of participants (36%) demonstrated similarities in their practice with that of a theorist. A majority of participants did not utilize any theoretical framework put forward by theorists. To a certain extent, the study is consistent with what Shadish and Epstein (1987) concluded, there is a low familiarity of theoretical evaluation knowledge, and there's still a gap between academic theorists and service-oriented practitioners.

Another study by Barela (2005), sought to develop an implicit prescriptive model of how evaluations were conducted in a school district. The researcher aimed to determine the contextual contingencies (political and otherwise) that influence how evaluators made sense of their practices, and how they make evaluation-related decisions in a school district. The researcher conducted 17 interviews and observed six additional evaluators as they went about their evaluation work. The study also gauged evaluators' knowledge about prescriptive models.

Barela's finding that evaluators were only vaguely aware of prescribed evaluation approaches was consistent with those of Shadish (1987), Williams (1989), and Christie (2003). It also suggested that direct questioning should not be an effective way to ask evaluators about prescribed models. Instead, Barela (2005) observed one evaluator talking to a supervisor about capacity building, and concluded that evaluators do not make many decisions based on knowledge of formal theories.

Evaluation Practice as a Decision-Making Process. Researchers also investigated how evaluators made decisions about various aspects of their practice. Benkofske (1996) studied how evaluators made decisions regarding data collection methodologies. Using semi-structured

interviews, the researcher investigated interactions between clients and evaluators on types of data to collect, roles played in the process and influencing factors for these decisions. The study results revealed four types of decision-making behaviors among clients, evaluators, and the combination of both. Educational training and experiences of evaluators and clients played a crucial role in data collection strategies. The researcher concluded that time constraints, professional standards, views on different paradigms, cost, and client needs all influenced the decisions on data collection.

Kundin (2008) investigated how evaluators make decisions on how to approach evaluations in various situations, and how they adopted working logic-in-use and logic-in-action in such situations. The study proposed a conceptual framework to attempt to explain how evaluators make practice decisions. The framework, centering on situation awareness as an umbrella concept, is integration and application of naturalistic decision-making research with evaluation practice. Using semi-structured interviews, the researcher applied a naturalistic decision-making framework and studied 11 evaluators making practice decisions. The study suggested that evaluators' practical knowledge played the dominant part in practice, and evaluators did not particularly follow any specific theoretical guidance.

Tourman (2009) designed and carried out a qualitative study with extensive interviews to inquire how evaluators made design decisions, and how these decisions relate to evaluation theories. She proposed that the choices evaluators have to make in their practices require both technical knowledge and practical knowledge, as she pointed out, these choices were directly related to evaluation theories and assessment of various situations. The activity theory framework used in the investigation discovered crucial behaviors and practical strategies that evaluators engaged in their practice. The conclusions from the study resonated with Schwandt's

(2008) argument, that evaluation practice requires both technical knowledge and practical knowledge.

The three studies presented in this section had the same goal of attempting to discover practical knowledge bases for evaluation decision-making. Two studies utilized various forms of interviews to solicit evaluators' reasoning processes. However, one of the problems of using the interview as a method to uncover practical knowledge is that practical knowledge is innately implicit and tacit. Evaluators themselves might not know consciously how the decisions were made or may have just tried to come up with reasonable explanations (Carroll & Johnson, 1990, p. 32). Tourmen (2009) bridged this gap by combining observations of evaluator activities in real scenarios and simulated activities. The triangulation of observation data allowed the researcher to compare what evaluators did, what evaluators proposed to do, and reflections on what they did, and consequently increase the validity of the findings of the study.

Decision-making studies on evaluation practice presented above made several assumptions: 1) situations where evaluators made decisions are objective; 2) evaluators are self-aware while making various decisions, or at least can reflect on how and why they made those decisions; 3) heuristic rules facilitating evaluation decisions are simplistic. However, these assumptions are often not guaranteed in real situations. Carroll and Johnson (1990) summed up some critical findings in general decision research, which might shed some light to orient decision-making studies:

- Decisions are not consistently made based on rationality;
- Limited human mental capacity often simplifies situations, and results in limited decisions, which may not reflect the accuracy of the situational information;
- People's perceptions influence their decisions. Depending on how problems are

framed, different decisions can be made;

- Heuristic rules and strategies can facilitate the decision-making process, but also may not produce the most appropriate decisions in the situation;
- Heuristic rules may help decision-makers to avoid assessing trade-offs. However, this avoidance also obstructs decision makers from seeking out the best decisions.
- Decision makers are not self-aware, and they often do not understand their own implicit decision rules;
- Learning from past decisions is a long and slow process;
- Groups do not necessarily make better decisions than individuals.

The summary is not intended to discredit the progress made in decision research. Rather, these findings characterize research in the field of decision-making. In conducting decision studies, these characteristics should be heeded in designing and understanding future research.

Evaluator Practice as a Construct. In studying how evaluators practice, some researchers (Christie, 2003; Shadish & Epstein, 1987; Williams, 1989) viewed evaluator practice as a construct or a latent variable. They assumed that evaluators demonstrate certain similarities in their practices. This view is particularly useful in quantitatively studying how evaluator practices as a construct relate to other variables such as evaluators' education background, professional settings, and their competencies.

Using a higher-order factor analysis method, Shadish and Epstein (1987) studied practices of 604 evaluators and identified four distinct practice patterns. The academic pattern represents practice focusing on societal betterment and pursuit of basic scientific inquiry, and often utilize quantitative methods to make results available to the general public. The

stakeholder-service pattern characterizes practice that aims to fulfill obligations to clients/customers. Evaluators in this pattern of practice work closely with stakeholders for information needs, question formulation, and program success criteria. For the decision-driven pattern, evaluators typically make evaluation decisions on the basis of cost-benefit considerations, assume the role of a servant of program constituents, and formulate evaluation questions based upon the pending decisions and legislation. In the last, outcome pattern of evaluation practice, evaluators set the judgment of program merit and worth as a primary purpose and assume the role of a methodological expert to guide stakeholders.

Williams (1989) focused on 14 theorists' practices in seven aspects, and yielded two patterns of practices: 1) interpretative-descriptive versus scaled-causal; 2) general audience versus specific end user. The study concluded that theorist practices demonstrated a more unified pattern than their theoretical propositions.

Using the same method and approach, Christie (2003) compared eight theorists' theoretical propositions and practices with that of 138 evaluators. The results of the study yielded two patterns connecting theory and practice: stakeholder involvement and method proclivity. The study utilized a theory classification framework of the method, use, and value. The research concluded that evaluation practitioners did not conform to any particular theoretical guidance in their practices.

Although all three studies examined evaluator practice patterns, they differed in the focus and methodology. Shadish and Epstein viewed evaluation practice pattern as a latent variable manifested by eight variables and used factor analysis to generate four practice patterns. Further regression analyses were carried out treating practice patterns as outcome variables and educational/training background, work settings, and theoretical influences as predictors.

Williams and Christie, on the other hand, derived their practice patterns by using multi-dimensional scaling, and evaluation practice patterns were implied.

The inherited problems of Williams (1989) and Christie (2003) research may challenge future follow-up studies. The instruments used in the two studies were conducted by creating survey questions based on the practices of various theorists. This approach itself may be problematic because both studies attempted to align evaluator practices with those of a selected number of theorists. Even though these theorists only represented a small sample of available theorists, the researchers of the two studies tried to generalize the findings to all theorists. Christie claimed, for example, "only 36 percent of the evaluators were within meaningful proximity of a theorist, indicating that most do not use frameworks aligned with a specific theoretical model" (p. 33). It is possible, however, the evaluator practices could align with other theorists who were not included in the study because theorists in both studies simply were not exhaustive. Williams, in her study, made assumptions that the sampled theorists were knowledgeable about each other's theories and practices that they could adequately relate their own works with others. Furthermore, both studies identified discrepancies of theoretical positions and practices of theorists. If theorists did not practice the way they advocated, how and why should evaluators' practices be compared to theorists' practices?

In Shadish and Epstein (1987), the instrument to measure practice patterns was rooted in eight aspects of evaluation practice and provided the most comprehensive assessment of evaluation practice, compared to similar studies. Moreover, the results of the present study were based on a much larger sample, which made the generalization to other evaluators possible. The weakness of the study was a lack of reported psychometric properties and construct validity. However, the initial exploratory factor analysis provided a factor structure, which can be used in

confirmatory studies. The current study intends to further the investigation by Shadish and Epstein (1987) in three aspects: 1) establish construct validity of evaluation practice by confirming the factor structure in the current evaluator population; 2) conduct a multi-group analysis to test the psychometric property of measurement invariance; 3) explore the relationship between evaluation practice as a construct with evaluator competencies.

The Relationships between Evaluator Competencies and Evaluator Practice

The relationship between competencies and professional practices has been implicitly assumed in the research literature. For instance, Schön (1983) argued that being reflective was an essential competency for practitioners in their professional practice. In the evaluation field, conceptual frameworks such as *Fundamental Issues in Evaluation* (Smith & Brandon, 2008) link evaluator competencies with professional practice at a macro level. However, there have not been any empirical studies explicitly examining the relationship between the two constructs. The current study is one of few empirical studies to investigate the relationship and potentially provide evidentiary support to these conceptual frameworks.

Chapter Summary

The Chapter has detailed discussions on the two critical constructs of evaluator competency and evaluation practice. First, evaluator competencies have been defined; then a thorough literature review of evaluator competencies studies have been conducted, and the results have been compared at dimensional and item levels. The comparison revealed that ECPE provides the most comprehensive coverage of evaluator competencies. Then, a review of evaluation practice has been provided. Lastly, the relationship between the evaluator competencies and evaluation practice has been connected.

The next chapter describes the methodological details of the study, including participants, measurement instruments, data collection, and specific analysis strategies.

CHAPTER III: METHOD

In the previous chapter, a thorough review of related literature exhibited the status and development of the two constructs of interest. In addition, a detailed construct analysis was included for evaluator competency and evaluation practice. This chapter focuses on the methodological framework used in the study, structural equation modeling (SEM), and specific analytical procedures at three phases are detailed.

Review of Research Questions

The present study has three main goals: (1) establish the construct validity for the scale of evaluator competencies adapted from the ECPE framework; (2) confirm the factor structure of the evaluator practice scale adapted from Shadish and Epstein (1987); and (3) test hypothesized relationships between the two constructs. The research questions are addressed in three analytical stages.

Exploratory Phase

The goal at this phase is to explore the factor structures of the two scales and establish the measurement foundation for later analyses.

- **R1.** Does the factor structure of the ECPE scale conform to a six-factor structure conceptualized by Stevahn et al. (2005)?
- **R2.** Does the factor structure conform to the 22 first-order factor structure yielded in 8 sub-domains of evaluation practice in the Shadish and Epstein (1987) study?
- **R3.** Does the higher-order factor structure conform to the four patterns yielded in the Shadish and Epstein (1987) study?

Confirmatory Phase

The goal of this phase is to confirm the results in the previous stage and also to establish sound measurement models and examine psychometric properties such as reliability and measurement invariance.

- **R4.** Does the factor structure yielded from R1 achieve reasonably good model fit?
- **R5.** Does the factor structure yielded from R2 achieve reasonably good model fit?
- **R6.** Does the factor structure yielded from R3 achieve reasonably good model fit?
- **R7.** Does the aforementioned set of eight covariates have statistically significant effects on the measurement model established in R4?
- **R8.** Does the aforementioned set of eight covariates have statistically significant effects on the measurement model established in R6?

Structural Phase

Upon the establishment of valid measurement models in previous phases, the goal of this phase is to examine the relationship between evaluator self-assessed competencies and their practice patterns.

- **R9**. How do evaluator self-assessed competencies and evaluation practice patterns relate to each other?

Participants and Sample

Population

The population of interest in the study includes all practicing evaluators or evaluation practitioners who were members of the American Evaluation Association (AEA) at the time that

the sample was drawn. Practicing evaluators or practitioners are defined as evaluators who have conducted evaluations in any of the five broad capacities – *designing evaluations, implementing evaluations, reporting evaluation results, managing or supervising evaluation projects, and consulting on evaluations*. Since the list of evaluation practitioners is not readily available from the AEA membership directory, the identification of study participants was through a self-selection process by soliciting responses to two filter questions. The first filter question inquires whether in the past ten years participants have conducted evaluations in any of the five capacities. The second filter question inquires about the number of evaluations conducted in the past ten years. The target number of evaluations conducted is 3 or more for practice patterns to emerge. If participants responded "no" to the first question or less than 3 for the second question, they were directed to the end of the survey indicating that they were not the target participants for the study.

Sample Size

Sample size has always been a contentious issue for quantitative research, and a large sample is generally desirable to achieve better representativeness of the population (Raykov & Marcoulides, 2000). Particularly for studies adopting SEM methodological framework, it is crucial to achieve adequate sample size since SEM is a large sample statistical technique (Kelloway, 1998; Tabachinick & Fidell, 2007; Kline, 2016).

Approaches towards sample size in SEM research include the number of observable variables and the ratio of cases/observations to free parameters being estimated. Bentler and Chou (1987) suggested 5 to 10 cases per observable variables when a dataset follows the normal distribution. Jackson (2003) argued for the ratio (n:q) of cases/observations (n) to the number of

free parameters (q) to be estimated in the model should be 10:1 or ideally 20:1 when using Maximum Likelihood (ML) estimator to calculate sample size for SEM studies. Other researchers also provided similar rules of thumb with various prescriptions. A frequently referenced guideline by Comrey and Lee (1992) considers a sample of 50 as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 as excellent. Tabachnick and Fidell (2007) agreed and recommended, "as a general rule of thumb, it is comforting to have at least 300 cases for factor analysis" (p. 613). This study aimed to collect responses from 500 respondents. Since sample size directly relates to the accuracy of the model estimates, a post hoc power analysis has been included to justify sample size adequacy in Chapter IV.

Measures

Scaling

Two measures were adapted from previous research as the data collection tools of the study. The measure for evaluator competencies was adapted from the taxonomy of essential competencies for program evaluators (ECPE) created by King et al. (2001) and further developed by Stevahn et al. (2005). It has hypothesized that the construct of evaluator competency could be represented by 61 competencies in six domains: professional practice, systematic inquiry, situational analysis, project management, reflective practice, and interpersonal competence. For the ECPE measure, participants of the study are requested to complete two ratings for the set of evaluator competencies. Evaluators were first requested to rate their perceived importance of the 61 competencies to the evaluation profession on a 7-point Likert scale (*extremely important*, *very important*, *somewhat important*, *neutral*, *somewhat unimportant*, *very unimportant*, and *extremely unimportant*). And then, participants were requested to perform a self-assessment to

rate their own levels of competencies on a 5-point Likert scale (*expert*, *proficient*, *intermediate*, *advanced beginner*, and *beginner*).

It is important to note that only the perceived importance rating was used in the exploratory phase to explore the dimensionality of the evaluator competency construct. The self-assessed level of competencies rating on the same set of evaluator competencies was used in the confirmatory and structural phases. Because the self-assessed level of competencies is a relatively temporary and contextually driven state, evaluator ratings on the perceived importance of competencies are more appropriate than evaluator ratings on their self-assessed levels of competencies.

The latent variable of evaluator practice (Shadish and Epstein, 1987) were measured from eight aspects of, purposes of the evaluation, influences on decisions to conduct the evaluation, the role evaluators play, the sources of the questions asked in the evaluation, the kind of issues about which data were gathered, the sources of the dependent variables, the method used in evaluation, and actions taken to facilitate use of evaluation results. The final scale for evaluator practice included 74 items on a 5-point Likert scale to gauge the frequency of evaluators engaging in these aspects of their practices. Specific anchors were provided for each of the five levels: *always* (100% of the time), *often* (about 61% to 90% of the time), *sometimes* (31% to 60%) of the time), *rarely* (about 5% to 30% of the time), and *never* (less than 5% of the time). Regarding labeling the points on the scale, Streiner and Norman (2008) suggested that most research does not indicate much difference, but recommend labeling as a good practice. For both measures in the study, all the points on the scales were labeled with clearly stated descriptors.

An empirical consideration in scaling is to increase the variations in responses. When deciding on scale steps for continuous scales, Pett, Lackey, and Sullivan (2003) suggested

including five to seven response options and argued that fewer scale steps would potentially pose restrictions on item variances. They also favor odd numbers over even numbers of response options because odd-numbered scales often provide respondents with a middle/neutral choice without forcing them to make a selection as in even-numbered scales. Nunnally (1978) argued that an additional advantage of a higher number of scale steps is to "enhance scale reliability but with rapidly diminishing returns. As the number of scale steps increases from two to twenty, there is an initial rapid increase in reliability that tends to level off at about seven steps" (p. 149). Therefore, all measures adopt five-level scales as the research suggests, except the rating of perceived importance for evaluator competencies at seven-level. Because the content validity has been established in several studies, all evaluator competencies included in the ECPE were important as demonstrated in the Wilcox (2012) study. By adopting nuanced levels of responses, the researcher aimed to increase the variances of participant responses.

Reliability and Validity

Furr and Bacharach (2014) discussed three implications of reliability when conducting and interpreting behavioral research. The effect of reliability should be taken into consideration when interpreting effect sizes and statistical significance. Researchers should also report and use measures with high reliability when possible.

Although the ECPE scale face validity and content validity were sufficiently established in numerous past studies, most of the past research relied on qualitative methods and small sample sizes. There has been a lack of quantitative research using large sample methodologies to establish construct validity and examine psychometric properties of the ECPE scale, which largely restricts the usability and application of ECPE in broader contexts. The evaluator practice

scale faced a similar challenge. Shadish and Epstein (1987) established the construct validity of evaluator practice as a multidimensional and hierarchical construct but failed to report any psychometric properties. A pilot study was conducted in 2016 to examine the content validity of the evaluator practice scale. The results are presented at the end of this Chapter.

Data Collection

A survey method was employed to collect data for the study. The target population for this study was members of the American Evaluation Association (AEA), which consisted of approximately 7,700 members at that time of the data collection in 2017. Mills and Gay (2016, p. 147) advised that a sample size of 400 would be sufficient for a population size of 5000. This guideline also supports the final number of responses (n = 459) received in the study.

An application was submitted to the AEA to request a random sample of 2,000 evaluators on the AEA mailing list for research purposes in November 2016, with a target response rate of 50% (Babbie, 2007). The AEA approved the application and provided contact information (names and emails) for 2,000 members, randomly drawn from the membership directory. Three rounds of contacts were made by the researcher to maximize the response rate. The initial customized invitation emails described the nature and procedure of the present study and included an active URL link directing participants to a Qualtrics survey. Once providing their consents, participants progressed towards the actual questions, which were updated with the results from the pilot study. Two follow-up contacts were made with two-week intervals.

Several strategies were used to protect participants. First, the study was reviewed and approved by the AEA and Syracuse University Institutional Review Board (IRB). In addition, respondents' identifying information was removed by using Qualtrics' "Anonymizing Responses" function. Next, the online format increased the transparency; and no incentives

increased the data integrity and minimized the risks of coercion. Furthermore, no private personal information was collected, as only general biographical information was collected. Lastly, no identification files were stored, and the access to collected data was restricted exclusively to the researcher of the study.

Data Analysis

Analyses in this study were carried out using Mplus (Version 8.0; Muthén & Muthén, 1998-2017). Data analyses were conducted in three consecutive phases (exploratory phase, confirmatory phase, and structural phase) to address all the research questions. Table 3.1 provides an overview of data sources for each phase.

Table 3.1 Data sources of three consecutive analytical phases

Phase	Data Source
Exploratory	- Perceived importance of competencies rating (ECPE scale) - Evaluator practice scale
Confirmatory	- Perceived importance of competencies rating (ECPE scale) - Self-assessed level of competencies rating (ECPE scale) - Evaluator practice scale
Structural	- Self-assessed level of competencies ratings (ECPE scale) - Evaluator practice scale

Exploratory Phase

A series of exploratory factor analyses (EFA) was conducted to explore the initial dimensionality of evaluator competencies and evaluation practice. Although EFA is often used as an explorative method with no a priori hypotheses identified, it can also be used in a confirmatory manner (Kline, 2016). In structural equation modeling (SEM) framework,

particularly for confirmatory factor analysis (CFA), the line between EFA and CFA has become increasingly blurred.

Table 3.2 EFA steps and analytical details

Steps	Analytical Details
Step 1: Selecting an estimation method	Robust Maximum Likelihood (MLR [a]) estimator was adopted for two reasons: 1) MLR is a full information estimation method that allows goodness-of-fit evaluation; 2) advantage of handling missing data.
Step 2: Selecting the number of factors	Three methods were used jointly to determine the appropriate number of factors extracted: 1) Kaiser Criterion of Eigen-values greater than 1.0; 2) Scree plot; 3) parallel analysis [c].
Step 3: Rotating the factors	Geomin [b] (default in Mplus), an oblique rotation, was adopted.
Step 4: Refining the solution	Factor solutions were adjusted to address issues such as items with strong loadings on multiple factors, items with weak factor loadings, and internal consistency.
Step 5: Interpreting the findings	Relate the results to the research questions.

Note. [a] MLR: "maximum likelihood parameter estimates with standard errors and a chi-square test statistic (when applicable) that are robust to non-normality" (Muthén & Muthén, 1998-2017, p. 668). [b] "The GEOMIN rotation is recommended when factor indicators have substantial loadings on more than one factor" (Muthén & Muthén, 1998-2017, p. 678). [c] "a method that uses random data with the same number of observations and variables as the original data… to determine the optimum number of factors in an exploratory factor analysis. The optimum number of factors is the number of the original data eigenvalues that are larger than the random data eigenvalues" (Muthén & Muthén, 1998-2017, p. 682)

As Kline (2016) and Asparouhov and Muthén (2009) pointed out, once modifications have been made to CFA, the analysis reverts to the EFA framework. In this case, analyses at the exploratory stage aim to confirm the factor structures suggested in previous research. Stevahn and colleagues (2005) conceptualized a six-factor ECPE to represent the six dimensions of essential evaluator competencies. Similarly, for the evaluator practice scale, the first-order factor structures and the second-order factor structure were extracted to compare with or confirm the results in Shadish and Epstein (1987). Pett et al. (2003) and Brown (2015) prescribed a set of

steps for conducting EFAs adopted in the study (see Table 3.2). Alternatively, exploratory structural equation modeling (ESEM) analytical modeling approach (Asparouhov & Muthén, 2009) was attempted in this analytical phase, but the model did not converge due to sample size and the number of parameters to be estimated.

Confirmatory Phase

Once the exploratory phase was completed, the data analysis moved to the confirmatory factor analysis (CFA) phase. Both CFA and EFA are rooted from common factor model with the same purpose to "reproduce the observed relationships among a group of indicators with a small set of latent variables" (Brown, 2015, p. 11). However, CFA differs from EFA in several major ways. Firstly, CFA provides standardized and unstandardized factor solutions, whereas EFA only provides standardized solutions. Unstandardized solutions in CFA make it possible for other applications such as measurement invariance across groups and comparing means among multiple groups. Then, CFA solutions are more parsimonious than EFA solutions because most or all indicators are restricted to one primary factor in CFA. By fixing cross-loadings to zero, factor correlations in CFA are higher than in EFA. With fewer parameters to be estimated, CFA models are much more parsimonious than EFA models. Next, unique variances, e.g., error variances, can be correlated in CFA, but not in EFA. By specifying correlated measurement errors, researchers can achieve better factor solutions in CFA, while EFA tends to extract a common method factor, which often does not have any conceptual support. Finally, in CFA, the model comparison becomes possible so that researchers have greater flexibility in CFA to impose various restrictions on factor loadings, such as changing primary factors for indicators and constraining all factor loadings to be equal. Using the Chi-square difference test, nested models can be compared with Chi-square significance test.

Table 3.3 Six-step processes to conduct CFA and SEM analysis

Steps	Analytical Details
1. Specify the model.	Factor structure results from EFA was tested in CFA analysis for confirmation.
2. Evaluate model identification.	With large numbers of indicators in both the evaluator competency scale (61 items) and the evaluation practice scale (74 items), both CFA models were over-identified.
3. Select the measures and collect, prepare, and screen the data.	Multivariate normality and missing data patterns were inspected.
4. Estimate the model.	The general model fit was assessed as well as the localized fit, such as factor loadings and error covariances.
5. Re-specify the model.	CFA models were respecified (e.g., item reduction, incorporation of error covariances) to improve the model fit.
6. Report the results.	Global/local model-fitting indices and rationales for model modifications (see SEM Model Fit section on p. 78) were reported and interpreted.

Kline (2016) explains that there are two major components in a structural equation model (SEM), a measurement model and a structural model. A measurement model, also known as a CFA model, depicts how many latent factors are included in the model and how indicators relate to the latent factors. On the other hand, a structural model describes how the latent factors relate to each other, directly/indirectly or bi-directional/unidirectional. A well-fitting measurement model should be established before a structural model can be tested. This study followed Kline's six-step process model (see Table 3.3) in testing CFA and SEM models.

Structural Phase

CFA with covariates is also referred to as multiple indicators and multiple causes (MIMIC) models, where relationships of latent factors and covariates are investigated. In the structural phase, MIMIC models were tested to investigate the heterogeneity of mean structure of evaluator competency scale and evaluation practice scale with a set of covariates, such as evaluator years of experience, work settings, gender, and educational background. Wang and Wang (2012) suggested, "when any covariance/correlations between latent variables/factors are replaced with a causal effect, the model becomes a general SEM model, in which a specific factor can be specified to predict other factors or is influenced by other factors" (p. 90). The analytic process followed Kline's model testing steps laid out in Table 3.3.

SEM Model Fit

A series of goodness-of-fit statistics were used to judge the model fit. SEM methodologists (Brown, 2015; Byrne, 2012) summed up model-fitting indices in three major categories: incremental (also comparative), absolute, and predictive or parsimony-corrective indices. According to Hu and Bentler (1999), the main difference between incremental and absolute indices is that incremental fit indices "measure the proportionate improvement in fit by comparing a target model with a more restricted, nested baseline model" (p. 2). Two of the most commonly used incremental indices are the Comparative Fit Index (CFI) and Tucker & Lewis Index (TLI). For absolute fit indices, Chi-square test and Standardized Root Mean Square Residual (SRMR) are the two most popular indices. The parsimony-corrective indices take model fit and model complexity into consideration and are particularly useful when comparing nested models. The Root Mean Square Error of Approximation (RMSEA) and Akaike

Information Index (AIC) are the two widely used parsimony-corrective indices. Since RMSEA provides a confidence interval, it has become one of the most recommended fit indices in SEM research and has been used in simulation studies to estimate statistical power and sample size (MacCallum, Browne & Sugawara, 1996).

There have not been any established rules to determine which fit indices to use since different indices often focus on different aspects of model fitting (Byrne, 2012; Brown, 2015; Jöreskog & Goldberger, 1975; Kline, 2016). The following guidelines were followed in the study to assess model fit.

- *Absolute fit.* Two fit indices are used to evaluate model fit in this category. Chi-square statistic is often used in conjunction with other fit indices due to its sensitivity to large sample sizes. SRMR values close to .08 or below indicate an adequate model fit; and (Hu and Bentler 1999).
- *Comparative/Incremental fit.* CFI and TLI close to .95 or greater indicate good fit (Hu & Bentler, 1999).
- *Parsimony correction*. RMSEA values less than .08 suggest adequate model fit; furthermore, RMSEA values less than .05 indicate good model fit; and models with RMSEA values equal or larger than 1 should be rejected (Browne & Cudeck 1993; Brown, 2015). An additional fit index developed by Browne and Cudeck using RMSEA, "close" fit probability (CFit *p*), tests the hypothesis of whether RMSEA values are less than or equal to .05. Non-significant CFit *p* values indicate an acceptable model fit. However, just as RMSEA, the power of the CFit *p* test can be affected by small sample size and model saturations (fewer *df*s). Akaike Information

Criterion (AIC) is used to compare non-nested models where lower AIC values indicate a better model fit.

Model fitting indices, provided in most SEM packages, offer evaluation criteria to judge whether the hypothesized model fit the sample data adequately. In cases when a model achieves an inadequate fit, modification indices (MIs) can guide the process of improving model fitting. Based on the recommendations of expected parameter change (EPC), minor re-specification of the model may result in improved model-fit. However, methodologists (Byrne, 2012) warned that using MIs to improve model fit should be done with caution as any adjustments in the model should be grounded with theoretical considerations.

Pilot Study – Evaluator Practice Scale

Since the evaluator practice scale was first established over thirty years ago, a pilot test was implemented to establish the scale content validity before the full study took place. Content validity refers to "the extent to which the items on a measure assess the same content or how well the content material was sampled in the measure" (Rubio, Berg-Weger, Tebb, Lee & Rauch, 2003, p. 94). Experts often consider content coverage or content relevance as a more appropriate term (Streiner & Norman, 2008). A content validity study can uncover issues with the measure, provide suggestions on revisions, and ultimately ensure the collection of quality and analyzable data in the full study (Rubio et al., 2003). Moreover, a measure of high content validity leads to more accurate inferences drawn from collected data (Streiner & Norman, 2008).

The pilot study was initiated on August 10th, 2016 and continued through September 30th. The initial email invitations were sent by the researcher of the current study to 100 American Evaluation Association (AEA) members randomly selected from the online AEA membership directory. There were no undeliverable email invitations. The study utilized Qualtrics—a cloud-

based survey system—to collect and manage data. All communications including the initial invitations and two follow-up reminders were implemented via Qualtrics' built-in email system. "Anonymizing Responses" function in Qualtrics was invoked to remove respondents' identifying information such as IP addresses, emails, and names once respondents submit their responses. Follow-up communications can still be sent to those who have not responded.

At the closing, there were 56 responses, of which eight were removed because they did not respond to the two filter questions on evaluator background and years of experience in evaluation. Therefore, the pilot study analyses were based on 48 responses. Since the focus of the pilot study was on the content validity not on evaluators' practice patterns, the minimum of 3 evaluations was not included in the filtering criteria.

All 74 items in the evaluation practice scale were included in the pilot study and were presented on a four-point Likert scale (*highly relevant, relevant, somewhat relevant, and irrelevant*). A textbox was also included for each rating question to encourage respondent elaboration or comments if any items were rated irrelevant.

Several demographic questions were included to provide additional information about respondents. Of the 48 who responded, 41 reported a mean of 12 years of evaluating with a range from 1.25 to 35 years. In addition, 26 (54%) respondents had a Doctorate, 20 (42%) respondents had a Master's degree, and 2 (4%) respondents had a Bachelor's degree. When asked about the number of evaluations that they have conducted in the past ten years, 47 respondents reported a range from 2 to 412 evaluations (Mean = 19.8; Median = 11). Of all respondents, 16 (33%) worked in colleges or universities; 20 (43%) worked for for-profit research, evaluation, or consulting firms; 4 (8%) were students conducting evaluations; 4 (8%) worked for companies in

business and industry; and the last 4 (8%) worked for non-profit organizations. Because the purpose of the pilot study was to examine the content validity, the sample representativeness was not the concern.

According to Rubio and colleagues (2003), item-level content validity index (CVI) could be calculated by dividing the number of "*Highly Relevant*," "*Relevant*," and "*Somewhat Relevant*" ratings by the total number of ratings. A CVI of 80% or higher indicates high content validity. Scale-level CVI was derived by averaging item-level CVIs for that subscale. Overall, 62 (84%) of all items achieved CVI of 80% or higher. 12 items (16%) yielded CVIs below 80% with the lowest of 43% and were subject for revision.

Evaluation Purpose

All item CVIs were presented in Table 3.4, and CVIs above the threshold of 80% indicate a high level of content validity. The subscale of evaluation purpose had nine items and yielded a scale-level CVI of 95%. One respondent commented that "to measure program effects" was probably the most common evaluation purpose. Four respondents rated the purpose of "to influence decisions makers" as somewhat relevant, as one explained, "funders often require evaluation, but then rarely utilize evaluation findings when making decisions about renewal funding." For the two items with lowest CVIs (88% and 87%), "to identify solution to social problems" and "to build social science theory", respondents explained that the low relevancy was due to how evaluators interpret the difference between evaluation and research, and academic interest versus practicality that clients are concerned about.

Table 3.4 Content validity of evaluation purpose subscale

Items	CVI
To measure program effects	100%
To improve program performance	100%
To influence decision makers	100%
To judge program value	96%
To provide information to clients that they can use	96%
To explain how programs work	96%
To identify a solution to social problems	88%
To meet the needs of disadvantage program clients	91%
To build social science theory	87%
Mean	95%

Factors Influencing Decisions to Evaluate

The subscale (9 items in Table 3.5) achieved a scale-level or mean CVI of 87% indicating a high content validity. Of the three items with CVIs below 80%, the lowest one, "clients paid to conduct the evaluations," yielded a low CVI of 64%. Respondents also commented on the difficulty in understanding this item, which was revised as "the evaluations were conducted because clients paid all the expenses" in the full study.

Table 3.5 Content validity: factors influencing decisions to evaluate subscale

Items	CVI
Evaluator's interest in the program being evaluated	92%
Evaluator's interest in basic research questions addressable through evaluation	100%
Evaluator's interest to publish in the area	79%
Managers/supervisors decided to conduct the evaluation	91%
Clients paid to conduct the evaluations	64%
Whether the results of the evaluation would be used to change the program	100%
Whether a good evaluation can be done within budget	92%
Whether the fiscal benefits of the evaluation would exceed its costs	83%
Whether the money to be spent on evaluation could better be spent on something else.	79%
Mean	87%

Evaluator Roles in Evaluation

The eight-item sub-scale achieved a scale-level CVI of 85% (Table 3.6). More than half of the respondents (57%) rated the evaluator role "a servant of the program manager" as irrelevant, hence it had the lowest CVI of 43% (100% minus 57%). Respondents had difficulty understanding the question and particularly disliked the word "servant." However, for the other two items where "servant" was used, respondents thought they were appropriate. The two items with the word "servant" were revised as "an achiever working with the program manager" and "a resource for program stakeholders."

Table 3.6 Content validity: evaluator roles in evaluation subscale

Items	**CVI**
A methodological expert	100%
An educator to my clients	100%
A facilitator of local change	92%
A judge of the program	79%
A servant to the public good	92%
Part of the program team	96%
A servant of the program manager	43%
A servant of program stakeholders	80%
Mean	85%

Influences on Decisions on Evaluation Questions

This subscale with nine items (Table 3.7) yielded a scale-level CVI of 96%, which indicated a high level of content validity. Although the item "pending decisions" received a CVI of 100%, two respondents commented that the item was unclear and difficult to respond. Meanwhile, there was no additional clarification from the original study (Shadish & Epstein, 1987). Based on the results of the exploratory factor analysis, the item was loaded on decision-driven factor and were revised as "pending decisions on the program being evaluated."

Table 3.7 Content validity: Influences on decisions on evaluation questions subscale

Items	CVI
Information needs of supervisors or of the client who paid for the evaluation	100%
Past research/evaluation	100%
Evaluator's own experience about which questions are usually most important	88%
Information needs of program manager	100%
Information needs of program staff	100%
Information needs of program clients	100%
Pending decisions	100%
Social science theory	96%
Pending legislation	88%
Mean	96%

Central Issues about Which Evaluation Data Were Gathered

The subscale with six items (see Table 3.8) yielded a high scale-level CVI of 94%. Despite the high CVI for all items, respondents thought that the question stem and responses were not matching well, and suggested refining the question stem to match the responses better. In the full study, the question stem was updated as "how frequent did you gather data about the following issues in your evaluations."

Table 3.8 Content validity: central issues about which evaluation data were gathered subscale

Items	CVI
Manner in which the program is actually implemented	100%
Changes in service recipients brought on by the program	92%
Explanation of variables that mediate the relationship between program implementation and effects	96%
Number and characteristics of real and potential service recipients	96%
Cost and fiscal benefits of the program	92%
Changes in other people or in other institutions that interact with the program client	92%
Mean	94%

Criteria for Program Effectiveness

The subscale (10 items in Table 3.9) achieved a high scale-level CVI of 97%, which indicated a strong content validity. Nevertheless, one respondent pointed out "there's an underlying assumption here about evaluation purpose. It is not always to determine program

effectiveness." Therefore, the question stem was revised as "when judging program effectiveness, how frequently did you use the following as your evaluation criteria?"

Table 3.9 Content validity: criteria for program effectiveness subscale

Items	CVI
Program goals	100%
Criteria used in past evaluations of the program or similar programs	100%
Criteria in relevant program regulations or legislation	92%
Criteria suggested by relevant social science theory	96%
Criteria selected by program managers	100%
Criteria selected by program staff	100%
Criteria selected by clients who paid for the evaluation	100%
Criteria selected by program clients	100%
Unintended side effects	92%
The needs of the disadvantaged	88%
Mean	97%

Methods Used in Evaluations

The 14 items in this sub-scale encompassed a wide range of evaluation methods, and the subscale yielded a scale-level CVI of 87% (see Table 3.10). Among all methods, "conducting meta-analysis had the lowest CVI of 58%. Respondents suggested that "constructing logic model" is too specific a method. Rather, "program theory" or "theory of change" might be a broader term. Hence, the item was revised as "constructing program theory/theory of change." Additionally, respondents also indicated that "sample survey" is confusing. In this case, "survey" was used in the full study.

Table 3.10 Content validity: methods used in evaluations subscale

Items	CVI
Inspecting program documents/records	100%
Onsite observation	96%
Sample survey	96%
Interviews with stakeholders	100%
Program monitoring (e.g., Management Information system)	100%
Client needs assessment	92%
Constructing logic models	92%
Randomized Experiment	78%
Quasi-experimental design	78%
Participant observation	92%
Achievement tests	96%
Constructing a Meta-evaluation	71%
Casual modeling (e.g., Path analysis)	75%
Conducting meta-analysis	58%
Mean	87%

Activities to Facilitate Evaluation Use

The subscale had nine items and achieved a high scale-level CVI of 93% (see Table 3.11). Only two items yielded CVIs below 80%. “Publish results in books/journals” (79%) and “publicize results in the media” (75%) both are highly related. The reasons that both had lower levels of relevancy, according to the respondents, were because the evaluation findings belong to evaluation clients, who would make publication decisions.

Table 3.11 Content validity: activities to facilitate evaluation use subscale

Items	CVI
Disseminate a written report of results	100%
Translate results into action recommendation	100%
Provide oral briefing to clients	100%
Keep in frequent contact with users during the conduct of the evaluation	91%
Provide feedback to clients during the evaluation	100%
Ask the clients how potential evaluative information would be used to make change	100%
Identify potential users in order to include their questions in the evaluation	96%
Publish results in books/journals	79%
Publicize results in the media	75%
Mean	93%

Overall, the evaluation practice scale has performed well in content validity test. The eight sub-scale CVIs are all above 80% cut-point. The final CVI for evaluation practice scale was 92% indicating high content validity. Revisions to 12 items (with CVIs below 80%) were made to strengthen the instrument content validity. Additionally, question stems for subscales of "issues of which evaluation data were collected" and "criteria for program effectiveness" were updated.

Chapter Summary

This chapter focused on the methodological logistics of the study. First, a description of the population of interest and participants of the study was provided. Then, a brief review of sample size was presented. Next, two measures of evaluator competencies and evaluator practices, adapted from prior research, were introduced. The results of the pilot study on the evaluator practice scale were presented and revealed that the scale achieved reasonably high content validity with a small number of revisions on item wordings. A simple random survey design was used to collect data for the study, and the data collected were analyzed in three—exploratory, confirmatory, and structural phases—to address research questions of each phase. Regarding methodological frameworks, the exploratory phase operated under exploratory factor analysis methodology to explore appropriate factor solutions of the two constructs. On the other hand, confirmatory and structural phases advanced into a more rigorous methodology of testing the goodness-of-fit of measurement and structural models. In the next chapter, analytical results from the three phases are presented in detail.

CHAPTER IV: RESULTS

This research aims to validate factor structures of two critical constructs in the field of program evaluation. This chapter presents the findings by research questions in three phases.

In the exploratory phase, factor structures of evaluator competencies and evaluation practice were tested using exploratory factor analysis (EFA). In the confirmatory phase, the factor structure yielded from the exploratory stage were imposed and confirmed using confirmatory factor analysis (CFA). In the structural phase, measurement invariance was also examined using multiple indicators multiple causes (MIMIC) model, with the purpose of confirming the stability of factor structures across different population/group heterogeneity. In the final structural phase, two structural regression models were applied to examine the relationship between evaluator self-assessed levels of competencies and their practice patterns. At the end of the chapter, results from power analyses using RMSEA are presented to justify the sample size and power of the analyses in the study.

Procedure & Sample Representativeness

A random sample of n = 2,000 participants from the American Evaluation Association (AEA) membership directory (n = 7,700) was provided as of January 19, 2017, as the sampling frame. After the initial contact by the researcher, two additional rounds of reminders were sent to maximize the number of responses. After removing 258 unreachable respondents, the usable sample of AEA members was reduced to n = 1742, and the response rate was 54.6% (n = 952). Since 235 out of 952 respondents opted out of the study, the final response rate was 47.6% (717/1507).

Fowler (2002) argues that the representativeness of the sample depends on how representative the sampling frame is of the population. The respondent traits and characters of the current study are congruent with the trait and characteristics of 2014 AEA member population (n = 7,026) reported by Coryn et al. (2016, p. 162 in Table 1) in gender, the highest level of education, primary work setting, and country settings presented in Table 4.1. For example, the current study had 68.6% (315) female and 27.5% (126) male respondents, compared with 64.56% female and 26.27% male members in AEA population. While 41.52% and 41.94% of AEA population had Doctorate and Master's degrees, 58.6% (269) and 35.1% (161) respondents in the study received Doctorate and Master's degrees. Additionally, 30.9% (142) respondents worked in college/university, compared with 30.84% in AEA population. This consistency suggested the study sample sufficiently represented the population.

Additional data cleaning procedure removed those respondents who did not consent for the use of their data (n = 17), those who have not conducted evaluations in the past ten years (n = 5), those who have conducted fewer than 3 evaluations in any capacity (n = 18), and those with missing data on more than 5% (n = 162) of all variables. Using Mahalanobis distance test (Tabachnick & Fidell, 2007), 56 cases were identified as outliers and hence removed from the analytical dataset. Consequently, the final usable responses were from 459 (30.5%) respondents. In addition to the post hoc power analysis presented at the end of the chapter, recommendations from Comrey and Lee (1992) and Tabachnick and Fidell (2007) suggested that the study acquired a reasonable sample size suitable for SEM analytical framework.

Table 4.1 Demographic and professional background information

	Study Sample (Total N = 459)	%	AEA Member Population (N = 7,026) (Coryn et al., 2016)
Gender			
Female	315	68.6%	64.56%
Male	126	27.5%	26.27%
Missing	18	3.9%	-
Professional Identity			
Evaluator	374	81.5%	-
Other	84	18.3%	-
Missing	1	0.2%	-
Primary Affiliation			
American Evaluation Association	309	67.3%	-
Other	149	32.5%	-
Missing	1	0.2%	-
Highest Degree			
Doctorate	269	58.6%	41.52%
Master's	161	35.1%	41.94%
Other	4	0.9%	-
Bachelor	8	1.7%	5.61%
Missing	17	3.7%	-
Field of Highest Degree			
Education	108	23.5%	-
Psychology	56	12.2%	-
Public Policy/Administration	54	11.8%	-
Health/Public Health	50	10.9%	-
Evaluation	41	8.9%	-
Sociology	32	7.0%	-
Business & Economics	18	3.9%	-
Social Work	12	2.6%	-
Other	72	15.7%	-
Missing	16	3.5%	-
Work Setting			
College/University	142	30.9%	30.84%
Independent Consulting	82	17.9%	-
Non-Profit Organization	76	16.6%	21.02%
For-Profit Company	60	13.1%	-
Federal Government	27	5.9%	5.31%
Local/State Government	22	4.8%	5.27%
Business & Industry	11	2.4%	
Student in Evaluation	6	1.3%	-
Other	18	3.9%	-
Missing	15	3.3%	-

Evaluation Background			
Reside in USA & USA Programs	353	76.9%	-
Reside in USA & International Programs	34	7.4%	-
Sub-total residing in USA	387	84.3%	80.03%
Reside outside USA & USA Programs	39	8.5%	-
Reside Outside USA & International	7	1.5%	-
Sub-total residing outside USA	46	10%	14.86%
Other	11	2.4%	-
Missing	15	3.3%	-
Percent of Current Evaluation Work			
≤ 25%	75	16.3%	-
30% - 50%	103	22.4%	-
55% - 75%	65	14.2%	-
80% - 95%	107	23.3%	-
98% - 100%	104	22.7%	-
Missing	5	1.1%	-
Number of Evaluations Conducted			
3 - 10	156	34.0%	-
11 - 20	112	24.4%	-
21-50	134	29.2%	-
51 - 100	35	7.6%	-
> 100 (Max = 1000)	22	4.8%	-
Evaluation Experiences (Years)			
≤ 3	39	8.5%	-
4 - 10	156	34.0%	-
11 - 20	160	34.9%	-
21 - 30	58	12.6%	-
> 30 (Max = 51)	40	8.7%	-
Missing	6	1.3%	-

Data Recode, Missing Data & Multivariate Normality

The ECPE importance of competencies was initially on a 7-point Likert scale (*7 = extremely important; 6 = very important; 5 = somewhat important; 4 = neutral; 3 = somewhat unimportant; 2 = very unimportant; 1 = extremely unimportant*). The univariate descriptive analysis suggested heavy skewness and kurtosis. Further examination of the data revealed that a small number/percentage of respondents rated unimportant (1-3), which is consistent with the expectations that the majority of the competencies included in the ECPE were somewhat

important. For this reason, the importance ratings were recoded to a 5-point Likert scale by combining the three unimportant rating categories (*3 = somewhat unimportant; 2 = very unimportant; 1 = extremely unimportant*). The recoding significantly reduced the skewness and kurtosis. However, the data still did not achieve univariate normality and consequently multivariate normality.

Other than the initial data cleaning procedure of removing cases with more than 5% of missing data on all variables, another strategy adopted to deal with missing data was the use of Mplus, which has the capacity of handling up to 50% of missing data (Muthén & Muthén, 1980). These two measures ensured the accuracy of the analyses. Multivariate normality, a critical assumption for SEM analyses, can be difficult to detect. Byrne (2012) argued that the violation of the multivariate normality assumption would result in inaccurate estimates. MLR estimator (Maximum Likelihood parameter estimates with robust standard errors) in Mplus, an extension from MLM estimator (Robust Maximum Likelihood Estimator) introduced by Satorra and Bentler (1988). MLM uses the listwise deletion to exclude missing data. MLR, similar to the full-information maximum likelihood (FIML), imputes rather than deletes the missing information. It also incorporates a scaling correction factor to adjust for the non-normality of categorical Likert data.

Descriptive Statistics

Item means and standard deviations were examined to provide initial information before moving into inferential statistical analyses. Since the ECPE perceived importance ratings were recoded from a 7-point Likert scale to a 5-point Likert scale, the means and standard deviations presented reflected this change.

ECPE Perceived Importance Rating. Evaluators rated the competency of "acts ethically and strive for integrity and honesty in conducting evaluation" (M = 4.91) the most important, followed by two competencies of "uses verbal/listening communication skills" (M = 4.81) and "respect clients, respondents, program participants, and other stakeholders" (M = 4.80). On the other hand, evaluators rated the competency of "conducts meta-evaluation" (M = 2.62) least important, followed by the competency of "contribute to the knowledge base of evaluation" (M = 3.34). Additionally, respondents also rated 12 other competencies less important (M < 4.0), shown in Table 4.2.

For the self-assessed levels of competencies, evaluators rated highest levels of competencies in the two same competencies, "acts ethically and strives for integrity and honesty in conducting evaluations" (M = 4.55) and "respect clients, respondents, program participants, and other stakeholders" (M = 4.46). Similarly, the same two competencies "conducts meta-evaluation" (M = 2.81) and "contribute to the knowledge base of evaluation" (M = 3.53) were also rated the lowest by evaluators. Amongst the 12 competencies that were rated as less important (M < 4.0), evaluators' self-assessed mean levels on ten competencies were also lower than 4.0. Overall, the results were consistent with the findings in Galport and Azzam (2017), where these low self-assessed competencies were identified as gaps in training and professional development.

Table 4.2 Descriptive statistics for ECPE scale

ECPE Scale	Perceived Importance		Self-assessed Level of Competencies	
	Mean	**SD**	**Mean**	**SD**
Professional Practice:				
Applies professional evaluation standards	4.44	0.799	4.13	0.779
Acts ethically and strives for integrity and honesty in conducting evaluations	4.91	0.354	4.55	0.583
Conveys personal evaluation approaches and skills to potential clients	3.86*	0.971	4.13	0.805
Respects clients, respondents, program participants, and other stakeholders	4.80	0.501	4.46	0.672
Considers the general and public welfare in evaluation practice	4.08	0.920	3.97*	0.789
Contribute to the knowledge base of evaluation	3.34*	1.093	3.53*	0.976
Systematic Inquiry:				
Understands the knowledge base of evaluation (terms, concepts, theories, assumptions)	3.72*	1.131	3.89*	0.869
Knowledgeable about quantitative methods	4.15	0.799	3.82*	0.888
Knowledge about qualitative methods	4.21	0.742	4.00	0.825
Knowledge about mixed methods	4.32	0.707	3.92	0.822
Conducts literature reviews	3.79*	0.999	4.23	0.831
Specifies program theory	3.81*	1.117	3.97*	0.972
Frames evaluation questions	4.75	0.515	4.37	0.743
Develops evaluation design	4.71	0.561	4.23	0.803
Identifies data sources	4.45	0.734	4.29	0.761
Collects data	4.70	0.593	4.39	0.684
Assesses validity of data	4.42	0.784	3.93*	0.919
Assesses reliability of data	4.35	0.834	3.90*	0.931
Analyze data	4.71	0.563	4.19	0.743
Interprets data	4.79	0.471	4.33	0.685
Makes judgments	3.89*	1.045	4.02	0.823
Develops recommendations	4.25	0.904	4.12	0.786
Provides rationales for decisions throughout the evaluation	4.26	0.790	4.15	0.774
Reports evaluation procedures and results	4.41	0.759	4.34	0.673
Notes strengths and limitations of the evaluation	4.36	0.786	4.19	0.729
Conducts meta-evaluation	2.62*	1.147	2.81*	1.176
Situational Analysis:				
Describes the program	4.46	0.765	4.42	0.736
Determines program evaluability	4.05	1.041	3.91*	1.037
Identifies the interests of relevant stakeholders	4.39	0.768	4.19	0.840
Serves the information needs of intended users	4.52	0.672	4.18	0.866
Addresses conflicts	4.05	0.937	3.72*	1.051
Examines the organizational context of the evaluation	4.12	0.886	3.94*	0.969
Analyzes the political considerations relevant to the evaluation	3.81*	1.003	3.60*	1.141
Attends to issues of evaluation use	4.11	0.921	3.93*	0.988
Attends to issues of organizational change	3.64*	1.007	3.63*	1.092
Respects the uniqueness of the evaluation site and client	4.22	0.927	4.13	0.908
Remains open to input from others	4.49	0.706	4.32	0.788
Modifies the study as needed	4.32	0.786	4.22	0.867
Project Management:				
Responds to requests for proposals	3.59*	1.347	3.76*	1.106
Negotiates with clients before the evaluation begins	4.08	1.074	3.77*	1.076
Writes formal agreements	3.80*	1.221	3.61*	1.121
Communicates with clients throughout the evaluation process	4.58	0.711	4.30	0.797

ECPE Scale	Importance		Self-assessed Level of Competencies	
	Mean	**SD**	**Mean**	**SD**
Budgets an evaluation	4.24	1.046	3.67*	1.129
Justifies cost given information needs	3.94*	1.042	3.64*	1.105
Identifies needs resources for evaluation, such as information, expertise, personnel, instruments	4.47	0.679	4.00	0.937
Uses appropriate technology	4.01	0.851	3.78*	0.880
Supervises others involved in conducting the evaluation	4.08	1.033	4.00	0.987
Trains others involved in conducting the evaluation	3.92*	1.050	4.00	0.948
Conducts the evaluation in a nondisruptive manner	4.26	0.885	4.10	0.838
Presents work in a timely manner	4.52	0.655	4.30	0.785
Reflective Practice:				
Aware of self as an evaluator (knowledge, skills, dispositions)	4.64	0.596	4.27	0.727
Reflects on personal evaluation practice (competencies and areas for growth)	4.28	0.808	4.03	0.749
Pursues professional development in evaluation	4.12	0.899	3.96*	0.809
Pursues professional development in relevant content areas	3.92*	0.922	3.86*	0.809
Builds professional relationships to enhance evaluation practice	4.01	0.934	3.71*	0.972
Interpersonal Competence:				
Uses written communication skills	4.78	0.440	4.45	0.679
Uses verbal/listening communication skills	4.81	0.437	4.40	0.674
Uses negotiation skills	4.11	0.912	3.80*	0.944
Uses conflict resolution skills	4.14	0.902	3.76*	0.951
Facilitates constructive interpersonal interaction (teamwork, group facilitation, processing)	4.35	0.768	4.10	0.810
Demonstrates cross-cultural competence	4.34	0.913	3.85*	0.882

Note. * Indicates M < 4.0.

Evaluator Practice Scale. Evaluators reported how they conducted evaluations in eight different aspects by completing the evaluator practice scale on a 5-point Likert frequency scale (5 = *Always* and 1 = *Never*). The most frequently and least frequently reported practice patterns by eight practice domains were included in Table 4.3. For example, the most frequently reported evaluation purpose was “to provide information to clients that they can use” (M = 4.43), and the least frequently reported evaluation purpose was “to build social science theory” (M = 2.16). Additionally, evaluators reported that they most frequently assumed the role as a methodological expert; and least frequently assumed the role of an achiever working with the program manager.

Table 4.3 Evaluator practice scale: most and least frequently reported practice items

Domains	Items	Mean	SD
Evaluation purpose:			
Most Frequent:	To provide information to clients that they can use	4.43	.880
Least Frequent:	To build social science theory	2.16	.995
Factors influencing decisions to evaluate:			
Most Frequent:	Evaluator's interest in the program being evaluated	3.51	1.072
Least Frequent:	Whether the money to be spent on evaluation could better be spent on something else	2.11	1.084
Evaluator roles:			
Most Frequent:	A methodological expert	3.98	.942
Least Frequent:	An achiever working with the program manager	2.85	1.265
Reported sources of questions & issues:			
Most Frequent:	Information needs of the client who paid for the evaluation	4.40	.980
Least Frequent:	Pending legislation	2.26	1.058
Central Issues on which evaluation data were collected:			
Most Frequent:	Manner in which the program is actually implemented	4.39	.742
Least Frequent:	Cost and fiscal benefits of the program	2.78	1.097
Dependent variables for program effectiveness:			
Most Frequent:	Program goals	4.64	.592
Least Frequent:	Criteria selected by program clients	3.05	1.131
Methods:			
Most Frequent:	Interviews with stakeholders	4.19	.786
Least Frequent:	Conducting meta-analysis/ Randomized Experiment	1.75/ 1.75	.899/ .969
Activities to facilitate use:			
Most Frequent:	Disseminate a written report of results	4.76	.535
Least Frequent:	Publish results in books or journals	2.35	1.072

The Roadmap to Research Questions & Analyses

A large number of analyses were conducted to address the nine research questions in three analytical phases. To be specific, the exploratory phase focused on three research questions

with purposes of exploring factor structures of the two constructs; analyses in the confirmatory phase concentrated on answers to five research questions aiming to confirm the factor structures derived from the exploratory phase; and the confirmed measurement structures were then brought into the structural phase to address the final research question on the interaction between the two constructs. Table 4.4 takes stock of all analyses carried out in all three phases.

Table 4.4 Roadmap to research questions and analyses

Phases	Research Question	Type of Analysis	Analysis Description
Exploratory Phase	R1	EFA	EFA aimed to explore the factor structure of the ECPE scale, and a five-factor model emerged.
	R2	EFA	Eight separate first-order EFA models were tested to explore the factor structures of eight evaluation practice subscales.
	R3	EFA	Two second-order EFA models in two approaches: ▪ Mean score by factor approach, with factor score indeterminacy issue discussed.
Confirmatory Phase	R4	CFA	Two CFA models: ▪ Five-factor model for the perceived importance of competencies, ▪ Five-factor model for self-assessed level of competencies.
	R5	CFA	Eight CFA models for evaluator practice subscales.
	R6	CFA	Two second-order CFA models were tested for evaluator practice scale: ▪ item-level model, ▪ composite mean score approach.
	R7	MIMIC	Five-factor perceived importance of evaluator competencies measurement model.
	R8	MIMIC	Four-factor second-order evaluator practice patterns measurement model.
Structural Phase	R9	Structural Regression Model	Two models testing two hypotheses: ▪ Whether self-assessed evaluator competencies as predictors affect their practice patterns; ▪ Whether evaluator practice patterns as predictors impact their self-assessed level of competencies.

Research Questions and Analytical Results

R1. Does the factor structure of the ECPE scale conform to the 6-factor structure conceptualized by Stevahn et al. (2005)?

EFA with MLR estimator and Geomin rotation was carried out using Mplus 8.0 to examine ECPE factor structure. Meyers, Gamst, and Guarino (2014) identified six aspects that can facilitate how many factors should be extracted: a) theoretical and empirical milieu; b) screen plot of the eigenvalues; c) amount of variance accounted for by different solutions; d) number of variables used to represent factors; e) strength of the coefficients; and f) reasonableness of the factor interpretation. Additionally, parallel analyses (PA) (Horn, 1965) were carried out for all EFAs to provide additional confirmation as PA is one of the most accurate methods to determine the number of actors to retain (Zwick & Velicer, 1986).

Before conducting EFA analysis, the correlation matrix was inspected for singularity or extreme multicollinearity (Tabachnick & Fidell, 2007). The high correlation of .84 between item 17 “Assess validity of data” and item 18 “Assess reliability of data” might be problematic. Therefore, item 17 was removed on the ground that the calculation of validity is often based on reliability statistics such as discriminant validity. The initial extraction based on Kaiser criteria (Eigenvalues > 1) produced 15 factors that were uninterpretable. Additional factor structures were tested and revealed that the most interpretable factor structures ranged from five to ten factors, with the five-factor solution being the most interpretable. Indicators with loadings lower than .30 and indicators with strong cross-loadings above .30 were eliminated (Hair, Anderson, Tatham & Black, 1995; Costello & Osborne, 2005; Tabachnick & Fidell, 2007). The final five-

factor EFA model had 44 indicators and achieved a moderate fit: χ^2 (736) = 1651.139, CFI = .838, TLI = .820, RMSEA = .052, CFit *p*=.155, SRMR = .041.

Parallel analysis (p. 682, Muthén & Muthén, 1998-2017) with 95th percentile and 1000 iterations were carried out in Mplus and confirmed the five-factor solution. As shown in the following table, the eigenvalue for the first factor in the actual data is 11.396, while it is 1.668 for simulated data. Up until the fifth factor, all the eigenvalues of the sample data are larger than those of simulated data. From the sixth factor on, all the eigenvalues for the actual data are smaller than those of the simulated data. This shift suggests that five factors should be extracted, and explain 46.18% of the total variance. The percentage of variance explained is calculated by dividing the sum of the five Eigenvalues (19.00208) by 44 (the total number of factors extracted).

Table 4.5 Parallel analysis Eigenvalues of actual data and simulated data

Factor	Eigenvalues of Sample Data	Eigenvalues of Simulated Data	Variance Explained
1	11.39614	1.66799	25.90%
2	2.87372	1.59661	6.53%
3	2.42462	1.54266	5.51%
4	1.98623	1.49537	4.51%
5[a]	1.63912	1.45433	3.73%
6	1.40708	1.41523	3.20%
7	1.28904	1.38057	2.93%
8	1.17219	1.34747	2.66%
9	1.09879	1.31494	2.50%
10	1.03846	1.28422	2.36%

Note. [a]There are five optimum factors whose original eigenvalues that are larger than the simulated data eigenvalues (Muthén & Muthén, 1998-2017, p. 682)

While Table 4.5 only exhibits eigenvalues up to ten factors, Figure 4.1 provides a comprehensive view of all eigenvalues produced.

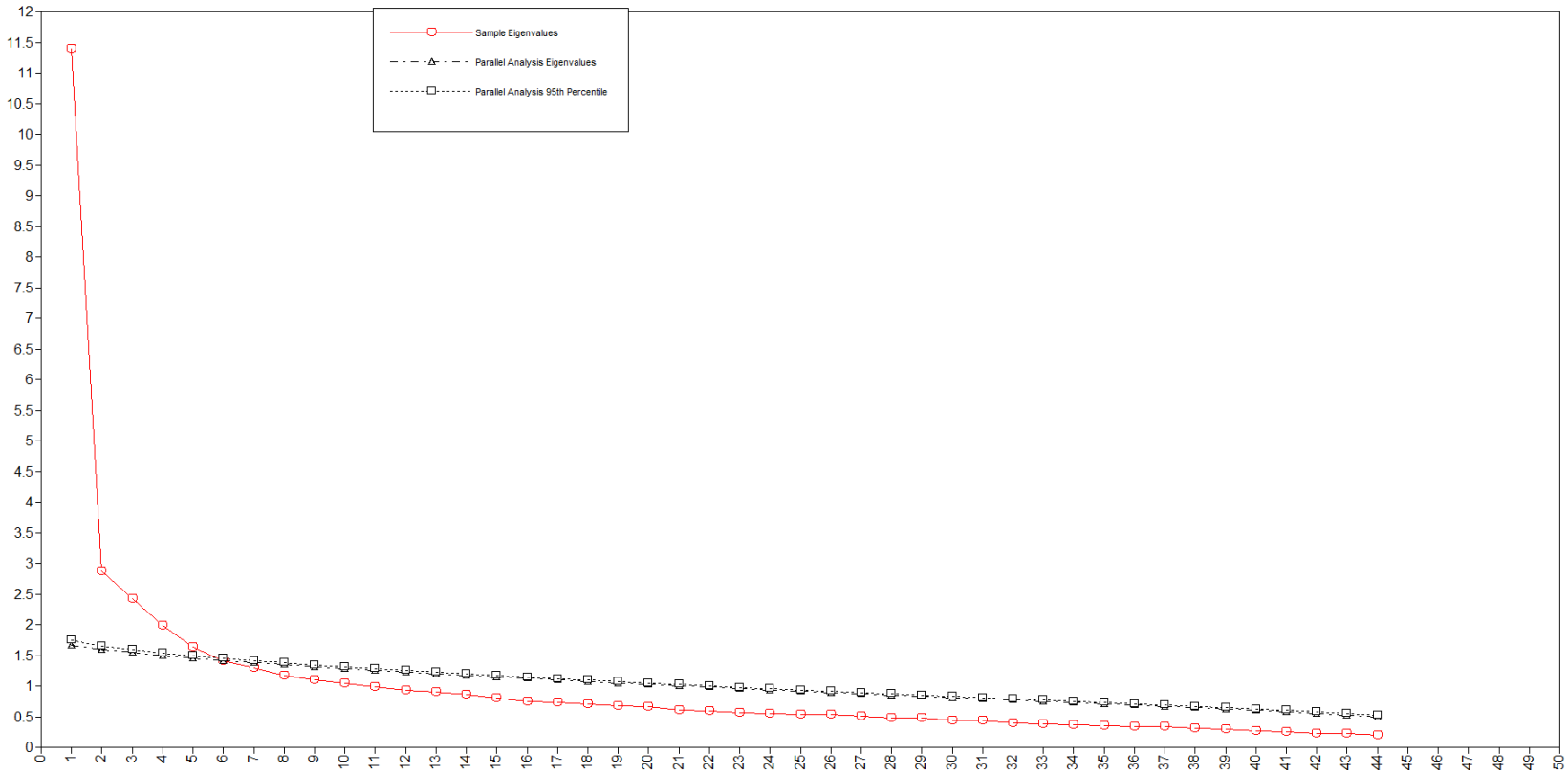


Figure 4.1 Parallel analysis scree plot of the ECPE 5-factor structure

Researchers in Psychology often use .30 or .40 as the cutoff criterion for factor loadings (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Kahn, 2006; Preacher, Zhang, Kim, & Mels, 2013). Factor loadings lower than .30 criterion suggest that roughly less than 10% of the total variance can be explained by a factor. In the current study, items with factor loadings lower than .30 and items with equally strong factor loadings on multiple factors were eliminated. Item-level communalities were also taken into consideration in the item reduction process. Items with low communalities suggest low item correlations with all other items and the target factors do not sufficiently explain item variances.

Factor 1 had 12 indicators with factor loadings ranging from .464 to .612. As indicated in Table 4.5, Factor 1 explains the largest portion and 29.08% of the total variance. The item contents centered focused heavily on competencies that were crucial in conducting evaluations, such as determining program evaluability, specifying program theory, and attending to issues of evaluation use. Therefore, the first factor was named *evaluative practice*.

Factor 2 had 12 indicators with factor loadings ranging from .316 to .715. Items for this factor centered around competencies that were general and appeared to be as critical for all evaluators, such as competencies using verbal/listening communication skills, presenting work in a timely manner, and remaining open to input from others. Hence, the second factor was named *meta-competencies.*

Factor 3 had ten indicators with factor loading ranges from .338 to .822. Items for this factor focused on knowledge and skills in research methods and data analysis. All items in this factor were from systematic inquiry dimension. For this reason, this factor was names as *evaluation knowledge base*.

Factor 4 was named project management since all six indicators were from the *project management* dimension of ECPE. Similarly, ***Factor 5*** had four indicators that all focused on professional development and was named as such.

Table 4.6 EFA results: range of factor loadings and number of indicators per factor

Sub-scale	Ranges of Factor Loadings	# of items
Evaluative Practice	.464 – .612	12
Meta-competencies	.316 – .715	12
Evaluation Knowledge Base	.338 – .822	10
Project Management	.343 – .875	6
Professional Development	.343 – .818	4
	Total:	44

The examination of factor correlations revealed that *evaluative practice, meta-competencies,* and *evaluation knowledge base* factors had low to moderate correlations with other factors (.183 - .622). *Project management and professional development* factors had the lowest correlation (r = .183).

Table 4.7 The ECPE factor correlation matrix

	Factor	1	2	3	4	5
1	Evaluative Practice	1.000				
2	Meta-competencies	.622	1.000			
3	Evaluation Knowledge Base	.567	.505	1.000		
4	Project Management	.468	.398	.386	1.000	
5	Professional Development	.499	.458	.351	.183	1.000

The ECPE researchers hypothesized six dimensions underlying the 61 competencies: professional practice, systematic inquiry, situational analysis, project management, reflective practice, and interpersonal competence. While in the EFA analysis, only five factors were extracted. The discrepancy could be attributed mainly to the methodological differences. Previous ECPE research was conducted with a focus on content validity with small sample sizes

and mostly qualitative. Therefore, the cross-dimensional item content overlapping was methodologically challenging to detect. In factor analytical frameworks, the ECPE dimensionality became more distinct as items with low item-total correlations, subsequently low communalities, and factor loadings were removed. Furthermore, the elimination of items with high cross-loadings assisted in the interpretation of the factor structure substantially.

Table 4.8 The ECPE five-factor structure: factor loadings and item means/standard deviations

Items	F1	F2	F3	F4	F5	Mean/SD
Determines program evaluability	.612					4.05/1.041
Analyzes the political considerations relevant to the evaluation	.587					3.81/1.003
Addresses conflicts	.581					4.05/0.937
Examines the organizational context of the evaluation	.578					4.12/0.886
Specifies program theory	.568					3.81/1.117
Conducts literature reviews	.558					3.79/0.999
Attends to issues of organizational change	.540					3.64/1.007
Attends to issues of evaluation use	.536					4.11/0.921
Uses conflict resolution skills	.517					4.14/0.902
Conducts meta-evaluation	.504					2.62/1.147
Frames evaluation questions	.467					4.75/0.515
Uses negotiation skills	.464					4.11/0.912
Respects clients, respondents, program participants, and other stakeholders		.729				4.80/0.501
Uses verbal/listening communication skills		.715				4.81/0.437
Remains open to input from others		.562				4.49/0.706
Acts ethically and strives for integrity and honesty in conducting evaluations		.548				4.91/0.354
Uses written communication skills		.547				4.78/0.44
Aware of self as an evaluator (knowledge, skills, dispositions)		.504				4.64/0.596
Presents work in a timely manner		.439				4.52/0.655
Facilitates constructive interpersonal interaction (teamwork, group facilitation, processing)		.412				4.35/0.768
Demonstrates cross-cultural competence		.391				4.34/0.913
Conducts the evaluation in a nondisruptive manner		.339				4.26/0.885
Modifies the study as needed		.328				4.32/0.786
Respects the uniqueness of the evaluation site and client		.316				4.22/0.927
Analyze data			.822			4.71/0.563
Interprets data			.736			4.79/0.471
Knowledgeable about quantitative methods			.579			4.15/0.799
Assesses reliability of data			.503			4.35/0.834

Knowledgeable about mixed methods	.503			4.32/0.707
Collects data	.498			4.70/0.593
Reports evaluation procedures and results	.492			4.41/0.759
Develops evaluation design	.385			4.71/0.561
Knowledgeable about qualitative methods	.355			4.21/0.742
Notes strengths and limitations of the evaluation	.338			4.36/0.786
Writes formal agreements		.875		3.80/1.221
Budgets an evaluation		.810		4.24/1.046
Justifies cost given information needs		.788		3.94/1.042
Negotiates with clients before the evaluation begins		.781		4.08/1.074
Responds to requests for proposals		.704		3.59/1.347
Supervises others involved in conducting the evaluation		.343		4.08/1.033
Pursues professional development in evaluation			.818	4.12/0.899
Pursues professional development in relevant content areas			.686	3.92/0.922
Reflects on personal evaluation practice			.533	4.28/0.808
Builds professional relationships to enhance evaluation practice			.343	4.01/0.934

Note. F1 = Evaluative Practice; F2 = Meta-competencies; F3 = Evaluation Knowledge Base; F4 = Project Management; F5 = Professional Development.

R2. Does the factor structure conform to 22 first-order factor structure yielded in 8 sub-domains of evaluation practice in Shadish and Epstein (1987) study?

Eight separate first-order EFAs were carried out in Mplus 8.0 with MLR estimator and Geomin rotation, the default oblique rotation in Mplus, to examine the evaluator practice factor structure. Parallel analysis results were used to guide the decision on the number of factors to be extracted. In addition, correlation matrix determinants, Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) and Bartlett's test of sphericity were also examined to confirm whether factor analyses were appropriate (Pett et al., 2003). Consistent with criteria suggested by Pett et al. (2003) and Tabachnick and Fidell (2007), all determinants were larger than zero; all KMOs were larger than .6 ranging from .601 to . 827 and consequently adequate for factor analysis; and all Bartlett's Tests of Sphericity were significant, rejecting identity matrix hypotheses. The EFA results for the eight subscales of evaluation practices are presented in the following sections. The convention for naming the factors closely followed the tradition of Shadish and Epstein (1987) for similar factors.

Evaluation Purposes

Competing factor structures were also fitted to the data. The model-fitting indices suggested that two factors should be retained and was confirmed by the results of the parallel analyses. Item 5 "to provide information to clients that they can use" had equally weak loadings (.159) on both factors, and hence was eliminated.

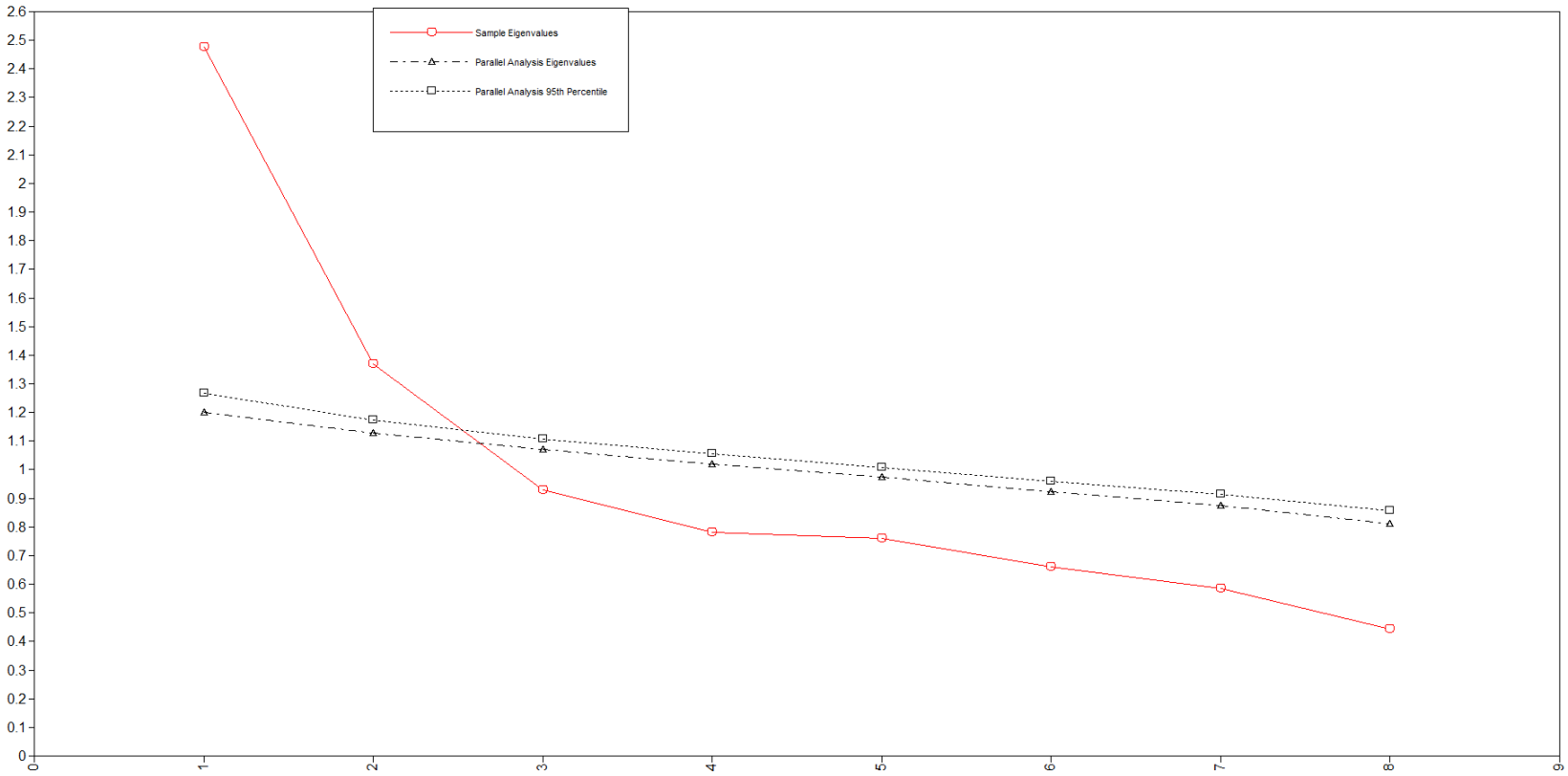


Figure 4.2 Parallel analysis plot for evaluation purposes subscale

All other items loaded strongly on their respective factors without any significant cross-loadings. The model-fitting indices indicated good fit: χ^2 (13) = 30.444, CFI = .959, TLI = .911, RMSEA = .054, CFit *p* = .359, SRMR = .029, and all factor loadings were statistically significant at .05 level. *Factor 1* had three of the four indicators that mainly focused on evaluands or program- driven purposes, hence was named *program-focused purposes*. *Factor 2* had four indicators with content areas focusing on broad societal and social science purposes and was named *scientific idealistic*. The two factors had a low but statistically significant correlation (*r* = .364).

Table 4.9 Evaluation purposes factor structure and loadings

Items	***Mean/SD***	***Factor 1***	***Factor 2***
To judge program value	3.78/.953	.577	
To measure program effects	4.34/.762	.531	
To improve program performance	4.33/.712	.468	
To influence decision makers	3.91/.875	.360	
To identify solutions to social problems	2.98/1.112		.830
To meet the needs of disadvantage program clients	3.27/1.172		.568
To build social science theory	2.16/.995		.523
To explain how programs work	3.47/1.039		.471
To provide information to clients that they can use	4.43/.880	dropped	

Note. Factor 1 = Program-focused purpose; Factor 2 = Scientific idealistic purpose.

Factors Influencing Decisions to Evaluate

After eliminating three items due to low loadings and cross-loadings, two factors were retained and was confirmed by the results of the parallel analysis. The model fitting indices indicated a good fit: $\chi^2(4)$ = 12.808, *p* = .0123, CFI = .981, TLI = .928, RMSEA = .069, CFit *p* = .186, SRMR = .021, and all factor loadings were statistically significant at .05 level. The content area for items loaded on *Factor 1* reflected influences rooted in *basic scientific interests*, while the indicators for *Factor 2* represented the *cost/benefit of evaluation*. The two factors had a low but statistically significant correlation (*r* = .253).

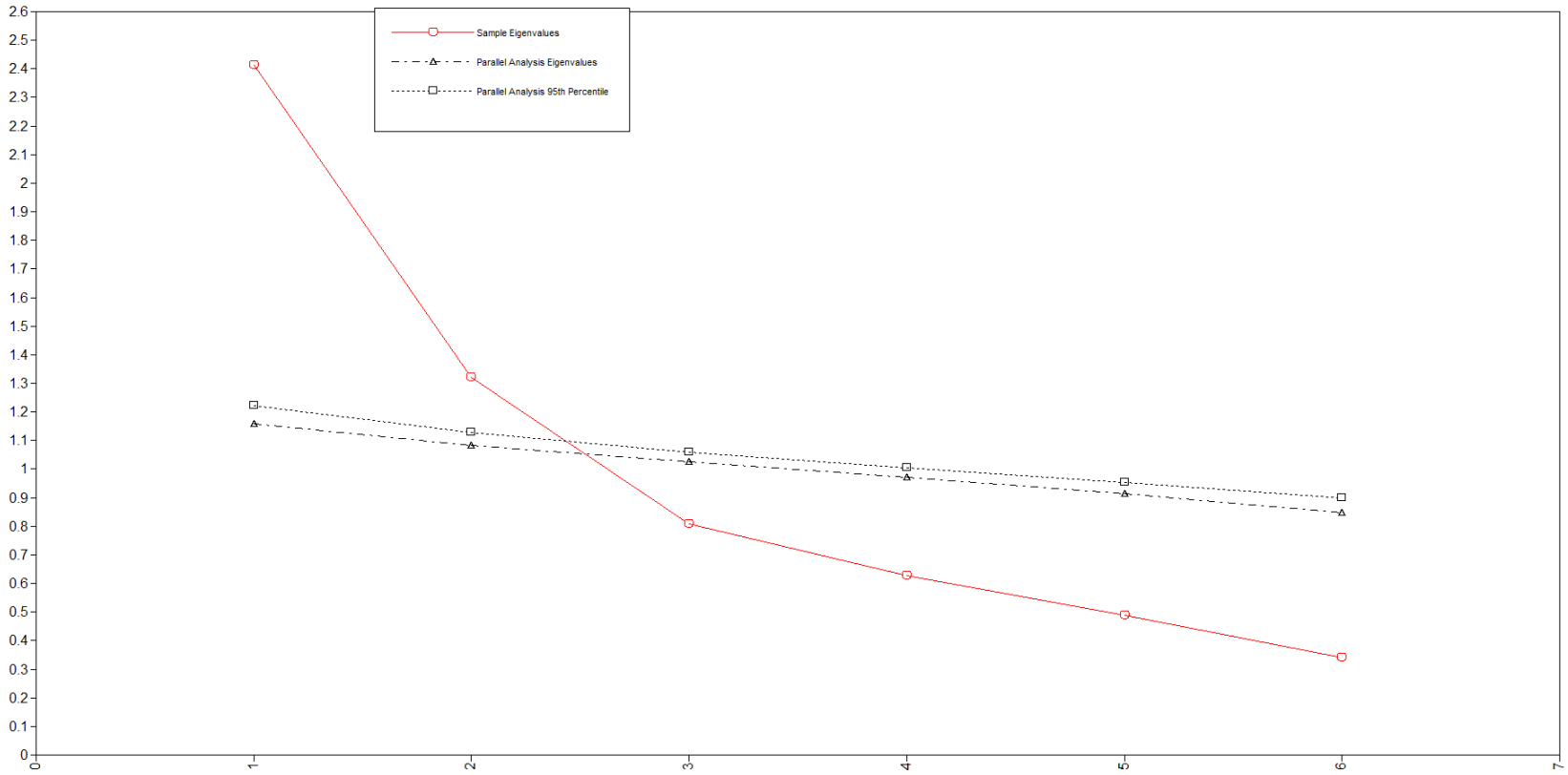


Figure 4.3 Parallel analysis plot for factors influencing decisions to evaluate subscale

Table 4.10 Factors influencing decisions to evaluate loadings

Items	***Mean/SD***	***Factor 1***	***Factor 2***
Evaluator's interest in basic research questions addressable through evaluation	3.17/1.153	.991	
Evaluator's interest in the program being evaluated	3.51/1.072	.617	
Evaluator's interest in publishing in this area	2.25/1.112	.364	
Whether the fiscal benefits of the evaluation would exceed its costs	2.54/1.203		.792
Whether the money to be spent on evaluation could better be spent on something else.	2.11/1.084		.611
Whether a good evaluation can be done within budget	3.42/1.220		.515
Whether it can be shown how the results of evaluation would be used to change the program	3.49/1.150	dropped	
Whether it can be shown how the results of evaluation would be used to change the program	3.49/1.150	dropped	
The evaluations were conducted because clients paid all the expenses	3.17/1.455	dropped	

Note. Factor 1 = Basic scientific interests; Factor 2 = Cost/benefit of evaluation

Evaluator Roles in Evaluation

After three items were eliminated due to low communalities and low loadings (< .30), two factors were extracted, and the factor solution was confirmed by parallel analysis. The two-factor model had a good fit with $\chi^2(8) = 23.163$, CFI = .957, TLI = .886, RMSEA = .064, CFit *p* = .235, SRMR = .031.

All indicators had loaded significantly on their respective factors with loading ranges from .50 to .804. While *Factor 1* has three indicators that focus on evaluator roles as a *change agent* at the local or public level, *Factor 2* has two indicators focus on *team-oriented evaluator roles*. The two factors have a low correlation of .318.

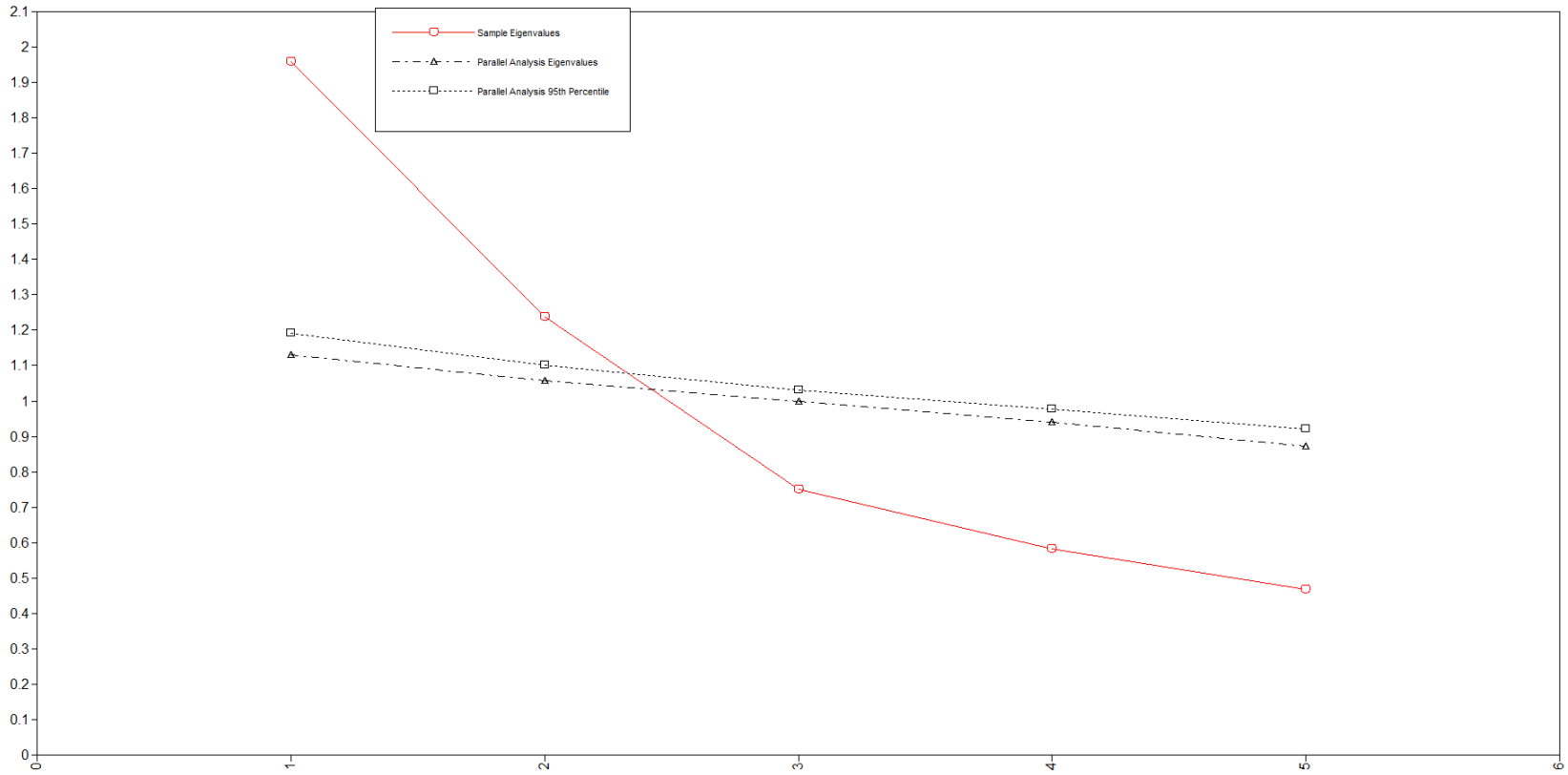


Figure 4.4 Parallel analysis plot for evaluator roles subscale

Table 4.11 Factor loadings for evaluator roles subscale

Items	***Mean/SD***	***Factor 1***	***Factor 2***
A facilitator of local change	2.89/1.099	.804	
A shepherd to the public good	2.97/1.203	.528	
An educator to my clients	3.74/1.077	.500	
Part of the program team	3.27/1.312		.740
An achiever working with the program manager	2.85/1.265		.640
A methodological expert	3.98/.942	dropped	
A judge of the program	3.05/1.123	dropped	
A resource of program stakeholders	3.83/1.049	dropped	

Note. Factor 1 = Change agent/external roles; Factor 2 = Team-oriented/internal roles

Reported Sources of Questions & Issues

Three items were eliminated due to low communalities and cross-loadings. A two-factor solution was confirmed (Figure 4.5) and yielded a reasonable model fit: $\chi^2(4) = 22.9$, $p = .0001$, CFI = .963, TLI = .928, RMSEA = .102, CFit $p = .014$, SRMR = .031.

All factors loadings are significant ranging from .405 to .918 (see Table 4.12). While *Factor 1* contains two indicators that center around *stakeholder information needs*, *Factor 2* has four indicators focusing on *research and theory*. The factors have a low correlation of .381.

Table 4.12 Factor loadings for the subscale of reported sources of questions and issues

Items	***Mean/SD***	***Factor 1***	***Factor 2***
Information needs of program manager	4.29/.716	.918	
Information needs of program staff	4.05/.860	.834	
Past research/evaluation	3.89/.829		.747
Social science theory	3.03/1.196		.490
Evaluator's own experience about which questions are usually most important	3.93/.851		.435
Pending legislation	2.26/1.058		.405
Information needs of program clients	3.77/1.144	dropped	
Information needs of the client who paid for the evaluation	4.40/.980	dropped	
Pending decisions on the program being evaluated	3.60/1.051	dropped	

Note. Factor 1 = Stakeholder information needs; Factor 2 = Research/theory.

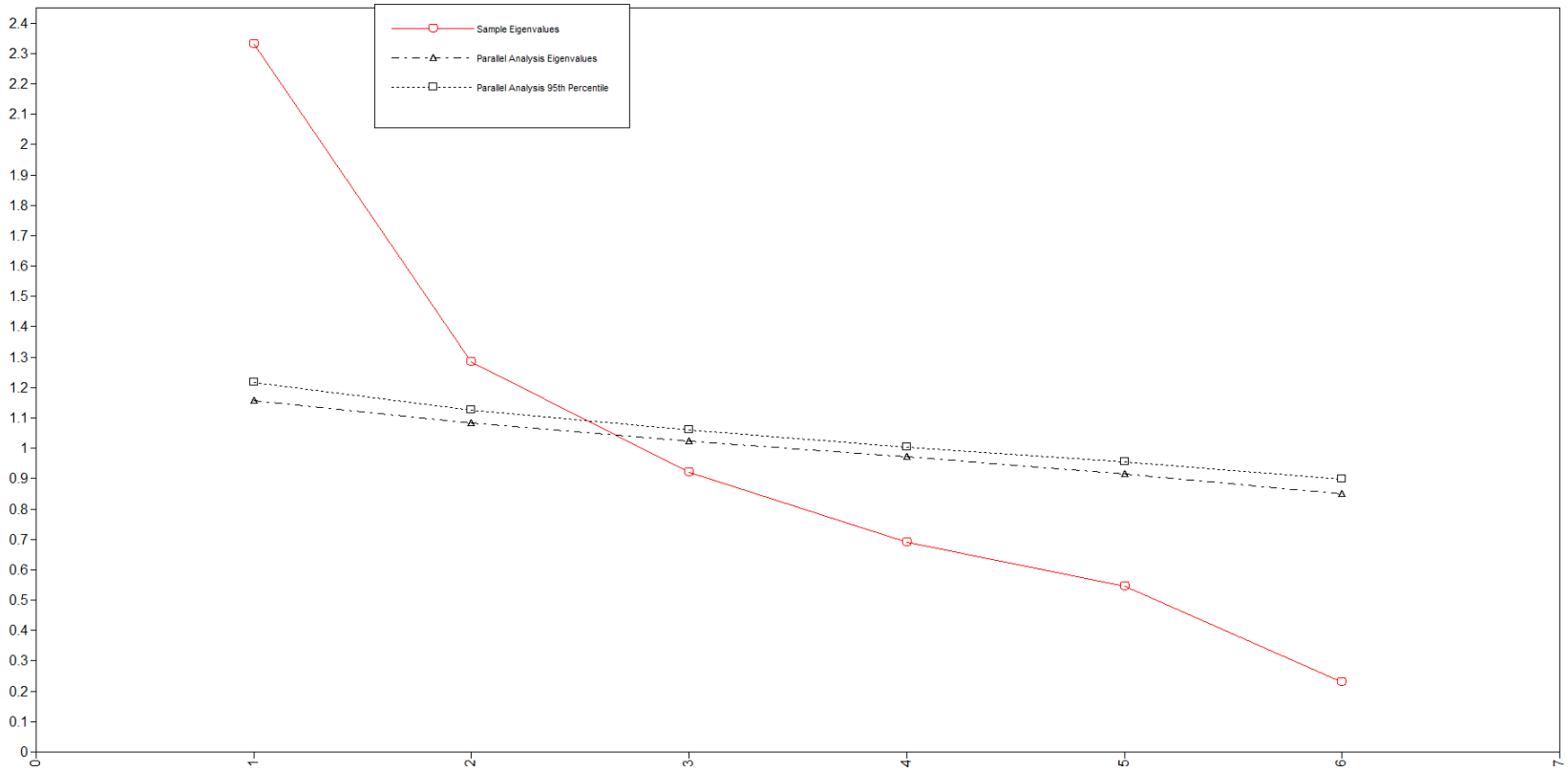


Figure 4.5 Parallel analysis plot for reported sources of questions and sources subscale

Central Issues On which Evaluation Data Were collected

Only one factor was extracted and all items were retained with significant factor loadings ranging from .467 to .666 (see Table 4.13). The one-factor model fit the data adequately: $\chi^2(9) = 24.309$, $p = .0038$, CFI = .949, TLI = .915, RMSEA = .061, CFit $p = .235$, SRMR = .037.

Table 4.13 Factor loadings for subscale of central issues

Items	***Mean/SD***	***Factor 1***
Number and characteristics of real and potential service recipients	3.84/1.089	.666
Changes in service recipients brought on by the program	3.82/1.068	.636
Explanation of variables that mediate the relationship between program implementation and effects	3.77/.979	.560
Manner in which the program is actually implemented	4.39/.742	.509
Cost and fiscal benefits of the program	2.78/1.097	.501
Changes in other people or in other institutions that interact with the program client	3.23/1.073	.467

Note. Factor 1 = Central Issues Factor.

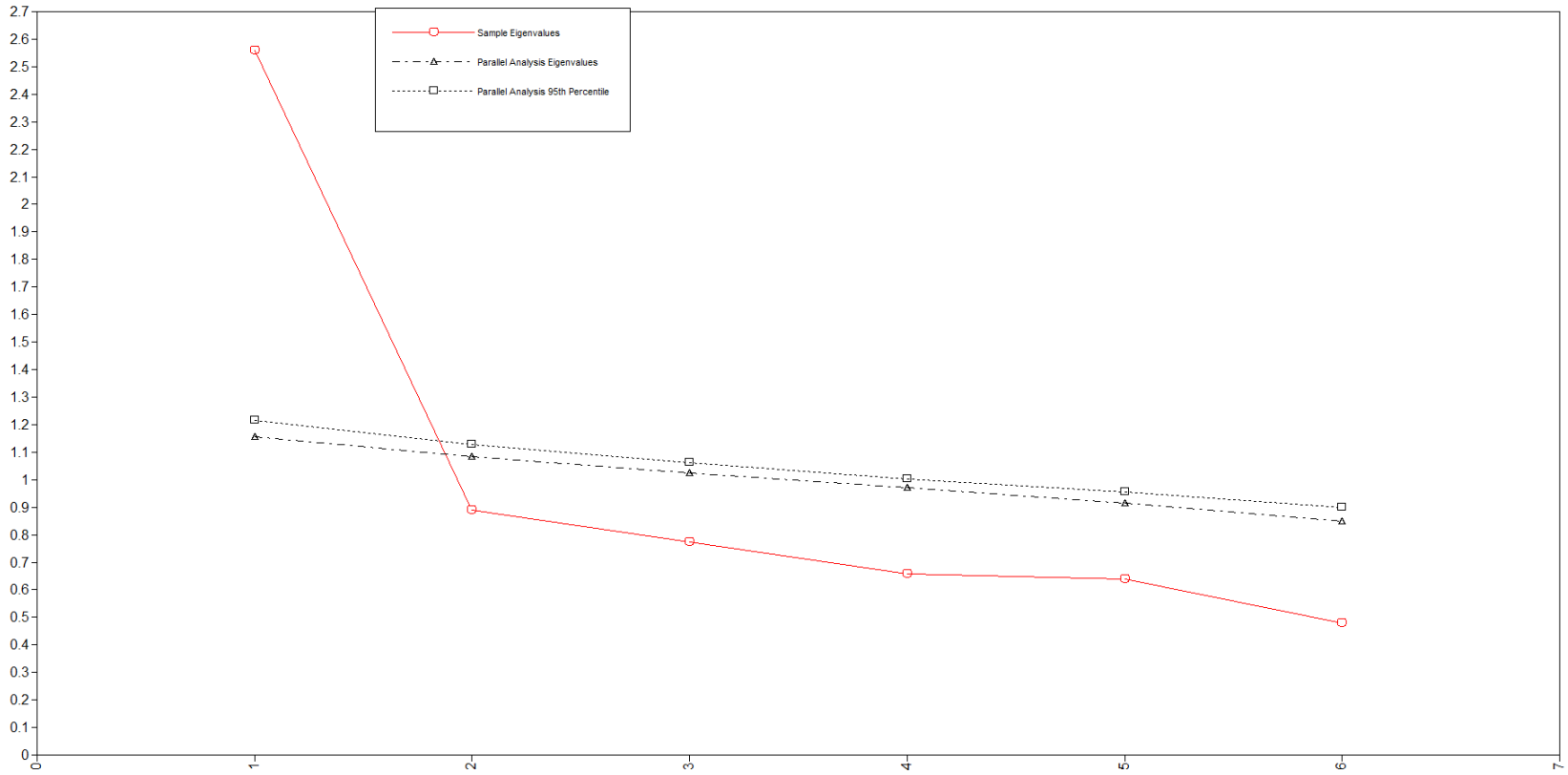


Figure 4.6 Parallel analysis Scree plot for the central issues on which evaluation data were collected subscale

Source of Dependent Variables for Program Effectiveness Questions

After two items were eliminated due to low communality and cross-loadings, a two-factor structure emerged and was confirmed by parallel analysis (see Figure 4.7) with χ^2 (13) = 82.183, $p < .0001$, CFI =.882, TLI = .745, RMSEA = .108, CFit $p = .0001$, SRMR = .041. However, the RMSEA and CFit *p* indicated that the model could be further improved.

All items loaded significantly on their respective factors with loading ranges from .467 to .824 (see Table 4.14). *Factor 1* has four indicators being *stakeholder-based*, while *Factor 2* has four indicators being *literature and research-based*. A moderate correlation of .433 was discovered between the two factors.

Table 4.14 Factor loadings for the subscale of dependent variables for program effectiveness

Items	***Mean/SD***	***Factor 1***	***Factor 2***
Criteria selected by program staff	3.51/.946	.824	
Criteria selected by program managers	3.77/.860	.719	
Criteria selected by program clients	3.05/1.131	.520	
Criteria selected by clients who paid for the evaluation	3.75/1.007	.510	
Unintended side effects	3.17/1.058		.702
The needs of the disadvantaged	3.33/1.110		.480
Criteria suggested by relevant social science theory	3.18/1.037		.467
Criteria in relevant program regulations or legislation	3.44/1.112		.401
Program goals	4.64/.592	dropped	
Criteria used in past evaluations of the program or similar programs	3.60/.819	dropped	

Note. Factor 1 = Stakeholder-based; Factor 2 = Literature/research-based.

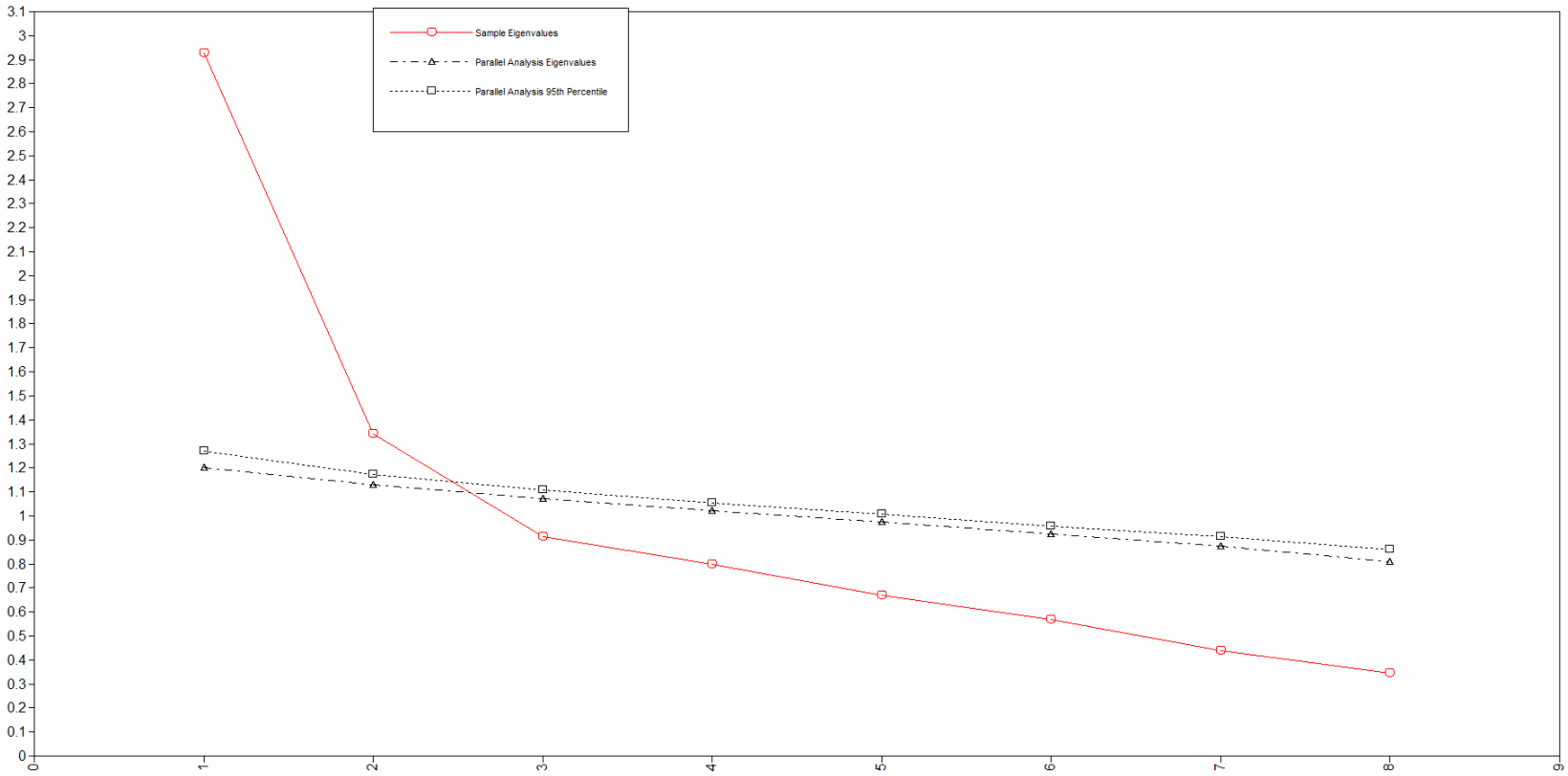


Figure 4.7 Parallel analysis plot for dependent variables for program effectiveness subscale

Methods Used in Evaluations

After eliminating three items, a four-factor model emerged. Although the parallel analysis (see Figure 4.8) suggested a three-factor solution, the four-factor solution was retained because of interpretability. The retained model achieved a good model fit: $\chi^2(17) = 33.327$, p = .011, CFI =.985, TLI = .951, RMSEA=.046, CFit p = .578, and SRMR = .018).

All factor loadings were significant with ranges from .306 to .988 (see Table 4.15). While the *secondary evaluation* factor correlated moderately (r = .501) with *quantitative methods* factor, all other factors had low correlations (.179 – .378).

Table 4.15 Factor loadings for the subscale of evaluation methods

	Mean/SD	***Factor 1***	***Factor 2***	***Factor 3***	***Factor 4***
Onsite observation	3.74/.914	.937			
Participant observation	3.15/1.053	.570			
Interviews with stakeholders	4.19/.786	.435			
Constructing a Meta-evaluation	1.95/.958		.906		
Conducting meta-analysis	1.75/.899		.801		
Randomized Experiment	1.75/.969			.701	
Casual modeling (e.g., Path analysis/Structural Equation Modeling)	1.92/.987			.692	
Quasi-experimental design	3.00/1.138			.580	
Program monitoring (e.g., Management Information system)	3.60/1.034				.988
Client needs assessment	3.24/1.069				.347
Inspecting program records	3.99/.905				.306
survey	4.12/.702	dropped			
Constructing program theory/Theory of Change	3.64/1.207	dropped			
Achievement tests	2.50/1.139	dropped			

Note. Factor 1 = Qualitative methods; Factor 2 = Secondary evaluation; Factor 3 = Quantitative methods; Factor 4 = Program Monitoring.

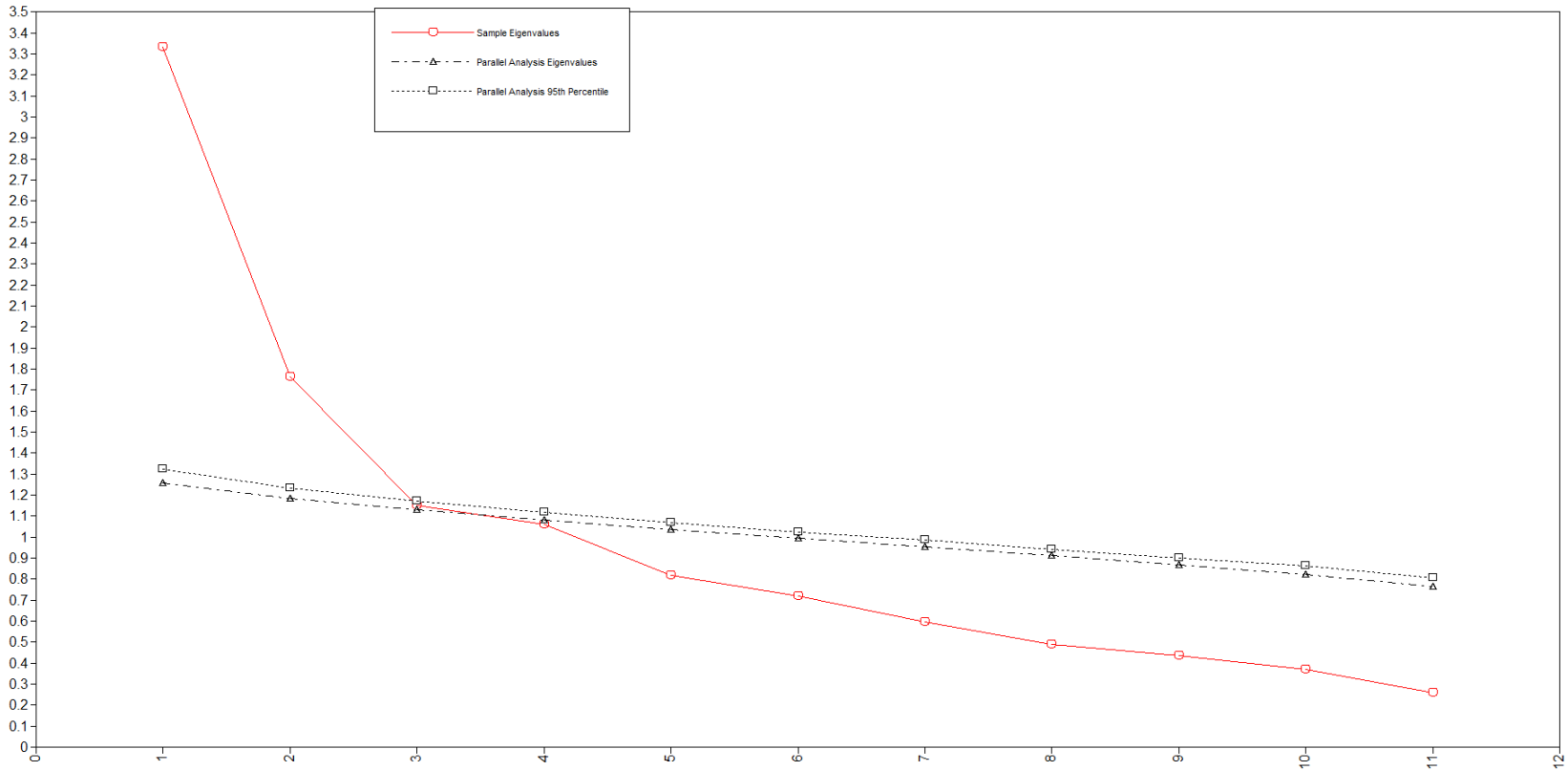


Figure 4.8 Parallel analysis plot for methods used in evaluations subscale

Activities to Facilitate Evaluation Use

After removing three indicators due to low communalities, two solutions (1-factor vs. 2-factor) were evaluated. Even though the parallel analysis indicated a 1-factor solution, the model fitting indices strongly suggested that the 2-factor structure should be favored because of the better model fit: $\chi^2(4) = .7.453$, $p = .113$, CFI = .994, TLI = .976, RMSEA = .044, CFit $p = .515$, SRMR = .015. The two factors has a moderate correlation of .649 with factor loadings ranging from .426 to .943.

Table 4.16 Factor loadings for subscale of activities to facilitate evaluation use

	Mean/SD	***Factor 1***	***Factor 2***
Keep in frequent contact with users during the conduct of the evaluation	4.35/.844	.779	
Provide oral briefings to clients	4.42/.770	.683	
Provide interim results to clients during the evaluation	4.15/.910	.612	
Translate results into action recommendation	4.37/.824	.426	
Ask the clients how potential evaluative information would be used to make change	4.07/1.020		.943
Identify potential users in order to include their questions in the evaluation	3.82/1.063		.549
Disseminate a written report of results	4.76/.535	dropped	
Publish results in books or journals	2.35/1.072	dropped	
Make evaluation results available to the public in the media	2.58/1.183	dropped	

Note. Factor 1 = Communication-oriented/Constant contact with stakeholder; Factor 2 = Participatory-oriented/Involving stakeholder in the evaluation process.

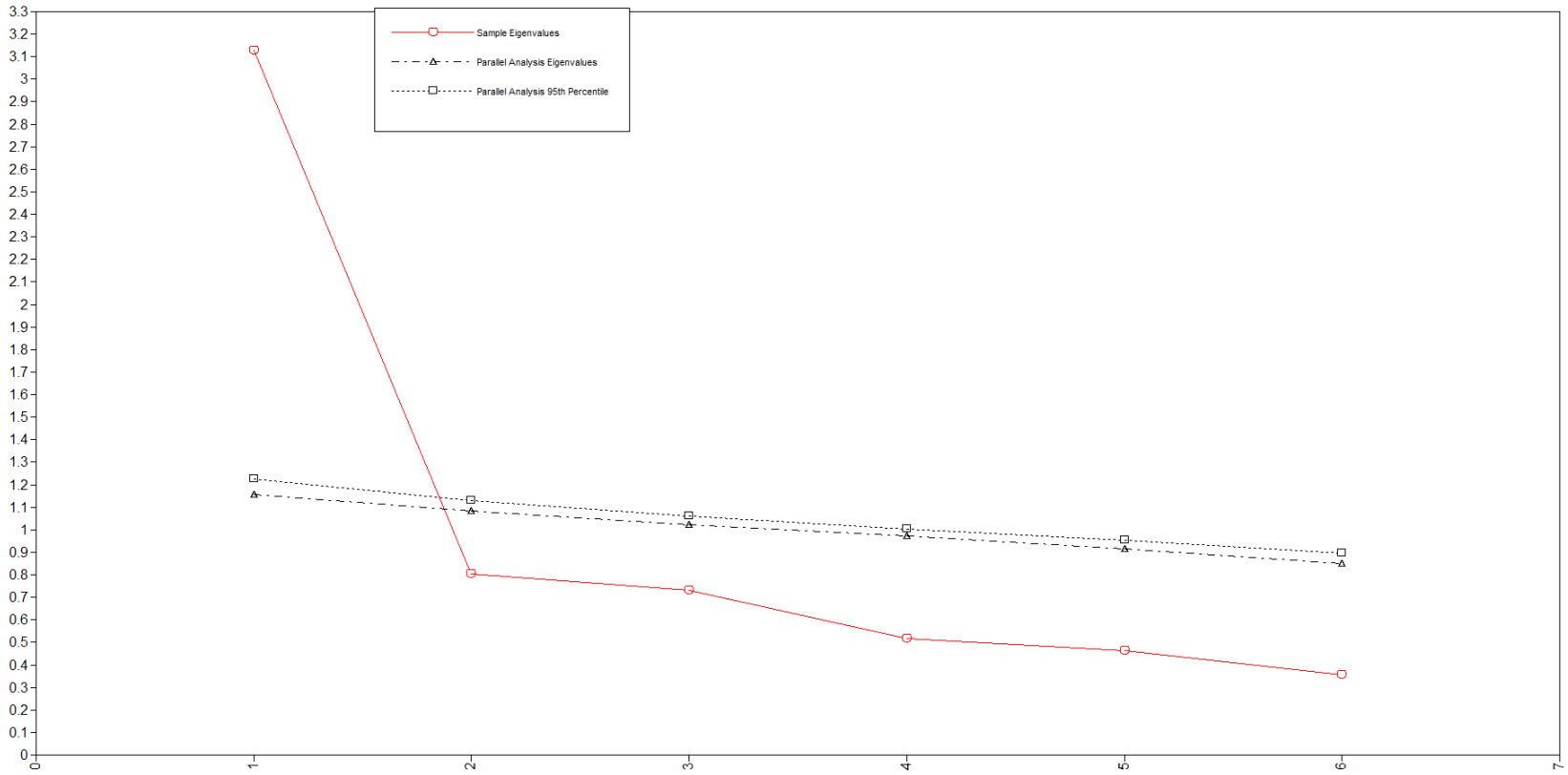


Figure 4.9 Parallel analysis plot for activities to facilitate evaluation use subscale

Eight EFA analyses generated 17 first-order factors. Although the 22-factor structure was not completely replicated, in each of the eight sub-domains, the results were approximately consistent. The reduced number of factors were primarily attributed to the item reduction approach adopted in the current study. Shadish and Epstein (1987) argued for the maximum number factors generated in the first-order analyses, and condensation effects on the second-order analysis, items with low communalities would inevitably muddle the shared variance or communalities and make factor interpretation more difficult. The subsequent EFAs revealed that overall the majority of factors had relatively low correlations and factor loadings with wide ranges.

R3. Does the higher-order factor structure conform to the four patterns yielded in Shadish and Epstein (1987)?

In examining second-order factor patterns, factor score approach adopted in Shadish and Epstein (1987) suffers from factor score indeterminacy issue (Gorsuch, 1983; Grice, 2001; Steiger, 1996), because an infinite number of correlation matrices could produce the same factor scores to explain the relationships between the indicators and factors. In other words, there is not a unique solution for the derived factor structure. Consequently, factor scores generated will differ by samples and studies with limited generalizability (Pett et al., 2003). Because of such indeterminacy issue, DiStefano, Zhu, and Mîndrilă (2009) caution researchers to draw conclusions using factor scores. In the current study, the approach with mean scores by factors was applied instead of the factor score approach for the reasons stated. A set of 17 composite mean scores were created by averaging the sum of indicator scores for each of the 17 factors in

the first-order EFA. Then, the 17 variables were subject to EFA analysis to examine the factor structure.

Though the results of the parallel analysis indicated a two-factor structure (see Figure 4.9), the four-factor solution revealed a good fit to the data: $\chi^2(74) = 143.751$, $p < .0001$, CFI = .960, TLI = .927, RMSEA = .045, CFit p = .747, SRMR = .028. *Factor 1* had eight first-order factors that loaded significantly with loading ranges from .426 to .601. The theme of this practice pattern was to fulfill basic research and scientific interest. Other than the *quantitative method factor*, *secondary evaluation method* factor also had a cross-loading on this practice pattern (λ = .277). This pattern was consistent with that of Shadish and Epstein (1987) and hence was also named *academic practice* pattern.

Table 4.17 Higher-order factor loadings for evaluation practice patterns

First-Order Factors	***Factor 1***	***Factor 2***	***Factor 3***	***Factor 4***
Purpose: Scientific idealistic	0.601			
Decision to Evaluate: Basic science interest	0.514			
Decision to evaluate: cost/benefit of evaluation	0.505			
Dependent variable: Literature-based	0.587			
Question source: Research & theory	0.535			
Role: Change agent	0.445			
Method: Quantitative	0.429	0.254		
Data gathered		0.686		
Method: Program monitoring		0.702		
Method: Secondary evaluations	0.277	0.349		
Method: Qualitative		0.395	0.276	
Purpose: Program-focused		*0.281*		
Activities to facilitate use: serving clients			0.792	
Activities to facilitate use: involving clients input			0.598	
Question source: Stake-holder Info need				0.805
Dependent variable: Stakeholder-based				0.532
Role: Team-oriented				*0.159*

Note. Factor 1 = Academic; Factor 2 = Method-driven; Factor 3 = Use-driven; Factor 4 = Stakeholder service.

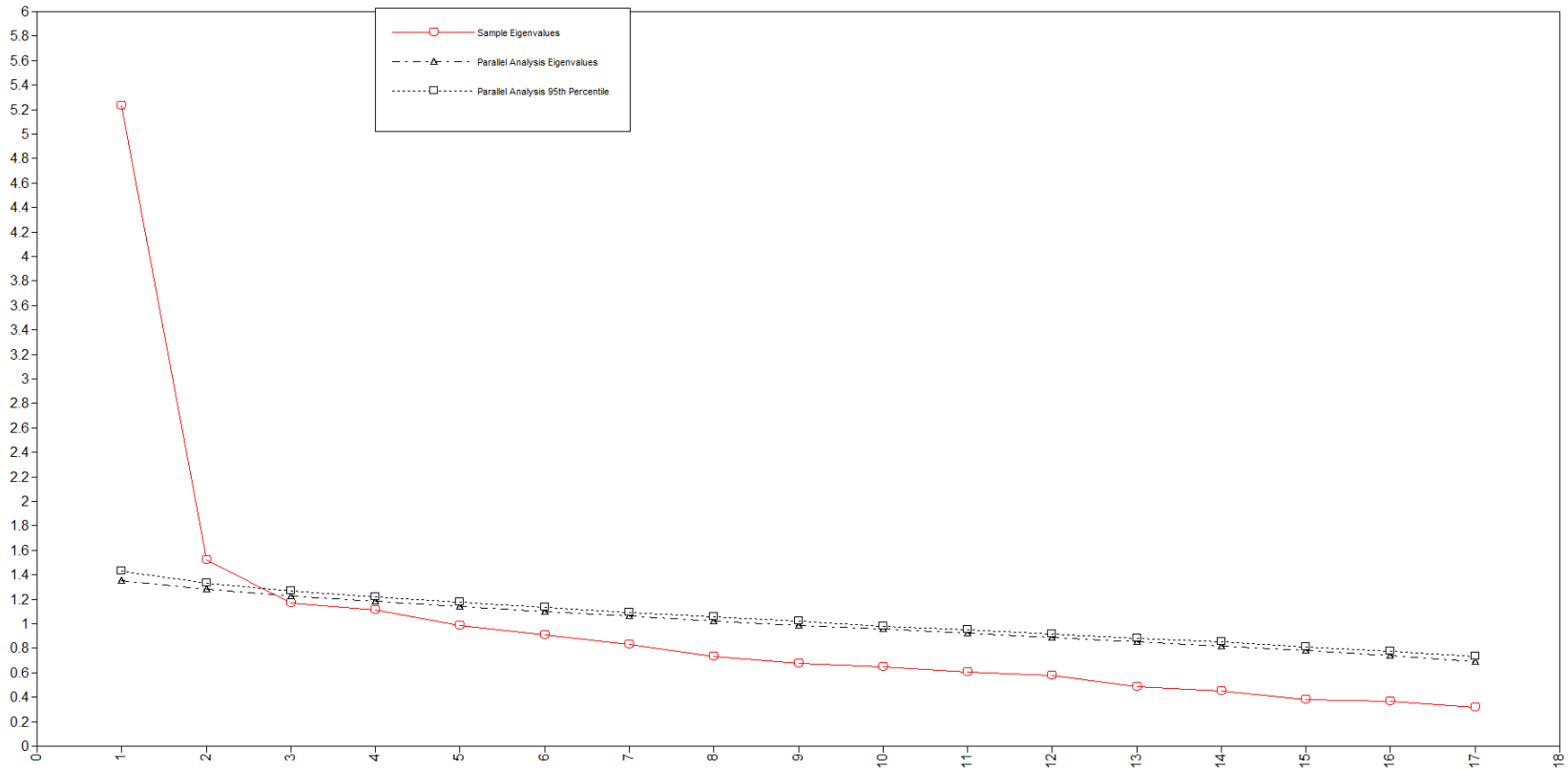


Figure 4.10 Parallel analysis plot for the higher-order factor structure

Factor 2 had four indicators that loaded significantly (> .30) with loading ranges from .349 to .702. This practice pattern focused predominantly on evaluation methods and source of data gathered. While all four first-order method factors loaded on this pattern, quantitative method factor loading (λ = .254) was just below .30. In this *method-driven* practice pattern, evaluators tend to emphasize the role of evaluation methods in their practice and utilize a wide variety of methods to address their central evaluation issues.

Factor 3 had two indicators that loaded significantly with loading ranges from .598 to .792. Both indicators focused on activities to facilitate evaluation use. The *use-driven pattern* also had a small loading (λ = .276) from qualitative method first-order factor, indicating that evaluators in this practice pattern have a methodological preference of qualitative methods.

Factor 4, stakeholder service pattern, had two indicators that loaded significantly with loading ranges from .532 to .805. Evaluators taking on this practice pattern tend to determine their evaluation questions based on stakeholder information needs and design evaluation studies with stakeholder

While the majority of first-order factors loaded significantly (> .30) on their corresponding second-order factors (Kahn, 2006), two failed to load significantly on any: *program-focused evaluation purpose* factor (highest loading = .281 on the *method-driven* pattern) and *team-oriented evaluator role* factor (highest loading = .159 on s*takeholder service* pattern).

Four higher-order factors or practice patterns also correlated differently. The *academic pattern* had a moderate correlation (.545, $p < .0001$) with the *method-driven* practice pattern. All other patterns had weak correlations ranging from .22 to .336.

Table 4.18 Evaluation Practice - Subscale reliabilities

Sub-scales	Cronbach's Alpha	Correlation Range	# of items
Evaluation Purposes	.682	.041 - .501	8
Influences on Decision to Evaluate	.702	.091 - .638	6
Evaluator Roles	.609	.049 - .477	5
Sources of Questions/Issues	.650	.108 - .766	6
Central Issues	.726	.173 - .495	6
Dependent Variables for program effectiveness	.732	.015 - .590	8
Methods used in Evaluations	.734	.026 - .70	10
Activities to Facilitate Use	.813	.308 - .63	6
		Total # of items:	55

Confirmative Phase

Research questions R4 to R8 were addressed in this phase by testing and presenting findings from a series of CFA models.

R4. Does the factor structure yielded from R1 achieve reasonably good model fit?

Perceived Importance of Competencies Rating

The results in R1 showed that the ECPE had five dimensions measured by 44 evaluator competencies: *evaluative practice*, *meta-competencies*, *evaluation knowledge base*, *project management*, and *professional development*.

All 44 indicators and the factor structure in evaluator perceived importance of competencies subscale were entered into CFA analyses. The test of the hypothesis that the ECPE has a five-factor structure yielded an inadequate fit with $\chi^2(892) = 2137.264$ and $p < .0001$. Given the sensitivity of the Chi-square test to sample size, other model-fitting indices were taken into consideration. Two most commonly used incremental indices of fit in SEM, comparative fit index (CFI) and Tucker-Lewis Index (TLI) were below .80 and hence did not reach acceptable

range. However, RMSEA (.055) and SRMR (.069) indices were within acceptable ranges. Overall, the model did not fit adequately, and a review of the modification indices (MIs) suggested that the hypothesized model could be modified to achieve a better fit by freeing up several parameters for estimates, such as cross-loaded items and error covariances.

Typically cross-loaded indicators may cause issues to interpret factor structures because it indicates overlapping contents between the factors. However, at this preliminary stage of the instrument development, the cross-loadings may be informative for redefining factors. Regarding correlated error variances, Byrne (2012) provides three contexts where it is appropriate to incorporate error covariance correlations into model respecification: a) there were significant item content overlaps; b) the same residual covariances were also included in previous research; and c) it is unreasonable to force large error terms uncorrelated as the correlations might indicate other common cause.

The largest MI indicated that the model Chi-square value would significantly drop if the residual covariance between item 19 “analyzing data” and item 20 “ interpret data” were to be freely estimated. With obvious content overlap, this modification was justified. Even though the overall fit was improved, χ^2 (891) = 2030.845, CFI = .817, TLI = .806, RMSEA = .053, CFit p = .065, SRMR = .065, AIC = 40738.681, the model did not achieve a reasonable fit.

Examination on MIs indicated that the model could be improved by incorporating additional residual covariances. For example, by estimating residual covariances between item 9 “Knowledgeable about qualitative methods” and item 10 “Knowledgeable about mixed methods”, the model fit was improved, χ^2 (890) = 1939.453, CFI = .832, TLI = .821, RMSEA = .051, SRMR = .065, AIC = 40623.218.

A series of modifications were incorporated in order to improve model fit. As a result, eight items were dropped to eliminate cross-loadings and content overlapping. For example, item 13 “Frames evaluation questions” strongly loaded on both evaluative practice and evaluation knowledge base factors; and item 31 “Address conflicts” and item 59 “Uses conflict resolution skills” had significant content overlap. Also, a large number of residual covariances were incorporated into the final model. The final ECPE measurement model had a total of 36 indicators and achieved a good fit: $\chi^2(565) = 872.650$, CFI = .935, TLI = .927, RMSEA = .034, CFit p = 1.000, SRMR = .053. Examination of the normalized residual covariance matrix did not reveal any problematic values (< 2). All indicators loaded statistically significantly on their corresponding factors, ranging from .446 to .799.

Table 4.19 Factor loadings for the final ECPE importance subscale

	Standardized Estimates	**Standard Errors**	**Critical Ratio**	***P*-value**
Evaluative Practice				
Conducts literature reviews	0.487	0.049	9.884	< .0001
Specifies program theory	0.528	0.044	11.953	< .0001
Conducts meta-evaluation	0.469	0.042	11.096	< .0001
Determines program evaluability	0.631	0.037	16.858	< .0001
Examines the organizational context of the evaluation	0.638	0.042	15.156	< .0001
Analyzes the political considerations relevant to the evaluation	0.554	0.045	12.322	< .0001
Attends to issues of evaluation use	0.61	0.043	14.303	< .0001
Attends to issues of organizational change	0.603	0.039	15.424	< .0001
Uses negotiation skills	0.658	0.041	15.88	< .0001
Uses conflict resolution skills	0.677	0.039	17.306	< .0001
Meta- Competencies				
Acts ethically and strives for integrity and honesty in conducting evaluations	0.540	0.124	4.347	< .0001
Respects clients, respondents, program participants, and other stakeholders	0.667	0.075	8.9	< .0001
Respects the uniqueness of the evaluation site and client	0.568	0.044	13.001	< .0001
Remains open to input from others	0.619	0.047	13.283	< .0001

Aware of self as an evaluator (knowledge, skills, dispositions)	0.611	0.069	8.848	< .0001
Uses written communication skills	0.492	0.08	6.118	< .0001
Uses verbal/listening communication skills	0.608	0.083	7.351	< .0001
Demonstrates cross-cultural competence	0.615	0.055	11.174	< .0001
Evaluation Knowledge Base				
Knowledgeable about quantitative methods	0.498	0.054	9.247	< .0001
Knowledgeable about qualitative methods	0.446	0.065	6.814	< .0001
Knowledgeable about mixed methods	0.511	0.061	8.323	< .0001
Develops evaluation design	0.663	0.052	12.689	< .0001
Collects data	0.537	0.079	6.772	< .0001
Assesses reliability of data	0.636	0.042	15.07	< .0001
Analyze data	0.68	0.058	11.693	< .0001
Interprets data	0.648	0.067	9.682	< .0001
Reports evaluation procedures and results	0.602	0.043	13.859	< .0001
Project Management				
Responds to requests for proposals	0.605	0.043	14.216	< .0001
Negotiates with clients before the evaluation begins	0.753	0.035	21.721	< .0001
Writes formal agreements	0.799	0.03	26.427	< .0001
Budgets an evaluation	0.761	0.029	25.824	< .0001
Justifies cost given information needs	0.762	0.033	23.415	< .0001
Professional Development				
Reflects on personal evaluation practice (competencies and areas for growth)	0.702	0.042	16.519	< .0001
Pursues professional development in evaluation	0.647	0.049	13.257	< .0001
Pursues professional development in relevant content areas	0.554	0.057	9.771	< .0001
Builds professional relationships to enhance evaluation practice	0.612	0.047	13.014	< .0001

All factor correlations ranged from moderate to strong (.414 – .722). *Evaluative practice* competencies correlated strongly with all other competencies (> .602), and the lowest correlation was between the *evaluation knowledge base* and *professional development* competencies.

Table 4.20 The ECPE factor correlation matrix

	Evaluator Competency Factors				
	Evaluative Practice	**Meta Competencies**	**Knowledge Base**	**Project Management**	**Professional Development**
Evaluative Practice	1.000				
Meta Competencies	.722	1.000			
Knowledge Base	.641	.574	1.000		
Project Management	.602	.448	.469	1.000	
Professional Development	.702	.675	.414	.433	1.000

The final ECPE measurement model achieved high internal consistency with Cronbach's Alpha of .909. All internal consistency measure – Cronbach's Alpha reached the critical cut point of .70. *Professional development* and *Meta-competencies* had slightly lower internal consistencies than other subscales. An alternative measure of reliability, omega (McDonald, 1999), was also included. Omega does not assume tau-equivalence and tends to be a "more sensible index of internal consistency" (Dunn, Baguley, & Brunsden, 2013). As such, omega's main advantages over Alpha include fewer and more realistic assumptions, less likely to underestimate or overestimate internal consistency, and more reflective of population estimates. Since the omega assumes unidimensionality, there are only reliability estimates for each subscale as the evaluator competencies scale is multidimensional. Table 4.21 shows that omega estimates, in this case, are more conservative than alpha estimates, but all within an acceptable range.

Table 4.21 Summary of the ECPE importance subscale reliability

Sub-scale	**Cronbach's Alpha**	**Omega**	**# of items**
Evaluative Practice	.835	.791	10
Meta-competencies	.752	.766	8
Evaluation Knowledge Base	.805	.759	9
Project Management	.854	.794	5
Professional Development	.749	.671	4
Scale Level:	.909		
		Total # of items:	36

Self-Assessed Level of Competencies Ratings

The final CFA model derived from the perceived importance of competencies ratings was also tested with the self-assessed level of competencies ratings. The goodness-of-fit indicated an adequate model fit: $\chi^2(565) = 1104.378$, p = .000, CFI = .931, TLI = .923, RMSEA = .046 (90% CI =.042 – .050; CFit *p* = .966), SRMR = .052. Although MIs presented several large values suggesting cross-loaded items and several residual covariances, no additional modifications were made to the measurement model for several reasons. Firstly, the incorporation of the cross-loaded item 11 "conducts literature reviews" on Factor Meta competencies rendered the measurement model empirically under-identified and consequently inestimable. Secondly, the close examination of item content suggested that modifications would only marginally be supported and indicated by low correlations with other items in this factor. Lastly, no modifications were made in order to maintain the factor structure consistency and the ease of comparison between the two ratings.

Factor correlations for self-assessed level ratings were significantly higher than those of the perceived importance ratings, ranging from .630 to .894. The difference in factor correlations could potentially suggest the bias in self-reporting nature of the ratings. In perceived importance scale, evaluators were requested to assess how important the competencies were to the entire evaluation profession; Whereas, the self-assessment ratings requested evaluators to evaluate their own levels of competencies. Given the previous moderate to strong factor correlations established for importance ratings, the correlations in self-assessment rating could have been magnified.

Table 4.22 Factor correlations for the self-assessed levels of competencies rating

	Evaluator Competency Factors				
	Evaluative Practice	**Meta-Competencies**	**Knowledge Base**	**Project Management**	**Professional Development**
Evaluative Practice	1.000				
Meta Competencies	.894	1.000			
Knowledge Base	.822	.841	1.000		
Project Management	.847	.730	.687	1.000	
Professional Development	.812	.850	.711	.630	1.000

Correlations of Perceived Importance and Self-Assessed Competencies Subscales

Correlations between the two ratings of the perceived importance of evaluator competencies and the self-assessed competencies were also examined. Overall, evaluators' perceived importance of competencies and self-assessed competencies had a statistically significant but weak correlation (r = .417). Furthermore, correlations of five subscales were also weak ranging from .007 to .444. The pattern of low correlations validated the objectivity of evaluators' responses. It suggested that evaluators' ratings on the importance of competencies had not greatly influenced their self-assessed levels on these evaluator competencies.

R5. Does the factor structure yielded from R2 achieve reasonably good model fit?

To confirm the factor structures of subscales, eight CFAs were carried out in Mplus 8.0 using the MLE estimator. The same set of model-fitting indices was used to evaluate how well each measurement model fit the data.

Evaluation Purpose

All goodness-of-fit indices suggested that the two-factor model fit the data very well with $\chi^2(19) = 32.069$, p = .0307, CFI = .969, TLI =.954, RMSEA = .039 (90% CI =.012 – .061; CFit

p = .774), SRMR = .037. Examination of normalized residuals and MIs did not indicate any localized areas of strain. Factor loading estimates revealed that all indicators were strongly related to their purported latent factors ranging from .428 to .813. The two factors were also moderately correlated (.423).

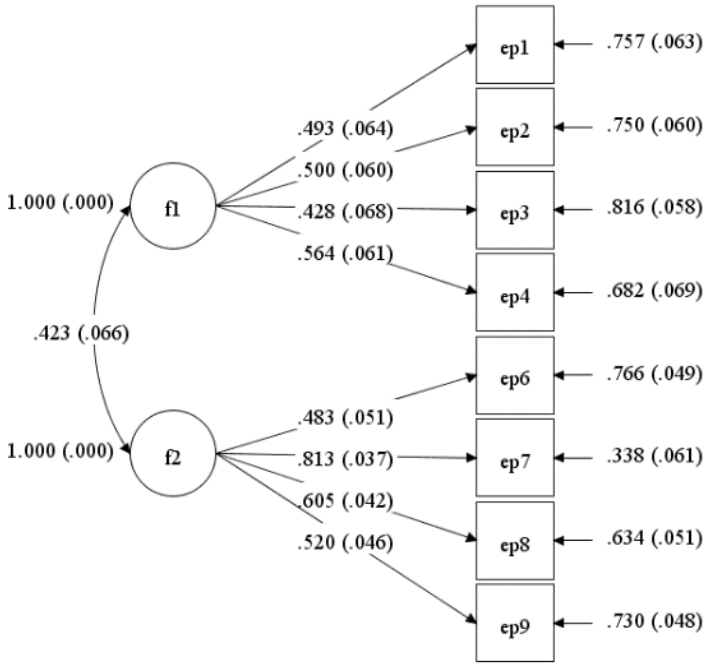


Figure 4.11 Path diagram for evaluation purpose subscale

Decision to Evaluate

The initial model fit indices suggested adequate fit: $\chi^2(8) = 31.185$, p = .0001, CFI = .950, TLI = .906, RMSEA = .080 (90% CI =.051 – .110; CFit *p* = .043), and SRMR = .051. The MIs suggested that the mode fit could be improved with residual covariances freely estimated between items 11 and 12, which had substantive content overlap regarding evaluator's interest. After respecifying the initial model, the goodness-of-fit indices achieved a reasonable fit with $\chi^2(7) = 18.415$, p = .0102, CFI = .975, TLI = .947, RMSEA = .060 (90% CI =.027 - .094; CFit *p* = .274), and SRMR = .036. All factor loadings were statistically significant and loaded strongly

on their purported latent factors ranging from .409 to .780. The two latent factors had a moderate correlation (.456), and additionally, the residual covariance was also statistically significant (.526).

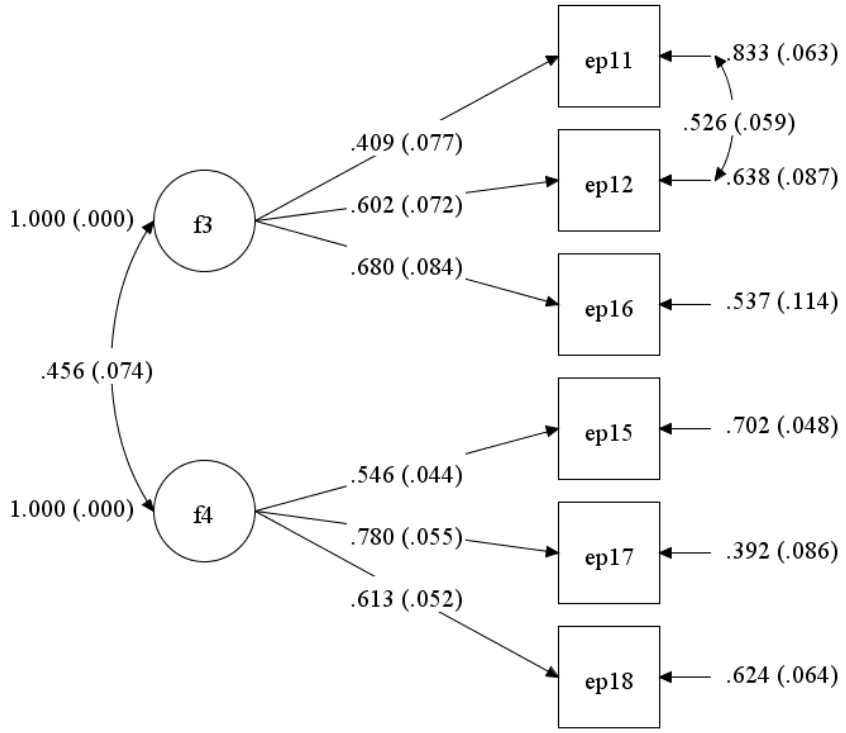


Figure 4.12 Path diagram for decisions to evaluate subscale

Evalautor Roles

The two-factor model achieved a very good fit: $\chi^2(4) = 7.319$, $p = .1199$, CFI = .986, TLI = .966, RMSEA = .043 (90% CI =.000 – .091; CFit $p = .527$), SRMR = .020. Examination of normalized residuals and MIs did not indicate any localized areas of strain. Factor loading estimates revealed that all indicators were strongly related to their purported latent factors ranging from .466 to .792. In addition, the two latent factors were also weakly correlated (.333).

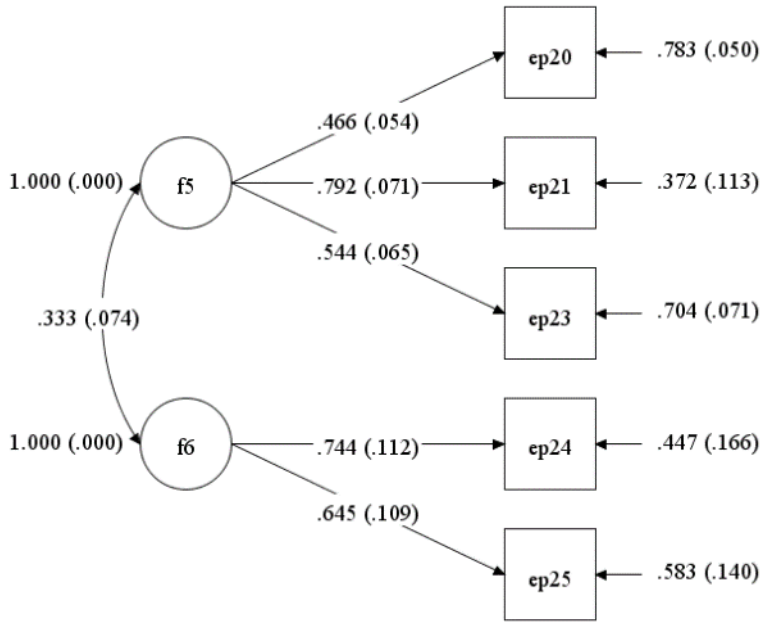


Figure 4.13 Path diagram for evaluator roles subscale

Sources of Questions/Issues

The two-factor model achieved a good fit: $\chi^2(8) = 22.226$, $p = .0045$, CFI = .972, TLI = .948, RMSEA = .062 (90% CI =.032 – .094; CFit $p = .222$), SRMR=.034. Examination of normalized residuals and MIs did not indicate any localized areas of strain. Factor loading estimates revealed that all indicators were strongly related to their purported latent factors ranging from .390 to .905. In addition, the two latent factors were weakly correlated (.392).

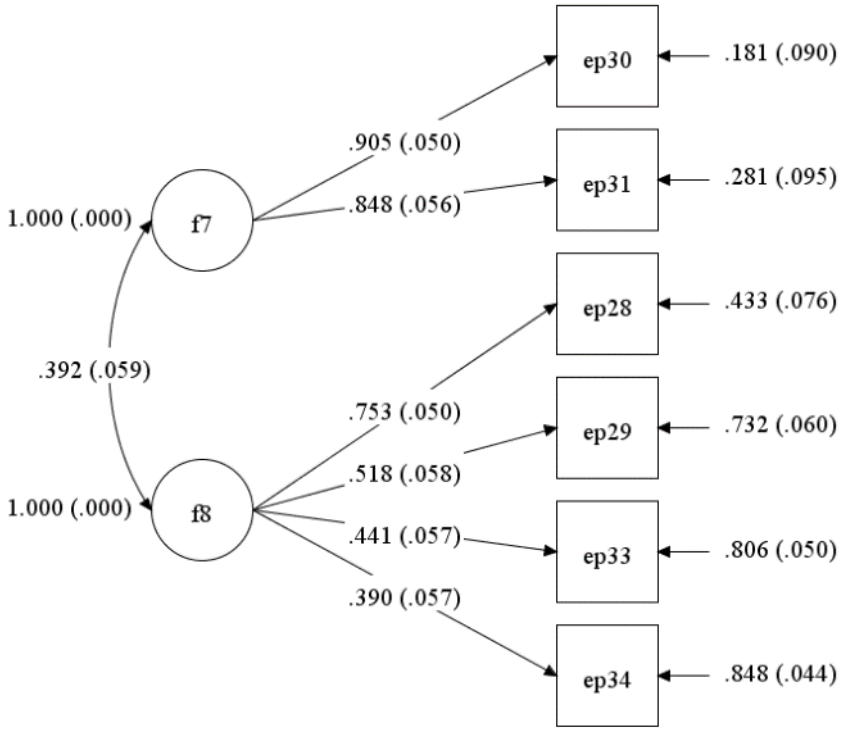


Figure 4.14 Path diagram for reported sources of evaluation questions/issues

Central Issues Data Collection

The resultant goodness-of-fit indices indicated a good model fit to the data: $\chi^2(9)$ = 24.309, p = .0038, CFI = .949, TLI = .915, RMSEA = .061 (90% CI =.032 – .091; CFit p = .235), SRMR = .037. No localized areas of strains were detected after examining normalized residual matrix and MIs. All factor loadings were statistically significant.

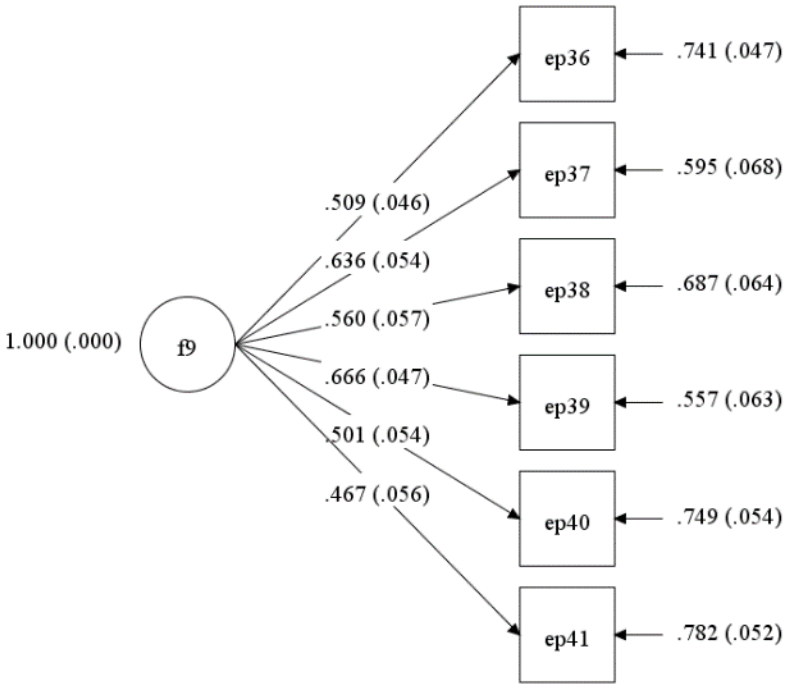


Figure 4.15 Path diagram for central issues subscale

Dependent Variables

The initial two-factor model did not achieve an adequate fit, $\chi^2(19) = 85.977$, $p < .0001$, CFI = .885, TLI = .831, RMSEA = .088 (90% CI =.070 – .107; CFit p = .001), SRMR = .053. The review of MIs revealed several relatively large values (>10): a cross-loading (item 49 "Criteria selected by program clients" on Factor 11 Stakeholder-based dependent variables) and a set of residual covariances. If item 49 were loaded onto the other factor, the overall model χ^2 statistic could decrease by 18.994, and the expected parameter change (EPC) indicated that the estimated parameter loading would be .322. However, the item 49 content did not thematically fit to cross-load on the factor, and hence was not included into the respecified model. Item 44 demonstrated multiple residual covariances, and hence eliminated. The final model only included one set of covariance between item 46 and item 49 that could be justified.

The respecified model achieved a goodt fit: $\chi^2(12) = 41.256$, $p < .0001$, CFI = .945, TLI = .914, RMSEA = .073 (90% CI =.049 – .098; CFit $p = .054$), and SRMR =.041. All factor loadings were statistically significant on their respective factors ranging from .424 to .786. The estimated value of error covariance was also statistically significant.

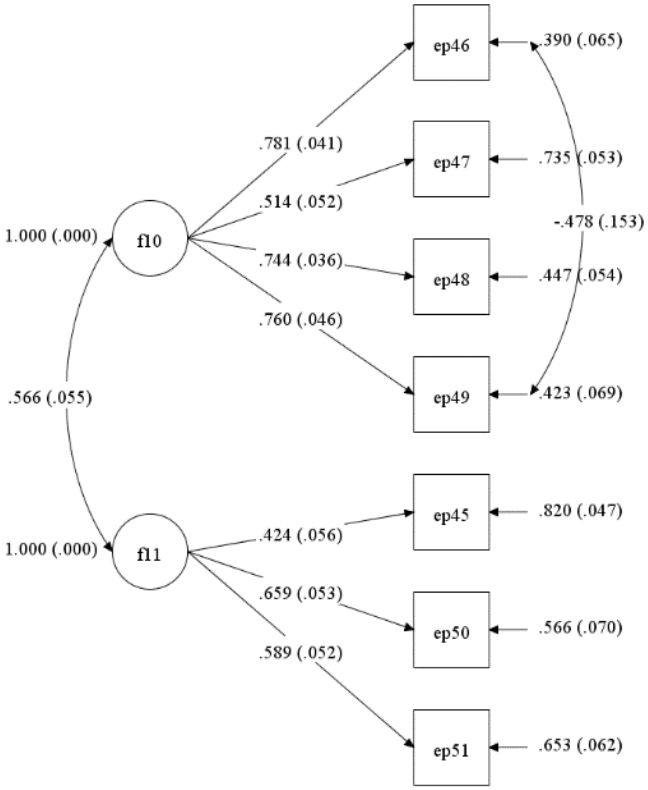


Figure 4.16 Path diagram for evaluation dependent variables subscale

Methods

The initial four-factor model did not achieve adequare fit with $\chi^2(38) = 162.938$, $p < .0001$, CFI = .884, TLI = .832, RMSEA = .085 (90% CI =.072 – .099; CFit $p < .0001$), and SRMR = .058. Although the review of MIs revealed several large values, three in particular stood out from the rest (MI = 42.271; MI = 27.118; MI = 16.348). All MIs signified residual

covariances, and involved item 52 with three other items, 57, 53 and 56. The fact that item 52 were involved in all three residual covariances indicated that there were substantive content overlapping. Consequently, item 52 was eliminated and the respecified model achieved an adequate fit, $\chi^2(29) = 81.919$, $p < .0001$, CFI = .945, TLI = .914, RMSEA = .063 (90% CI =.048 – .080; CFit p = .080), SRMR = .041. Factor parameter loadings on their purported latent factors were statistically significant ranging from .462 to .883. Additionally, the four latent factors were correlated ranging from .211 to .560.

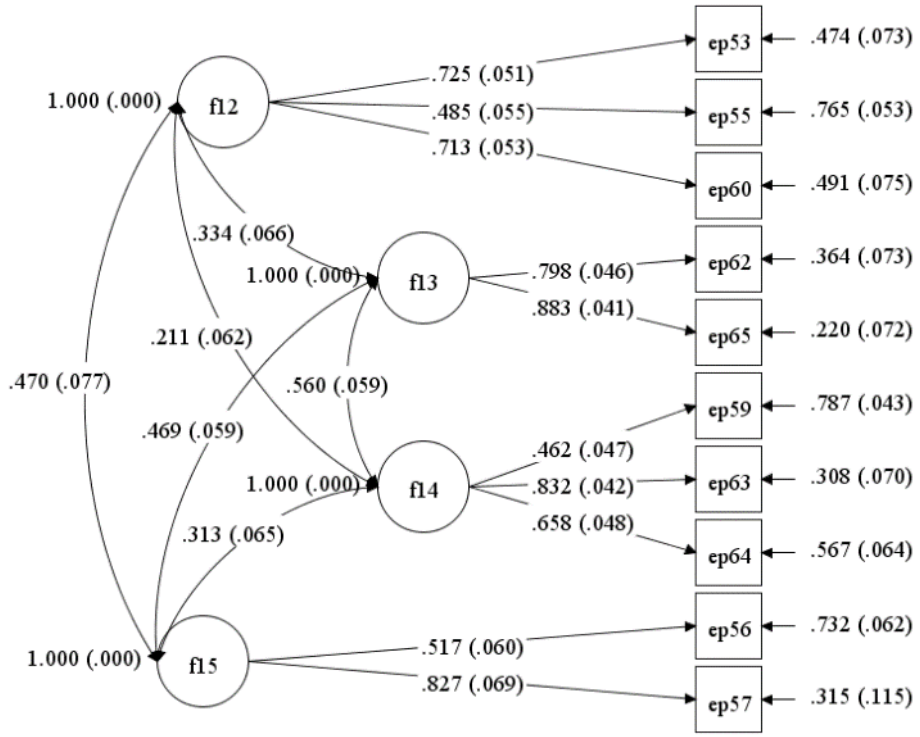


Figure 4.17 Path diagram for evaluation methods subscale

Activitites to Faciliate Use

The two-factor model was confirmed with excellent fit indices, χ^2 (8)= 10.924, p = .206, CFI = .995, TLI =.990, RMSEA = .028, and SRMR = .021. The two factors were strongly correlated (.754) and all the estimated factor loadings were statistically significant ranging from .507 to .805.

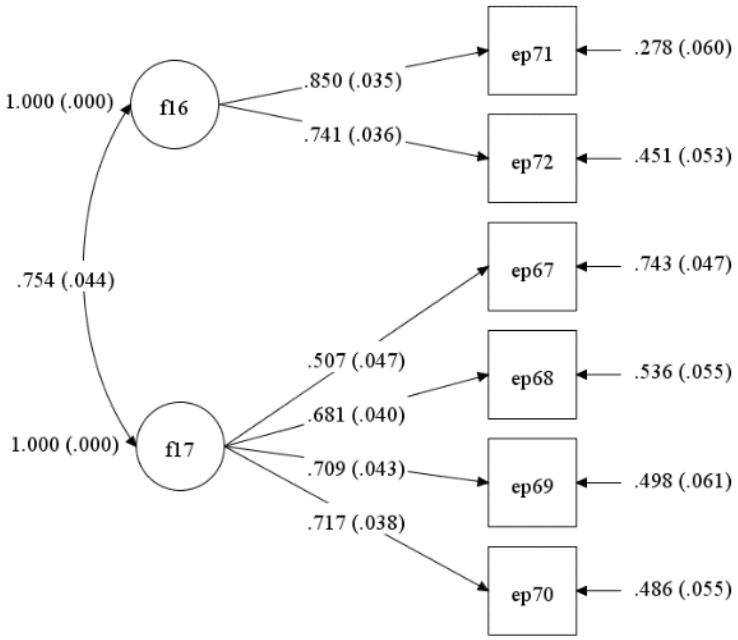


Figure 4.18 Path diagram for evaluation methods subscale

Even though some initial models did not fit adequately, the results from eight CFA final models confirmed the factor structures yielded from EFAs. All CFA models presented in Table 4.23 achieved an adequate fit with high CFIs (> .945), and particularly non-significant RMSEA's CFit $p > .05$ (Brown, 2015) indicating RMSEA values within the acceptable ranges. Additionally, the majority of all CFA models yielded non-significant χ^2 statistics, which typically extremely sensitive to large sample size. The next section describes how well the higher-order factor structure fits the data.

Table 4.23 Summary model fit statistics of CFA models for evaluator practice subscales

Models	X^2/df	***P***	**CFI/TLI**	**RMSEA (90% CI)**	**CFit *p***	**SRMR**
CFA: Evaluation Purpose	32.069/19	.0307	.969/.954	.039 [.012 .061]	.774	.032
CFA: Decision to Evaluate	18.415/7	.0102	.975/.947	.060 [.027 .094]	.274	.036
CFA: Evaluator Roles	7.319/4	.1199	.986/.966	.043 [.000 .091]	.527	.020
CFA: Source of Issues	22.226/8	.0045	.972/.948	.062 [.032 .094]	.222	.034
CFA: Central Issues	24.309/9	.0038	.949/.915	.061 [.032 .091]	.235	.037
CFA: Dependent Variables	41.256/12	< .0001	.945.904	.073 [.049 .098]	.054	.041
CFA: Methods	81.919/29	< .0001	.945/.914	.063 [.048 .080]	.080	.041
CFA: Activities for use	10.924/8	.2061	.995/.990	.028 [.000 .066]	.796	.021

Note. CFI = comparative fit index; TLI = Tucker-Lewis fit index; RMSEA = root mean square error of approximation; CI = confidence interval; CFit *p* = close fit; SRMR = standardized root mean square residual.

R6. Does the higher-order factor structure yield from R3 achieve reasonably good model fit?

Both item-level and subscale-level models were estimated to evaluate the fit of the four-factor structure/evaluator practice patterns emerged in EFA higher-order analysis.

The item-level model is a more complex model as it uses the item scores for each respondent to fit a 17-factor first-order model, which in turn served as the indicators for the four-factor model. Even though the item-level model yielded acceptable RMSEA and SRMR, the Chi-square test and comparative fitting indices did not meet the criteria: $\chi^2(1406) = 3232.600$, $p < .0001$, CFI = .746, TLI = .732, RMSEA = .053, SRMR = .066. The poor fit could be attributed to the complexity of the model, low correlations among factors, as well as multivariate abnormal nature of the data, even with MLR estimator adjustment.

Taking into account the CFA results (R5) confirming 17 first-order factor structure, the subscale-level model was fitted to assess the measurement model of four practice-patterns. Composite scores were calculated by averaging item scores for each of the 17 first-order factors. The initial model fit the data adequately, $\chi^2(97) = 207.856$, $p < .0001$, CFI = .932, TLI = .916, RMSEA = .050, and SRMR = .045. The MIs suggested that the model fit could be further improved. The largest value of MI (17.946) suggested residual covariances between *Factor 13* (Methods: Secondary Evaluations) and *Factor 14* (Method: Quantitative) might have certain content overlapping. *Factor 13* is measured by two indicators, one of which, meta-analysis, is usually considered a quantitative method.

Table 4.24 Correlations matrix of 17 evaluation practice first-order factors

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17
E1	1.000																
E2	0.431	1.000															
E3	0.374	0.406	1.000														
E4	0.215	0.453	0.335	1.000													
E5	0.327	0.565	0.229	0.451	1.000												
E6	0.112	0.115	0.145	0.156	0.344	1.000											
E7	0.213	0.108	0.077	0.173	0.156	0.197	1.000										
E8	0.358	0.535	0.344	0.491	0.425	0.125	0.355	1.000									
E9	0.558	0.531	0.254	0.422	0.399	0.226	0.229	0.532	1.000								
E10	0.264	0.234	0.175	0.23	0.274	0.307	0.592	0.338	0.38	1.000							
E11	0.429	0.754	0.447	0.581	0.601	0.155	0.353	0.986	0.815	0.504	1.000						
E12	0.287	0.252	0.296	0.273	0.345	0.193	0.208	0.368	0.552	0.278	0.518	1.000					
E13	0.204	0.374	0.176	0.25	0.368	0.12	0.071	0.329	0.453	0.165	0.513	0.336	1.000				
E14	0.291	0.357	0.274	0.389	0.24	0.169	0.041	0.42	0.469	0.155	0.510	0.221	0.541	1.000			
E15	0.546	0.445	0.288	0.366	0.387	0.409	0.302	0.389	0.763	0.379	0.621	0.515	0.504	0.340	1.000		
E16	0.341	0.214	0.178	0.29	0.386	0.06	0.322	0.31	0.435	0.272	0.451	0.470	0.101	0.057	0.365	1.000	
E17	0.335	0.4	0.298	0.357	0.514	0.128	0.385	0.389	0.434	0.416	0.516	0.395	0.220	0.082	0.440	0.734	1.000

Note. E1 = program-focused purpose; E2 = scientific idealistic purpose; E3 = basic scientific interest; E4 = cost/benefit of evaluation; E5 = Change agent/external roles; E6 = Team-oriented/internal roles; E7 = Stakeholder information needs; E8 = Research/theory; E9 = Central Issue on which data collected; E10 = Stake holder –based; E11 = Literature-based; E12 = Qualitative methods; E13 = Secondary evaluation; E14 = Quantitative methods; E15 = Program monitoring; E16 = Communication-oriented; E17=Participatory-oriented.

The respecified model fit was improved slightly after freeing the parameter estimates between *Factor 13* and *Factor 14* residual variances: $\chi^2(96) = 189.752$, $p < .0001$, CFI = .942, TLI = .928, RMSEA = .046 (90% CI =.036 – .056; CFit p = .736), SRMR=.042. Three additional modifications were considered substantive and thus incorporated. The final model had one crossloading of *Factor 12* (Methods: Qualitative methods) on the method-driven pattern and stakeholder-service pattern; two additional sets of residual covariances: *Factor 9* (Data gathered) with *Factor 14* (Quantitative methods) and *Factor 3* (Decision to Evaluate: Basic Science Interest). The obvious content overlap warranted the modifications (Byrne, 2012).

The final mode achieved a good fit: $\chi^2(94) = 167.773$, $p < .0001$, CFI = .955, TLI = .942, RMSEA = .041 (90% CI =.031 – .051; CFit p = .920), SRMR = .041. All estimated parameters, including factor loadings (.241 - .853), residual covariances (.156 - .256), and factor correlations (.501 - .760) were statistically significant.

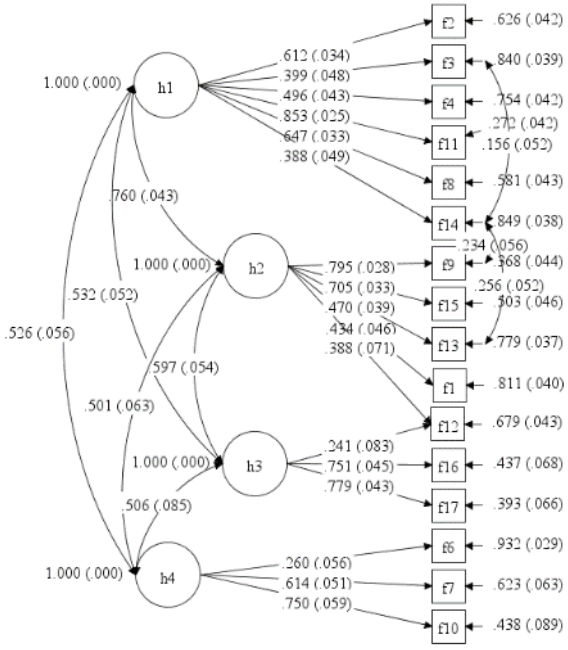


Figure 4.19 Path diagram for four evaluation practice patterns

R7. Does the aforementioned set of eight covariates have statistically significant effects on the measurement model established in R4?

In examining measurement invariance, multiple indicators multiple causes (MIMIC) modeling, or CFA with covariates, is a less commonly used method but has several advantages over multiple-group analysis (Brown, 2015). MIMIC models can examine a large number of comparison groups simultaneously; have smaller sample size requirements; can accommodate categorical or continuous predictors or covariates, and can be easily tested by adding covariates to well-validated CFA models. On the other hand, MIMIC models are limited in testing only the invariance of indicator intercepts, and factor means, assuming many other model parameters, such as equal factor loadings, error variances/covariances, or factor variances/covariances, equal across all grouping/covariate levels.

After the five-factor structure for ECPE was established in R1 and confirmed in R4, eight covariates were introduced to ECPE measurement model: years of experience, professional identity, primary affiliation, highest degree achieved, the field of highest degree, job setting, evaluation background, and gender. All covariates were coded as categorical variables, and the multicollinearity was checked using Pearson Chi-square tests. Results showed that the highest Phi/Cramer's V statistic was .357, which indicated no multicollinearity issues among the covariates.

The MIMIC model for the perceived importance of evaluator competencies fit the data adequately with χ^2 (875)= 1314.298, p < .0001, CFI = .908, TLI = .896, RMSEA = .034 (90% CI =.032 – .040; CFit p = 1.000), and SRMR = .050. Since all covariates were categorical, unstandardized estimates in Mplus was presented and interpreted. As Table 4.25 indicates, only a small number of covariates had significant direct effects on the five sub-scales of evaluator

competencies, which indicated measurement variances or population/group heterogeneity. Years of experience had significant positive effects on evaluative practice, knowledge base, and project management factors, suggesting that evaluators tend to rate the importance of competencies in evaluative practice, knowledge base and project management higher as their years of experience increase. Specifically, as years of experiences were dummy coded (*0 = 15 years or less; 1 = more than 15 years*), it can be interpreted that the evaluators with more than 15 years experiences tended to rate .109 units, .117 units, and .192 units higher on evaluative practice, knowledge base, and project management competencies than evaluators with 15 years or less experiences.

Table 4.25 Unstandardized estimates from the estimated MIMIC model for percived importance of evaluator competencies

Covariates	Evaluator Competency Factors				
	Evaluative Practice	**Meta-Competencies**	**Knowledge Base**	**Project Management**	**Professional Development**
Years of Experience	.109 (.041*)	.027 (.119)	.117 (.008**)	.192 (.023*)	-.105 (.101)
Professional Identity	.028 (.743)	-.011 (.516)	-.068 (.299)	-.078 (.500)	-.111 (.186)
Primary Affiliation	-.098 (.142)	.008 (.644)	-.041 (.442)	.016 (.874)	-.158 (.013*)
Highest Degree	-.177 0.016*)	.000 (.994)	-.131 (.032*)	-.140 (.175)	.029 (.665)
Field	-.033 (.568)	-.022 (.105)	-.008 (.881)	-.212 (.027*)	-.111 (.078)
Job Setting	.003 (.967)	-.008 (.587)	-.035 (.464)	.042 (.650)	.067 (.284)
Evaluation Background	.102 (.136)	-.001 (.952)	.038 (.490)	.076 (.490)	.075 (.332)
Gender	-.075 (.229)	-.069 (.027*)	-.008 (.880)	.047 (.600)	-.093 (.195)

Note. *$p < .05$; **$p < .001$.

Also, the primary affiliations (*0 = AEA; 1 = Other*) had a significant negative effect on professional development competencies. It can be concluded that the mean of evaluators identifying with AEA as their primary affiliation is .153 units higher than the mean of the non-AEA group on the importance of professional development competencies.

Furthermore, the degree variable (*0 = Bachelor & Master's; 1 = Doctorate*) had statistically significant effects on academic and knowledge base competencies, indicating that evaluators with doctorate degrees tended to rate higher on evaluative practice and knowledge base competencies than evaluators with Bachelor's or Master's degree.

The field variable (*0 = education/evaluation/psychology; 1 = others*) had a significant negative effect on project management competencies. To be specific, evaluators receiving their highest degrees in education, evaluation, and psychology rated the importance of project management competencies significantly higher than evaluators in other fields such as sociology and social work.

Lastly, the gender variable (*0 = female; 1 = male*) had a significant negative effect on Meta competencies, indicating that the mean rating of female evaluators on the importance of Meta-competencies was .069 units higher than that of male evaluators.

Despite some significant direct effects indicating a certain level of heterogeneity, the measurement invariance has mostly been achieved in the ECPE scale, and this further strengthens the scale's usability across different populations/groups.

R8. Does the aforementioned set of eight covariates have significant effects on the measurement model established in R6?

Similarly, an MIMIC model was fitted to examine the effects of covariates on four practice patterns of EP scale. Eight covariates were added to the four-factor measurement model established in R6, and the goodness-of-fit index confirmed a well-fitting model with $\chi^2(214) = 332.173$, $p < .0001$, CFI = .932, TLI = .911, RMSEA = .036 (90% CI =.028 – .043; CFit p = .999), and SRMR = .039. The examination of normalized residual matrix and MIs did not reveal any localized strains.

As Table 4.26 revealed, years of experience (*0 = 15 years and less; 1 = more than 15 years*) had positive effects on Academic and Use-driven practice patterns. The results suggested that evaluators with more evaluation experience engage in these two patterns (Academic and Use-Driven) of practice more frequently than evaluators with 15 years and less experience. Years of experience did not affect methodology-driven and stakeholder-service practice patterns.

Professional identity (*0 = Evaluator; 1 = other professionals*) had a significant positive effect on Method-driven pattern. The results showed that respondents who identify professionally with other professions have a higher mean (unstandardized estimate = .161) than respondents who identify as evaluators on method-driven practice factor. It suggested that non-evaluation professionals tended to take on method-driven patterns more frequently than professional evaluators did in their practices.

Table 4.26 Results of unstandardized estimates from the MIMIC model of evaluator practice patterns

Covariates	Evaluator Practice Patterns (Higher-Order)			
	Academic	**Method-Driven**	**Use-Driven**	**Stakeholder-Service**
Years of Experience	.126 (.017*)	.092 (.102)	.128 (.017*)	.074 (.066)
Professional Identity	.034 (.650)	.161 (.047*)	-.049 (.554)	-.027 (.651)
Primary Affiliation	.041 (.478)	-.121 (.065)	.024 (.721)	-.048 (.301)
Highest Degree	-.100 (.112)	-.052 (.429)	-.101 (.115)	.053 (.226)
Field	-.043 (.417)	.052 (.360)	-.037 (.517)	-.045 (.305)
Job Setting	-.002 (.977)	.129 (.032*)	.109 (.082)	-.055 (.211)
Evaluation Background	.044 (.535)	.281 (< .001**)	-.070 (.372)	-.089 (.107)
Gender	-.027 (.664)	.051 (.420)	-.111 (.073)	-.013 (.782)

Note. *$p < .05$; **$p < .001$.

Additionally, the highest degree achieved (*0 = Bachelor or Master's; 1 = Doctorate*) had positive effects on academic, methodological-driven, and use-driven practice patterns. However, the effects were not statistically significant. Furthermore, the job setting variable (*0 = Non-college/university; 1 = college/university*) had a significant negative effect on method-driven practice pattern, concluding that evaluators in settings, such as federal/state government and non-profit organizations more frequently take on method-driven practice pattern than evaluators in college/university setting.

Also, the evaluation background (*0 = U.S. based and US programs; 1= others*) had a positive and statistically significant effect on the method-driven practice pattern. The results indicated that U.S. based evaluators who predominantly evaluate U.S. programs were less frequently to engage method-drive practice than evaluators based elsewhere were. The absence of significant effects of primary affiliation, the field of highest degree and gender suggested measurement invariances in these population/group variables.

Structural Phase:

R9. How do evaluator competencies relate to their evaluation practice patterns? Specifically, do evaluator self-assessed competencies have significant effects on evaluation practice patterns? Alternatively, do evaluators' practice patterns have significant effects on their self-assessed competencies?

Two structural models were tested, a) whether evaluators' self-assessed competencies affect their practice patterns; and b) whether evaluators' practice patterns affect how they assessed their competencies.

Self-Assessed Evaluator Competencies as Predictors

The goodness-of-fit indices indicated a good fit, $\chi^2(154) = 232.028$, $p < .0001$, CFI = .953, TLI = .953, RMSEA = .035 (90% CI =.025 – .044; CFit $p = .998$), and SRMR = .039. Examination of normalized residual variance-covariance matrix and MIs did not reveal any localized fit issues.

The results showed that *evaluative practice* competencies had significant effects on academic and *method-driven* practice patterns, but not on the other two patterns. The results suggest that evaluators with higher self-assessed competencies on *evaluative practice* competencies are more likely to take on *academic* and *method-driven* patterns.

Table 4.27 Results from the estimated SEM model of self-assessed evaluator competencies as predictors

Evaluator Practice Patterns	**Evaluator Competency Factors**			
	Academic	**Method-Driven**	**Use-Driven**	**Stakeholder-Service**
Evaluative Practice	.442 (< .001**)	.539 (< .001**)	.217 (.133)	.222 (.106)
Meta-Competencies	-.028 (.774)	-.041 (.652)	.221 (.015*)	.139 (.184)
Knowledge Base	.035(.673)	-.211 (.021*)	-.160 (.123)	-.142 (.194)
Project Management	.128 (.116)	.076 (.342)	.096 (.256)	-.019 (.843)
Professional Development	-.093 (.207)	.034 (.665)	.103 (.208)	.061 (.492)

Note. *$p < .05$; **$p < .001$.

Meta-Competencies had a significant effect on the use-driven pattern, suggesting that evaluators with higher self-assessed *meta-competencies* are more likely to take on the *use-driven* practice pattern.

Additionally, the *evaluation knowledge base* competencies had a significant negative effect on the *method-driven* pattern, indicating that evaluators with higher self-assessed

evaluation knowledge base competencies tend to avoid method-driven practice pattern frequently.

The absence of significant effects of *project management* and *professional development* indicated that evaluator practice patterns were not influenced in any way by their self-assessed *project management* and *professional development* competencies.

Evaluator Practice Patterns as Predictors

The model with evaluator practice patterns as predictors achieved a good fit, $\chi^2(154) = 251.654$, $p < .001$, CFI = .967, TLI = .955, RMSEA = .037 (90% CI = .029 – .045; CFit $P = .996$), and SRMR = .039. Examination of the modifications indices did not indicate the presence of any localized areas of strains.

The results in Table 4.28 shows that *academic* pattern had significant effects on how evaluated rated their levels of competencies in three areas of *evaluative practice, Meta competencies*, as well as *knowledge base*. As it suggested, the more frequently evaluators engage in *academic* practice patterns, the higher they rate their competencies in those three areas.

Furthermore, the *use-driven* pattern had significant effects on all competencies except knowledge base. Evaluators engaging more frequently in this practice pattern tend to rate their competencies in *evaluative practice, meta-competencies*, *project management*, and *professional development* higher.

Table 4.28 Results of the estimated SEM model of evaluator practice patterns as predictors

Practice Patterns	Evaluator Competency Factors				
	Evaluative Practice	**Meta-Competencies**	**Evaluation Knowledge Base**	**Project Management**	**Professional Development**
Academic	.340 (< .001**)	.258 (.007*)	.473 (< .001**)	.360 (.001)	.115 (.275)
Method-Driven	.122 (.239)	-.063 (.553)	-.166 (.142)	.012 (.911)	.059 (.624)
Use-Driven	.179 (.022*)	.302 (.001*)	.147 (.097)	.191 (.020*)	.250 (.011*)
Stakeholder-Service	-.085 (.247)	-.032 (.697)	-.107 (.222)	-.121 (.118)	-.031 (.725)

Note. *$p < .05$; **$p < .001$.

Post-Hoc Sample Size Estimation & Empirical Power Analysis

Sample size in SEM is dependent on a wide variety of factors, such as the number of variables, correlations of variables, factor loading size, model complexity, reliability of observed indicators, multivariate normality, missing data handling, and model estimation methods (Raykov and Marcoulides, 2000; MacCallum, Widaman, Preacher, and Hong, 2001). Hence, commonly derived rules-of-thumb are difficult to generalize to specific models (Wolf, Harrington, Clark, and Miller, 2013).

Researchers conducted statistical simulation studies to investigate the sample size issue from various perspectives. For example, MacCallum, Widaman, Zhang, and Hong (1999) investigated how variable communalities influence the same size. With high communalities (higher than .60) among variables, sample size can be as low as 60 to reproduce the population loadings. Even when variable communalities are lower around .50, a sample size of 100 to 200 cases is required to reproduce the population estimates. A Monte Carlo simulation study by Muthén and Muthén (2002) examined the effects of normality and missing data on sample size and power of a two-factor CFA model. The study concluded that a sample size of 150 is sufficient for a power of .80 to reject the null hypothesis of zero factor correlation if variables are normally-distributed without any missing data. Under the condition of non-normal data with missing data, a sample size of 315 is needed for a power of .81.

Wolf et al. (2013) observed three major approaches to assessing sample size adequacy and statistical power: a) Satorra and Saris (1985) method based on the noncentrality parameters; b) the MacCallum, Browne, and Sugawara (1996) method based on RMSEA value; and c) the Monte Carlo simulation method (Muthén & Muthén, 2002). In this study, the second approach

was taken to evaluate the sample size and statistical power. Comparing with two other methods, MacCallum, Browne, and Sugawara (1996) method is simple to carry out and not model specific. In this method, a pair of Root Mean Square Error of Approximation (RMSEA) values is adopted to estimate sample size and the power to reject the null hypothesis. Preacher and Coffman (2006) implemented the methodology in a set of easy-to-use and convenient online simulators.

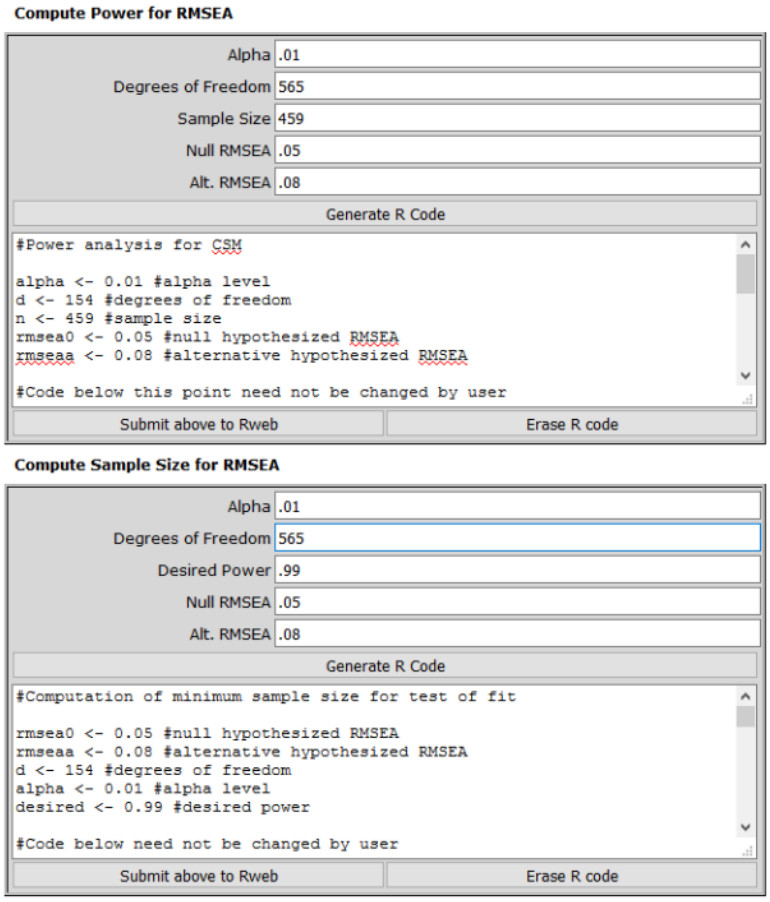


Figure 4.20 Simulation utility using RMSEA by Preacher and Coffman (2006)

To estimate statistical power, five model parameters, including the Alpha level, the degree of freedom, sample size, and null and alternative RMSEA indices, are necessary. Instead, when estimating sample size, the desired power of .99 has been provided. Post hoc analyses for

seven models were carried out: a) CFA model for perceived importance of evaluator competencies; b) CFA model for self-assessed level of competencies; c) CFA model for second-order evaluator practice patterns; d) MIMIC model for perceived importance of evaluator competencies; e) MIMIC model for evaluator practice patterns; f) structural model with evaluator practice patterns predictors; and g) structural model with self-assessed competencies as predictors.

Table 4.29 displays the estimation parameters for the seven models and the estimated results in the last columns. The results demonstrated that the sample size of the study (n = 459) far exceeded the minimum required sample size for all analyses conducted. In addition, all models achieved high statistical power in detecting type II error.

Table 4.29 Simulated results of statistical power analyses and minimum sample size for RMSEA

Models	Alpha	*DF*	Sample Size	Desired Power	Null/Alternative RMSEA	Statistical Power	Minimum Sample Size
Model a:	.01	565	459	.99	.05/.08	1	97
Model b:	.01	565	459	.99	.05/.08	1	97
Model c:	.01	94	459	.99	.05/.08	.999	350
Model d:	.01	875	459	.99	.05/.08	1	73
Model e:	.01	214	459	.99	.05/.08	1	188
Model f:	.01	154	459	.99	.05/.08	.999	238
Model g:	.01	154	459	.99	.05/.08	.999	238

Note. Model a-g corresponds to the models described above.

Summary

The research questions and analytical results were presented in three phases. In the exploratory phase, exploratory factors analyses were carried out to examine the factor structures of ECPE and EP scales. Even though ECPE researchers hypothesized a six-factor model, EFA results suggested that a five-factor structure emerged after testing alternative model structures and eliminating items with overlapping content coverage. Additionally, eight separate EFA analyses on the EP scale were conducted. The results approximated those in Shadish and Epstein (1987) study, and a total of 17 factors emerged. Subsequently, the mean sub-scale scores from the 17 factors were factor-analyzed, and four practice patterns emerged. Due to the factor score indeterminacy, the study adopted a composite score per factor approach in extracting "second-order" factors from the original study.

In the confirmatory phase, the factor structures yielded in exploratory phase were verified under the CFA framework. For the ECPE scale, two CFA models (perceived importance and self-assessment) with the same factor structure were tested and yielded adequate fit. For the EP scale, eight item-level and a subscale-level CFAs were carried out, and almost all models achieved a good fit to the data. Measurement invariance was also investigated on the ECPE perceived importance and the EP scales. Two MIMIC models achieved adequate fit, and overall both measurement models were invariant across populations with few exceptions.

In the structural phase, results on relationships of self-assessment level of competencies and evaluator practice patterns were presented. Both models yielded reasonably good model fitting. A small number of significant direct effects revealed how self-assessed evaluator competencies affect evaluators' practice patterns, as well as how evaluator practice patterns and

evaluator self-assessed level of competencies reciprocally influenced each other. Table 4.30 presents the model fitting indices for all final CFA and MIMIC models tested in the study.

Table 4.30 Summary of CFA Model fit indices

Model	**Chi-Square/*DF***	***p***	**CFI**	**TLI**	**RMSEA (90% CI)**	**CFit *p***	**SRMR**
CFA: Importance	872.650/565	<.001	.935	.927	.034 [.030 .039]	1.000	.053
CFA: Self-assessment	1104.378/565	<.001	.931	.923	.046 [.042 .050]	.966	.052
CFA: EP-Higher-order	167.773/94	<.001	.955	.942	.041 [.031 .051]	.920	.041
MIMIC: Self-assessment	1314.298/875	<.001	.908	.896	.034 [.030 .038]	.998	.050
MIMIC: EP	332.173/214	<.001	.932	.911	.036 [.028 .043]	.996	.039

Note. EP = evaluator practice; MIMIC = multiple indicators multiple causes; CI = confidence interval; CFI = comparative fit index; TLI = Tucker-Lewis fit index; RMSEA = root mean square error of approximation; CI = confidence interval; CFit *p* = close fit probability; SRMR = standardized root mean square residual.

CHAPTER V: DISCUSSION

Fundamental issues in evaluation are defined as those problems, subjects, topics, or themes that are of critical import to many or all aspects of evaluation. These issues are deeply rooted in the field, ever-evolving, cumulative in nature, and reappear in different forms at different times. Smith (2008) explained why a better understanding of these fundamental issues in evaluation helps advance the field of evaluation:

> If we can identify such issues in our work, we can then better examine their importance, reflect on how they impact our work, and develop more effective ways of dealing with them. Such examinations may help us keep our current problems in better historical perspective, support more thoughtful considerations of our present options, and enable us to create more effective alternatives for the future (p. 4).

The purpose of the study is to renew the understanding of three fundamental issues in evaluation by reviewing the historical perspectives and developing new strategies to shed lights on alternative solutions for future research. The results from previous research, although not fully verified, undoubtedly guided the analytical phases in this study. In the exploratory phase, the study first sought to replicate the factor structures resultant in previous research. After the initial factor solutions did not fit the data appropriately, the study explored the appropriate factor structures for ECPE and EP scales. In the confirmatory phase, the newly discovered factor structures were confirmed and psychometric properties, reliability as well as measurement invariance, were inspected. In the structural phase, the study explored the relationship between evaluators' self-assessed level of competencies and their evaluation practice patterns. Though the causal relationship was inconclusive (as the study does not definitely determine the directionality of the causal relationship between the two constructs), the results revealed statistically significant effects of self-assessed competencies of sub-domains affected evaluators practice patterns and

alternatively the statistically significant effects of evaluator practice patterns on self-assessed competencies.

Summary of Findings

While the primary goals of the exploratory and confirmatory phases were to establish the construct validity and measurement models for essential evaluator competencies and evaluator practice patterns, the structural phase aimed to examine structural relationships of these two critical constructs.

Table 5.1 Summary of research hypotheses and findings in three analytical phases

Analytical Phases	Research Questions	Hypotheses	Results
Exploratory Phase	R1.	ECPE importance scale was hypothesized as a 6-factor model.	Rejected
	R2.	EP was hypothesized to consist of 22 first-order factors in eight sub-scales.	Rejected
	R3.	EP scale was hypothesized to consist of 4 second-order factors.	Partially Accepted
Confirmatory Phase	R4.	ECPE importance scale has a 5-factor structure.	Accepted
	R5.	EP scale has 17 first-order factors in eight sub-scales.	Accepted
	R6.	EP scale has four higher-order factors.	Accepted
	R7.	ECPE importance scale is measurement-invariant with 8 covariate groups.	Partially Accepted
	R8.	EP's four higher-order factor structure is measurement-invariant with 8 covariate groups.	Partially Accepted
Structural Phase	R9a.	Evaluators' self-assessed competencies have no effects on their practice patterns.	Partially Rejected
	R9b.	Evaluators' practice patterns have no effects on their self-assessed competencies.	Partially Rejected

Essential Competencies for Program Evaluators

Whereas the essential competencies for program evaluators proposed as a six-dimension construct made conceptual sense, empirical results supported a five-factor solution. The

reduction of items and dimensions in the ECPE scale should not be surprising given the content overlapping of the original set of items. For example, item 31 “Addresses conflicts” in the original *Situational Analysis* dimension and item 59 “Uses conflict resolution skills” in *Interpersonal Competence* dimension; similarly, item 40 “Negotiate with clients before the evaluation begins” in *Project Management* dimension and item 58 “Uses negotiation skills” in *Interpersonal Competence* dimension. After eliminating repetitive items, the more concise ECPE scale revealed more interpretable factor structure. To a certain extent, the empirically derived five dimensions share commonalities with the conceptualized six dimensions, which are manifested by the items under each dimension.

Table 5.2 Comparison of conceptualized and empirically derived dimensions for the final items

Items	**Conceptual Dimensions**	**Empirical Dimensions**
Determines program evaluability	Situational Analysis	Evaluative Practice
Analyzes the political considerations relevant to the evaluation	Situational Analysis	Evaluative Practice
Examines the organizational context of the evaluation	Situational Analysis	Evaluative Practice
Specifies program theory	Systematic Inquiry	Evaluative Practice
Conducts literature reviews	Systematic Inquiry	Evaluative Practice
Attends to issues of organizational change	Situational Analysis	Evaluative Practice
Attends to issues of evaluation use	Situational Analysis	Evaluative Practice
Uses conflict resolution skills	Interpersonal Competence	Evaluative Practice
Conducts meta-evaluation	Systematic Inquiry	Evaluative Practice
Uses negotiation skills	Interpersonal Competence	Evaluative Practice
Respects clients, respondents, program participants, and other stakeholders	Professional Practice	Meta-Competencies
Uses verbal/listening communication skills	Interpersonal Competence	Meta-Competencies
Remains open to input from others	Situational Analysis	Meta-Competencies
Acts ethically and strives for integrity and honesty in conducting evaluations	Professional Practice	Meta-Competencies
Uses written communication skills	Interpersonal Competence	Meta-Competencies
Aware of self as an evaluator (knowledge, skills, dispositions)	Reflective Practice	Meta-Competencies
Demonstrates cross-cultural competence	Interpersonal Competence	Meta-Competencies
Respects the uniqueness of the evaluation site and client	Situational Analysis	Meta-Competencies

Analyze data	Systematic Inquiry	Knowledge Base
Interprets data	Systematic Inquiry	Knowledge Base
Knowledgeable about quantitative methods	Systematic Inquiry	Knowledge Base
Assesses reliability of data	Systematic Inquiry	Knowledge Base
Knowledgeable about mixed methods	Systematic Inquiry	Knowledge Base
Collects data	Systematic Inquiry	Knowledge Base
Reports evaluation procedures and results	Systematic Inquiry	Knowledge Base
Develops evaluation design	Systematic Inquiry	Knowledge Base
Knowledgeable about qualitative methods	Systematic Inquiry	Knowledge Base
Writes formal agreements	Project Management	Project Management
Budgets an evaluation	Project Management	Project Management
Justifies cost given information needs	Project Management	Project Management
Negotiates with clients before the evaluation begins	Project Management	Project Management
Responds to requests for proposals	Project Management	Project Management
Pursues professional development in evaluation	Reflective Practice	Professional Development
Pursues professional development in relevant content areas	Reflective Practice	Professional Development
Reflects on personal evaluation practice	Reflective Practice	Professional Development
Builds professional relationships to enhance evaluation practice	Reflective Practice	Professional Development

Table 5.2 depicts how items from the six original conceptualized dimension are included in the empirically derived five subscales. The evaluative practice subscale contains predominantly items from original situational analysis dimension, as well as items from systematic inquiry and interpersonal competence. The meta-competencies subscale contains items from five of the six originally-conceptualized dimensions, except systematic inquiry. All items in the knowledge base, project management, and professional development subscales are from the originally-conceptualized dimensions. Specifically, all items in knowledge base subscale are from systematic inquiry dimension; all items in project management subscale are

from project management dimension, and all items in professional development subscale are from reflective practice.

Psychometric property examinations showed the final scale had a strong internal consistency with an alpha coefficient of .909, and all subscales achieved reasonable internal consistency (> .70). Overall, CFA analyses provided the support for a five-factor solution for the perceived importance and self-assessed scales, and both measurement models achieved excellent fit.

Evaluation Practice

The initial evaluation practice scale consists of eight aspects of practice, and each contains a various number of factors. The original research (Shadish & Epstein, 1987) observed 22 factors from eight separate first-order EFA analyses and subsequent second-order analysis using factor scores on the first-order factors yielded four distinct evaluation practice patterns. In the exploratory phase, the current study replicated the analytical methodology and discovered 17 first-order factors in the eight separate EFAs. Instead of utilizing factor scores as the original study, this study took the composite score per factor approach because of the factor score indeterminacy and yielded four distinct practice patterns, which to a certain extent are consistent with the findings in the previous study.

Shadish and Epstein (1987) adopted the Kaiser (Eigenvalues > 1) approach and extracted as many first-order factors as possible without eliminating any items. The results from this over-extraction of the first-order factors supposedly offset the results of the second-order analysis. Although the overfactoring approach produces fewer errors and less inaccurate estimates than under factoring, Fabrigar et al. (1999) recommended avoiding overfactoring because “solutions

with too many factors might prompt a researcher to postulate the existence of constructs with little theoretical values and thereby develop unnecessarily complex theories" (p. 278). Furthermore, they argued that overfactoring compounded with the use of inappropriate extraction method, principal component extraction, in this case, may accentuate minor factors to appear to be major factors by inflating factor loadings and subsequently produce a false factor solution.

In the present study, however, the elimination of items with low item-total correlations and communalities made the extraction of major factors possible. Comparing the four practice patterns resulted in the original and current studies, two practice patterns—*academic and stakeholder-service* patterns—are present in both studies. In the *academic* pattern, evaluators decided to evaluate because of their basic science interest. Therefore, the dependent variables and question sources in the evaluation are often selected on the basis of research and theoretical foundations. The evaluation in academic practice pattern tends to use quantitative methods. Similarly, for the *stakeholder-service* pattern, evaluators tend to select evaluation questions and decide on dependent variables with stakeholder needs in mind. In addition, evaluators often assume team-oriented roles.

On the other hand, for *method-driven* practice pattern, evaluators utilize all different evaluation methods and gather a wide variety of data, to measure program effects, improve program performance, or judge program values. For the fourth, *use-driven* practice pattern, evaluators engage in all activities to facilitate evaluation use, and often adopt qualitative methods as their primary evaluation method.

Compatible with the findings in Shadish and Epstein (1987) study, while most inter-factor correlations are small, the academic practice pattern has the most substantial correlation

(.545) with the *method-driven* pattern, which implies that evaluators taking on academic practice pattern tend to inject methodological rigor in pursuing evaluative truth.

In interpreting these practice patterns, Shadish and Epstein (1987) cautioned that the practice patterns in their study aimed to characterize evaluations rather than evaluators. They argued that evaluators might engage in any one or more of these practice patterns in conducting their evaluations. They also suggested that the four-pattern interpretation might oversimplify the nature of practice due to methodological limitations. There is a major difference in the unit of analysis between the present study and Shadish and Epstein (1989) study. While the unit of analysis in Shadish and Epstein study is the most recent evaluation conducted, the unit of analysis in the present study is the evaluators' common practice patterns in their most recent 3-10 evaluations. In order to uncover evaluators' practice patterns in this study, evaluators were made aware of reflecting on their evaluation practices before responding to the questions. The Likert scale adopted in the current study, directly addressing the frequency of different aspects of evaluators' practices, as such, is also more appropriate.

Discussions, Limitations, and Implications for Future Research

Lack of Variance in Responses to the Perceived Importance of the ECPE

One concern emerged in the data analysis was how the lack of variability in responses to the ECPE importance ratings would affect the construct validity and psychometric properties of the scale. Primarily, respondents were requested to rate the importance of 61 evaluator competencies on a 7-point Likert scale. The results showed a restricted pattern on the range of item means (5.31 – 6.90) except item 26 "conducts meta-evaluations" (M = 4.49). After recoding data to a 5-point Likert scale to reduce skewness and kurtosis, the restriction was slightly

improved with item means ranging from 3.34 to 4.91 with the only exception of item 26 having a lower mean (M = 2.62). Even though the restricted response pattern made logical sense, as the previous research had already excluded unimportant competencies from the set of 61 competencies, there might be effects on the strength of the observed correlations and the quality of model fit. Despite the lack of response variability, the measurement model nevertheless exhibited a good model fit and reasonable psychometric properties.

Cross-validation

In the current study, CFA analyses for ECPE importance rating and EP scale were performed on the same data set as EFA analyses. When testing whether the factor structure resulted from EFA are consistent with that in CFA, experts recommend cross-validation using an independent sample (Byrne, 1989; Jöreskog & Sörbom, 1989). Research literature shows that in many instances EFA factor structures could not be confirmed by CFA with independent samples. Van Prooijen & Van Der Kloot (2001) contend that the approach using independent samples makes it difficult to explain the factor structure discrepancy between EFA and CFA. They presented three methodological explanations for inconsistent factor structures results in EFA and CFA: a) inadequate application of EFA; b) methodological differences between EFA and CFA; and c) inappropriate application of CFA.

First, many EFA studies may not be properly carried out by selecting inappropriate criteria to determine the number of factors to be extracted, inappropriate rotation method, and inappropriate procedures such as estimation methods. Next, while EFA is a data-driven analytic method, CFA is a theory-driven one. The difference can be reflected in how the parameters are set up in the models. In EFA, indicators are allowed to load freely on all factors and often load

on multiple factors with different strengths. However, in CFA, each indicator loads on its target factor, and cross-loadings are usually fixed to zero. Consequently, indicators that are free to load on all factors in EFA are hence constrained to their target factors without any cross-loadings in CFA. Because of this constraint, CFA is more parsimonious and conservative than EFA. At the same time, this may also lead to the CFA model misfit. Lastly, modifications are often made to improve CFA model fit, and the resultant final measurement model could be significantly different from the original hypothesized model in EFA.

When conducting CFA in a different sample to cross-validate EFA factor structure, it is challenging to differentiate whether the misfit can be explained by any of the three methodological possibilities. Therefore, Van Prooijen & Van Der Kloot (2001) suggested that a cross-validation process should be carried out with the same data set to account for lacking inconsistency between EFA and CFA factor structure. They concluded that "if a good fit is questionable when the factor structure is confirmatively tested on the same data, we cannot expect that a test of the factor structure in a confirmatively follow-up study, that is, on different data, will lead to a good fit" (p.790). In discussing cross-validation, Kline (2016) agreed that the discrepancy between EFA and CFA results could be explained by the condition of the area of research. For less-researched areas, factor structure derived in EFA may not be ready for CFA because it is more restricted. Additional EFAs in different samples are more appropriate to confirm the factor structure than CFA.

Measurement Invariance

As a critical psychometric property, measurement invariance tests whether the factor structure of a scale is consistent across heterogeneous population groups. This study utilized

MIMIC modeling approach and showed that both ECPE and EP scales were partially invariant across some of eight groups. Although MIMIC modeling has numerous advantages including small sample requirements, ease to be carried out, ability to test differential item functioning (DIF), and ease of accommodating a large number of covariates, only two measurement parameters, factor means and indicator intercepts, can be tested (Brown, 2015). Consequently, MIMIC modeling assumes the invariance of other model parameters, such as factor loadings, factor variances, factor covariances, and residual variances. Multiple-group modeling approach, however, has the ability to test all aspects of measurement invariance. For future research, multiple-group CFA should be carried out to examine all aspects of measurement invariance with larger sample sizes.

Self-reporting & Self-Assessment Data

The self-reporting nature of the data utilized in the study may invite debates on the issues of data reliability. However, as Chan (2009) argued, a self-report measure is the only possible way to conduct certain studies on certain topics. The argument fit the research context of the present study. Additionally, Chan further addressed the criticism and reservation of using self-report data as "urban legend," and provided a comprehensive analytical review.

Chan proposed that the common beliefs on the biased nature of self-report data could be attributed to common method variance in the form of measurement error described in Campbell & Fiske (1959). When the relationship between constructs is measured with the same method (self-report), the results may be biased due to the shared variance attributed to method effects, also known as measurement error impacting the accurate estimation of true relationships between constructs in the study. Chan also pointed out that studies on methods effects produced

considerably different conclusions on the effects of self-report measures. Additionally, four common misconceptions were summed up regarding self-report data.

- *Construct validity of self-report data.* Chan contended that that construct validity of self-report data is often questioned due to the inert susceptibility to systematic influences such as question wording and orders in the measurement instrument. However, he argued that not all self-report instruments suffer from systemic bias. There are many well-established self-report measurement scales such as the Big-Five personality traits.
- *Interpretation of correlations in self-report data.* Researchers argue that self-report data often fails to estimate parameters in question accurately. According to Chan, this is related to the common method variance that tends to inflate the estimation of correlations. He further demonstrated that this inflation might be a possibility, not a necessity.
- *Social Desirability*. Chan contends that not all self-report measures are susceptible to social desirability. Many factors contributed to socially desirable responses, such as item content, item wording, test instruction, or high-stakes or not.
- *Value of data collected from non-self-report measures*. Chan also pointed out that non-self-report measures may suffer a similar set of problems as self-report measures, such as artificially inflated or deflated correlations.

Contrasting with self-reporting, Kaslow and colleagues (2009) suggest that self-assessment might be a helpful tool in assessing professional competencies. They define self-assessment as a “process by which the person being assessed validly ascertains personal and professional strengths and areas in need of improvement across foundational and functional

competency domains, raises awareness of own limits of expertise and determines what to do when those limits are reached, and monitors own progress in the process of taking action to address specific developmental needs" (p S39). The implementation of successful self-assessments hinges on proper training of the person conducting self-assessment and his understanding of the rationale and the self-assessment methodology. The strengths of using self-assessment in competency assessment include increases in self-knowledge on the level of competency achievement and promotion of self-reflection. However, adopting self-assessment is challenging because of the difficulty of training the person to conduct an accurate self-appraisal without providing points of reference. People with lower competencies tend to inflate the results of self-assessment. This inflation leads to the questioning of the accuracy of the information. Future studies are recommended to use self-assessment as a supplementary method in addition to others.

Reflective and Formative Measurement Models

When developing measurements for constructs, researchers have to choose which measurement approach is appropriate, reflective or formative. While social science constructs are often conceptualized as a reflective measurement, a formative approach sometimes may be more appropriate (Hair, Hult, Ringle, & Sarstedt, 2016). A reflective measurement model, also known as a scale, assumes that the construct represents the commonality of the indicators, so the causal relationship is from the construct to indicators, also considered as effect indicators. On the other hand, a formative measurement model, also known as an index, considers that the indicators form the construct, and hence the construct is a composite of all indicators. Therefore, the causal relationship is from indicators to the construct.

Misspecification of measurement may have a serious effect on not only how the construct is conceptualized, but also potentially how the construct is operationalized (Diamantopoulos & Siguaw, 2006). In choosing measurement approaches, researchers should base their decisions on the auxiliary theory of the construct. While the decision can be straightforward in some cases, it will not be in others. Diamantopoulos and Siguaw hypothesized that "the probability of *erroneously* selecting a reflective perspective (and thus committing a Type I error) is currently much higher than the corresponding probability of erroneously opting for a formative perspective (Type II error)" (p 266). They compared formative and reflective approaches in three stages of measurement development, item generation, measure purification and measure validation. While Type I error would not be an issue in the initial item generation stage because there are no substantive differences suggested by conceptual guidelines, it becomes plausible in the purification and validation stages to commit errors. In the purification stage, for instance, the criteria to include or exclude items are completely opposite for the two approaches. For reflective measurement models, high inter-item correlations are desirable and indicative of high internal consistency. For formative models, however, lower intercorrelations are desirable and indicative of lacking redundancy. Also, reflective indicators are considered interchangeable, and elimination of any indicators does not change the conceptualization of the construct. Nevertheless, formative indicators are not required to correlate, and elimination of any indicators does alter the construct itself.

Even though the difference between formative and reflective measurement models has been well established (MacCallum & Browne, 1993), there is a lack of consensus on how to effectively choose the correct measurement approach (Edwards & Bagozzi, 2000; Howell, Breivik, & Wilcox, 2007; Bagozzi, 2007). To address this issue, Coltman, Devinney, Midgley,

and Venaik (2008) built upon the ideation by Diamantopoulos and Siguaw (2006) and offered a set of theoretical and empirical considerations. Their framework included three theoretical considerations: the nature of the construct, the direction of causality between items and latent construct, and characteristics of items used to measure the construct. In addition, three empirical considerations were also included, item inter-correlation, item relationships with construct antecedents and consequences, and measurement error and collinearity. The framework put forward by Coltman et al. (2008, p 1252) includes specific comparisons between the two measurement approaches on each of the considerations.

The framework is particularly applicable to the present study, in exploring which measurement approach is the most appropriate. For example, the low item inter-correlations, according to the framework, are indicative of formative measurement approach. Future research should explore how the ECPE and the EP can be modeled as formative constructs and how the final items can produce different results from the present study.

Implication on research on evaluation (RoE)

There has been an upward research trend on RoE in recent years (Coryn, Noakes, Westine, & Schroter, 2011; Galport & Galport, 2015; Lewis, Harrison, G. M., Ah Sam, & Brandon, 2015; Vallin, Philippoff, Pierce, & Brandon, 2015; Coryn, et al., 2016; Galport & Azzam, 2017). Smith (2015) argues that, although current RoE takes a wide variety of approaches and methods, many share similar limitations that can be grouped in four areas of definition, focus, evidence, and inference. As the present study aims to contribute to RoE, it is inherently bound by some of the limitations outlined. However, the study has also overcome some of the limitations and demonstrated better practices advocated by Smith (2015).

- *Definition.* Whereas Smith (2015) pointed out that the lack of a precise and comprehensive definition for evaluation practice has been one of the common weaknesses shared by recent RoE, the present study aimed to operationalize evaluation practice, and adopted the evaluation practice scale (Shadish & Epstein, 1987), which has provided a highly comprehensive operational definition for evaluation practice from eight aspects (evaluation purposes, reported influences on evaluation decisions, evaluator source of evaluation questions and issues, central issues, sources of dependent variables for program effectiveness questions, evaluation methods used, and reported activities to facilitate use). The present study has established that the evaluation practice is a multidimensional and hierarchical construct.
- *Focus.* Regarding focus, Smith (2015) contended that many RoE studies have not been able to address the complexity of evaluation settings and the complexity of evaluation practice process. Furthermore, these studies often focus on evaluators' common practices rather than individual practice and practice variations. Although the purpose of the present study is to uncover and confirm evaluators' common practice patterns, as Smith identified, the adoption of the evaluation practice scale and analytical procedures in the study have partially mitigated some of the weaknesses. As addressed earlier, the evaluation practice scale (Shadish & Epstein, 1987) has 74 items in eight practice aspects providing comprehensive coverage for evaluators' practices. Additionally, the adoption of the MIMIC method has incorporated a wide variety of contextual evaluation

covariates into the study, such as evaluators' years of experience, professional affiliation, work settings, educational background, and gender.

- *Evidence.* Self-reporting nature and overuse of survey design of many RoE studies resulted in unreliable conclusions about evaluation practice (Smith, 2015). In the previous section, an extensive discussion has been offered on self-report versus self-assessment data. While the present study does rely on self-reported data, the analytical framework has increased the reliability of the study results. As an example, compared with Galport and Azzam (2017) survey study that has relied predominantly on descriptive statistics, the current study has applied the advanced quantitative method to draw inferences. Even though the present study has not met the rigorous standards and suggestions set by Smith (2015), it is definitely one of the more robust RoE studies.
- *Inference.* The fourth limitation put forward by Smith (2015) is that most RoE studies fail to implement a multi-perspective approach in drawing the inference. Assessments and verifiable evidentiary support from other evaluation stakeholders will increase the credibility of the study conclusions. Future RoE studies should explore alternative designs, such as mixed methods, to achieve the suggested level of inference.

To overcome these limitations and advance RoE, Smith suggested an alternative approach—action design research, "an iterative process of problem clarification, design, development, testing, reflection, redesign, and so on" (p. 67). And during this process, RoE researchers seek evaluative input from all key stakeholders and collaboratively reach conclusions. The example action design research process provides a useful framework for future

research on studying how evaluator competencies affect their practices. Specifically, evaluator competencies in the current study are based on a set of general competencies established in the ECPE scale and similarly in the evaluation practice scale. By incorporating action design research process, researchers can investigate what specific competencies are applicable in a particular context, and what specific practice decisions evaluators make in this evaluative context. Therefore, move RoE from a generalized view to a case-based view of evaluation practice.

Conclusion

This study examines two critical constructs in the evaluation profession, essential competencies for program evaluators and evaluation practice, regarding their construct validity, psychometric properties, and the inter-relationships. As one of the few empirical studies investigating essential evaluator competencies and evaluator practices, this study builds upon previous research and extends the understanding of a set of fundamental issues involving in the two critical constructs. First, the finding adds support for construct validity of essential competencies for program evaluators scale and evaluation practice scale, specifically, the factor structures that are confirmed empirically. Second, the study establishes psychometric properties of the two scales, including reliability and measurement invariance, to support future research. Third, the study uncovers that evaluator practice patterns and their self-assessed level of competencies have a reciprocal relationship.

On the one hand, evaluators' self-assessed level of competencies directly relates to their practice patterns, for example, the higher rating on evaluative practice competencies indicates higher proclivity of evaluators take on academic and method-driven practice patterns; One the

other hand, evaluator practice patterns also have impacts on how evaluators' self-assessed level of competencies. For instance, evaluators in academic patterns tend to report higher self-assessed competencies in areas of evaluative practice, Meta, as well as knowledge base. The results also suggest an alternative measurement approach of formative and reflective indicators, which are two different measurement perspective in operationalizing the causal relationship between latent constructs and indicators.

Measurement is at the heart of social science research, and sound measurements ensure the validity of research findings. As DeVellis (2003) states, scale development is a continuous process. As a result, the findings and recommendations outlined in this study provide input for future research, help researchers make better decisions and advance our understanding of these fundamental issues in research on evaluation.

Appendix A: American Evaluation Association (AEA) Research Approval Letter

AMERICAN EVALUATION ASSOCIATION

Jie Zhang
Syracuse University
jzhang08@syr.edu

Dear Jie Zhang,

The American Evaluation Association is pleased to approve your research request to survey the AEA membership. AEA has approved the surveying of a random sampling of 2000 AEA members in support of your dissertation. The research list is provided for use only for the research as represented to AEA and only for the example correspondence provided to AEA. The list must not be used for any other purpose, including but not limited to personal communication, and must be destroyed immediately after use. The list should be used within 30 days of provision and if for any reason cannot be used within 30 days should be destroyed and a new list requested.

Please note that your letters to members would need to include the following footer

You are receiving this email as a member of the American Evaluation Association. This research request was reviewed by a Research Request Task Force consisting of tenured AEA members. If you have concerns about the survey and would like to express them to the AEA leadership, please email info@eval.org. Any concerns raised will be shared, confidentially, with the Executive Committee of the association. AEA allows its membership list to be used infrequently for research that focuses on the field of evaluation. If you would like to opt-out of AEA's research list, please send an email request to info@eval.org. Please note that we encourage you to consider remaining on the list as such research strengthens and furthers the field's knowledge base

We wish you success in your research.

Thank you,

B Denise Roosendaal

Denise Roosendaal
Executive Director, American Evaluation Association

Appendix B: Syracuse University IRB Approval

SYRACUSE UNIVERSITY

INSTITUTIONAL REVIEW BOARD
MEMORANDUM

TO:	Nick Smith
DATE:	January 23, 2017
SUBJECT:	**Determination of Exemption from Regulations**
IRB #:	16-372
TITLE:	*A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice: An Application of Structural Equation Modeling*

The above referenced application, submitted for consideration as exempt from federal regulations as defined in 45 C.F.R. 46, has been evaluated by the Institutional Review Board (IRB) for the following:

1. determination that it falls within the one or more of the five exempt categories allowed by the organization;
2. determination that the research meets the organization's ethical standards.

It has been determined by the IRB this protocol qualifies for exemption and has been assigned to category 2. This authorization will remain active for a period of five years from **January 20, 2017** until **January 19, 2022**.

CHANGES TO PROTOCOL: Proposed changes to this protocol during the period for which IRB authorization has already been given, cannot be initiated without additional IRB review. If there is a change in your research, you should notify the IRB immediately to determine whether your research protocol continues to qualify for exemption or if submission of an expedited or full board IRB protocol is required. Information about the University's human participants protection program can be found at: http://orip.syr.edu/human-research/human-research-irb.html Protocol changes are requested on an amendment application available on the IRB web site; please reference your IRB number and attach any documents that are being amended.

STUDY COMPLETION: Study completion is when all research activities are complete or when a study is closed to enrollment and only data analysis remains on data that have been de-identified. A Study Closure Form should be completed and submitted to the IRB for review (Study Closure Form).

Thank you for your cooperation in our shared efforts to assure that the rights and welfare of people participating in research are protected.

Tracy Cromp, M.S.W.
Director

DEPT: Instructional Design, Development & Evaluation, 259 Huntington Hall **STUDENT:** Jie Zhang

Appendix C: Informed Consent

Statement of Informed Consent for Study Participants

Introduction

You are invited to participate in a dissertation study, conducted by Jie Zhang, a Ph.D. candidate from the School of Education at Syracuse University. This survey study offers an opportunity for you, as an evaluator, to share your perception of essential evaluator competencies and reflection on your professional practice. Your responses will help us achieve a better understanding of these two critical constructs in program evaluation. The study results will directly contribute to the general knowledge base for evaluation and the advancement of evaluation as a profession.

Purpose of the Research

The study has three goals: 1) to gain a better understanding of the validity of a set of proposed and tested essential professional competencies for program evaluators; 2) to investigate the patterns of evaluator practices; 3) lastly, to uncover how evaluators' self-assessed levels of competencies impact their evaluation practices.

These three issues are fundamental to the evaluation profession. With the increasing professionalization and interdisciplinary nature of evaluation, it has become extremely crucial to achieve a better understanding of these important issues. The results of the study will contribute to the advancement of the profession, provide input to the development of evaluation curriculum and training programs, facilitate new and experienced evaluators to reflect on their professional competencies and practices, and add to the general literature in certification and licensing of professional evaluators.

Procedure & Voluntary Participation

You will be responding to two sets of questions online. The first set of questions will inquire about your perceptions on a list of evaluation competencies and your self-assessed level of competencies. The second set of questions will seek your reflections on how you have conducted evaluation studies in terms of purposes, evaluation questions asked, roles assumed, methodologies adopted, use of evaluation results, and so on.

My pilot study suggested that it would take about 30 to 40 minutes to respond to all the questions. It may seem like a lengthy process, however, you will find the process truly rewarding, not only because your responses will promote a better understanding of the two critical evaluation constructs, and at the same time you will also be able to develop a better sense of your professional competencies as an evaluator and various aspects of your evaluation practice.

Your participation in this study is completely voluntary, and your decision to participate or not has no negative impacts in any way possible. You have the right to withdraw from this study at any time, and you have the right not to answer any question(s) for any reason without prejudice or penalty.

Anonymity

All data gathered in the study will be completely anonymous by using the "Anonymize Responses" function in Qualtrics, the data collection system of the study. This function enables the researcher to remove all identifying information including respondent IP addresses, emails, and names as soon as participants complete and submit their survey responses. At the same time, Qualtrics is still able to track non-respondents and allows the researcher to send future reminders to follow up with them.

Whenever one works with email or the Internet, there is always the risk of compromising privacy, confidentiality, and/or anonymity. Your confidentiality/anonymity will be maintained to the degree permitted by the technology being used. It is important for you to understand that no absolute guarantees can be made regarding the interception of data sent via the Internet by third parties.

Contact Information

If you have any questions or would like additional information about this research, please contact the researcher, Jie Zhang, directly at jzhang08@syr.edu. You can also contact my dissertation research adviser, Dr. Nick L. Smith at nlsmith@syr.edu or 315-443-3703. If you have any questions, concerns, or complaints about your rights as a research participant, and you wish to address your concerns to someone other than the investigator, or you cannot reach the investigator, please contact the Syracuse University Office of Research Integrity and Protections at orip@syr.edu or 315-443-3013.

IRB Approval

The Institutional Review Board (IRB) approval can be downloaded here.

Statement of Consent (Download the Statement of Consent here)

I verify that I am 18 years of age or older and agree to participate in this research. I understand the above statement about the research and grant the researcher permission to use my responses provided in this survey for research and publication purposes.

- o Yes
- o No

Appendix D: Data Collection Contacts

First Invitation to Study Participation

Dear ${m://FirstName} ,

My name is Jie Zhang, a Ph.D. Candidate in the Instructional Design, Development and Evaluation (IDD&E) program at Syracuse University. I would like to invite you to participate in my dissertation research studying two critical constructs in the evaluation profession, evaluator competencies, and evaluation practice.

My study aims to address three important research questions: 1) what constitutes as essential professional evaluator competencies? 2) Are there any recognizable patterns of how evaluators conduct evaluations? 3) How do evaluator competencies influence their practice patterns?

As the evaluation profession grows rapidly, it becomes extremely crucial to have a set of validated evaluator competencies and develop a better understanding of how evaluators conduct evaluations in their professional practices. The recent (2015) call from the American Evaluation Association (AEA) Board of directors on feedback on AEA draft competencies manifested the importance and timeliness of the three fundamental issues to be addressed in this study.

You will be asked to respond to questions in two scales: 1) essential competencies of program evaluation (ECPE); 2) evaluation practice (EP). The ECPE scale was adapted from a taxonomy and a series of research published by King, Stevahn and colleagues. The set of competencies identified have been the most comprehensive by far. While the content validity has been established, the construct validity has yet to be examined. The EP scale was adapted from Shadish and Epstein (1987) study of the patterns of program evaluation practice. The questions in the scale remain highly relevant despite that the research was conducted almost thirty years ago.

Your responses will be **completely anonymous** by using “Anonymize Response” function in Qualtrics survey system As soon as you complete and submit your responses, all identifying information (your name, IP address, and email) will be removed.

I hope you will take **20 - 30** minutes to join in this anonymous study because you will find the process truly rewarding. Not only will your responses promote a better quality of research on evaluation, but also you will be able to take this opportunity to reflect on your current evaluation practice.

The Syracuse University Institutional Review Board (IRB) Approval can be viewed after you start the survey by clicking the URL below.

Take the Survey

Or copy and paste the URL below into your internet browser:
${l://SurveyURL}

Thank you for your participation and valuable input!

Best regards,

Jie Zhang
Ph.D. Candidate
Instructional Design, Development & Evaluation
Syracuse University

You are receiving this email as a member of the American Evaluation Association. This research request was reviewed by a Research Request Task Force consisting of tenured AEA members. If you have concerns about the survey and would like to express them to the AEA leadership, please email info@eval.org. Any concerns raised will be shared, confidentially, with the Executive Committee of the association. AEA allows its membership list to be used infrequently for research that focuses on the field of evaluation. If you would like to opt-out of AEA's research list, please send an email request to info@eval.org. Please note that we encourage you to consider remaining on the list as such research strengthens and furthers the field's knowledge base.

Follow-up Invitation to Study Participation

Dear ${m://FirstName},

A week ago, I invited you to participate in my dissertation study on your perceptions of essential evaluator competencies and reflection on your own practice. If you are in the process of completing and submitting your responses, I would like to express my gratitude!

If you have not responded yet, I would like this opportunity to strongly encourage your participation. I understand that I have asked you to answer quite a few questions and how precious your time must be. However, your responses will be really instrumental to the validity of the study. Your responses represent and reflect your unique professional knowledge and experience, and will help to ensure the quality and completeness of the study. In turn, the findings of the study will add to our general knowledge base and the overall evaluation profession.

To show my appreciation for your participation, I will be happy to share the study findings and my dissertation upon your request as soon as the study concludes.

I sincerely hope you will take 20 - 30 minutes of your time to participate in this anonymous survey. Meanwhile, it is my hope that your thoughtful responses will also make this process a truly rewarding experience for you!

${l://SurveyLink?d=Take the survey}

Or copy and paste the URL below into your internet browser:
${l://SurveyURL}

I really appreciate your time and effort! Please don't hesitate to contact me if you have any questions or concerns. Thank you!

Best regards,
Jie Zhang
Ph.D. Candidate
Instructional Design, Development & Evaluation
Syracuse University

You are receiving this email as a member of the American Evaluation Association. This research request was reviewed by a Research Request Task Force consisting of tenured AEA members. If you have concerns about the survey and would like to express them to the AEA leadership, please email info@eval.org. Any concerns raised will be shared, confidentially, with the Executive Committee of the association. AEA allows its membership list to be used infrequently for research that focuses on the field of evaluation. If you would like to opt-out of AEA's research list, please send an email request to info@eval.org. Please note that we encourage you to consider remaining on the list as such research strengthens and furthers the field's knowledge base.

Final Invitation for Study Participation

Dear ${m://FirstName},

In the early part of the year 2017, I can imagine that you must be busy setting up new goals and kicking off new projects. Why not start your year by participating in this crucial study to contribute to the general knowledge base of the evaluation profession?

This research investigates fundamental issues of evaluator competencies and evaluation practice. By responding to the survey, you can bring your unique perspectives to these issues; and make sure that your professional experience and opinions are represented and reflected in this study.

As this study aims to be comprehensive, the results will not be genuinely representative without your input. I sincerely hope you will take 20 - 30 minutes (estimated by respondents who have already responded) to participate in this anonymous survey.

Take the Survey

Or copy and paste the URL below into your internet browser:
${l://SurveyURL}

To show my appreciation for your participation, I will be happy to share the study findings and my dissertation upon your request as soon as the study concludes.

Thank you again for your valuable input and have a productive New Year!

Best regards,

Jie Zhang
Ph.D. Candidate
Instructional Design, Development & Evaluation
Syracuse University

You are receiving this email as a member of the American Evaluation Association. This research request was reviewed by a Research Request Task Force consisting of tenured AEA members. If you have concerns about the survey and would like to express them to the AEA leadership, please email info@eval.org. Any concerns raised will be shared, confidentially, with the Executive Committee of the association. AEA allows its membership list to be used infrequently for research that focuses on the field of evaluation. If you would like to opt-out of AEA's research list, please send an email request to info@eval.org. Please note that we encourage you to consider remaining on the list as such research strengthens and furthers the field's knowledge base.

Appendix E: Study Survey Instrument

SYRACUSE UNIVERSITY

Dissertation Study: A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice

Introduction

You are invited to participate in a dissertation study conducted by Jie Zhang, a Ph.D. Candidate from Syracuse University School of Education. This study offers an opportunity for you, as an evaluator, to share your perception of essential evaluator competencies and reflection on your professional practice. Your responses will help achieve a better understanding of these two critical evaluation constructs. The study results will directly contribute to the general knowledge base for evaluation and the advancement of evaluation as a profession.

Purpose of the Research

The study has three goals: 1) the study aims to gain a better understanding of a set of previously proposed essential professional competencies for program evaluators; 2) the study also aims to examine evaluator practice patterns; 3) lastly, the study seeks to uncover how evaluators' self-assessed evaluation competencies relate and impact their evaluation practices.

These three issues are fundamental to the evaluation profession. With the increasing professionalization and interdisciplinary nature of evaluation, it has become extremely crucial to achieve a better understanding of these important issues. The results of the study will facilitate the advancement of the profession, support the development of evaluation curriculum and training programs, provide guidelines to new and experienced evaluators, and establish benchmarks for certification and licensing of professional evaluators. Practitioners can also improve their professional competencies and performances to conduct evaluation more efficiently.

Procedure & Voluntary Participation

You will be asked to respond to two sets of questions online. The first set of questions will inquire about your perceptions on a list of specific evaluation competencies and your self-assessed level of competencies. The second set of questions will seek your reflections on how you have conducted evaluation studies in terms of purposes, roles assumed, methodologies adopted, use of evaluation results, and so on.

My pilot study suggested that it would take about **30 minutes** to respond to all the questions. It may seem like a lengthy process, however, you will find the process truly rewarding, not only because your responses will promote a better understanding of the two critical evaluation constructs, and at the same time you will also be able to develop a better sense of your professional competencies as an evaluator and various aspects of your evaluation practice.

Your participation in this study is completely voluntary. Your decision whether or not to participate will have no negative impact in any way possible. You have the right to withdraw from this study at any time and you have the right to refuse to answer any question(s) for any reason without prejudice or penalty.

Anonymity

All data gathered in the study will be completely anonymous by using the "Anonymize Response" function in Qualtrics, the data collection system of the study. This function enables the researcher to remove all identifying information including respondent IP addresses, emails, and names as soon as participants complete and submit their survey responses. At the same time, Qualtrics is still able to track non-respondents and allows the researcher to send future reminders to follow up with them.

Whenever one works with email or the Internet, there is always the risk of compromising privacy, confidentiality, and/or anonymity. Your confidentiality/anonymity will be maintained to the degree permitted by the technology being used. It is important for you to understand that no absolute guarantees can be made regarding the interception of data sent via the Internet by third parties.

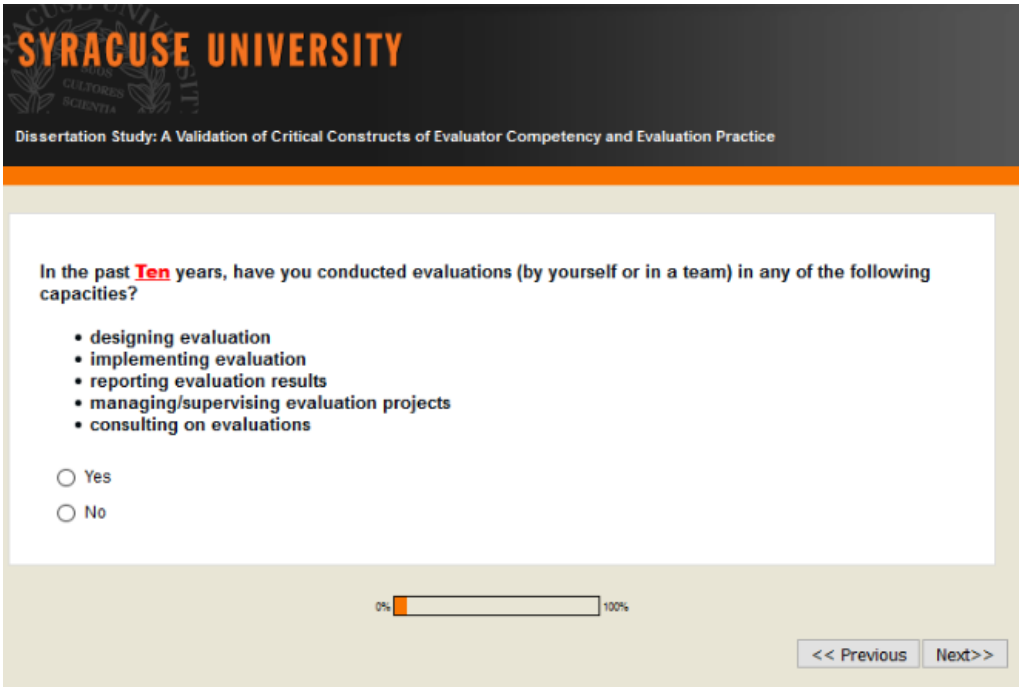
Contact Information

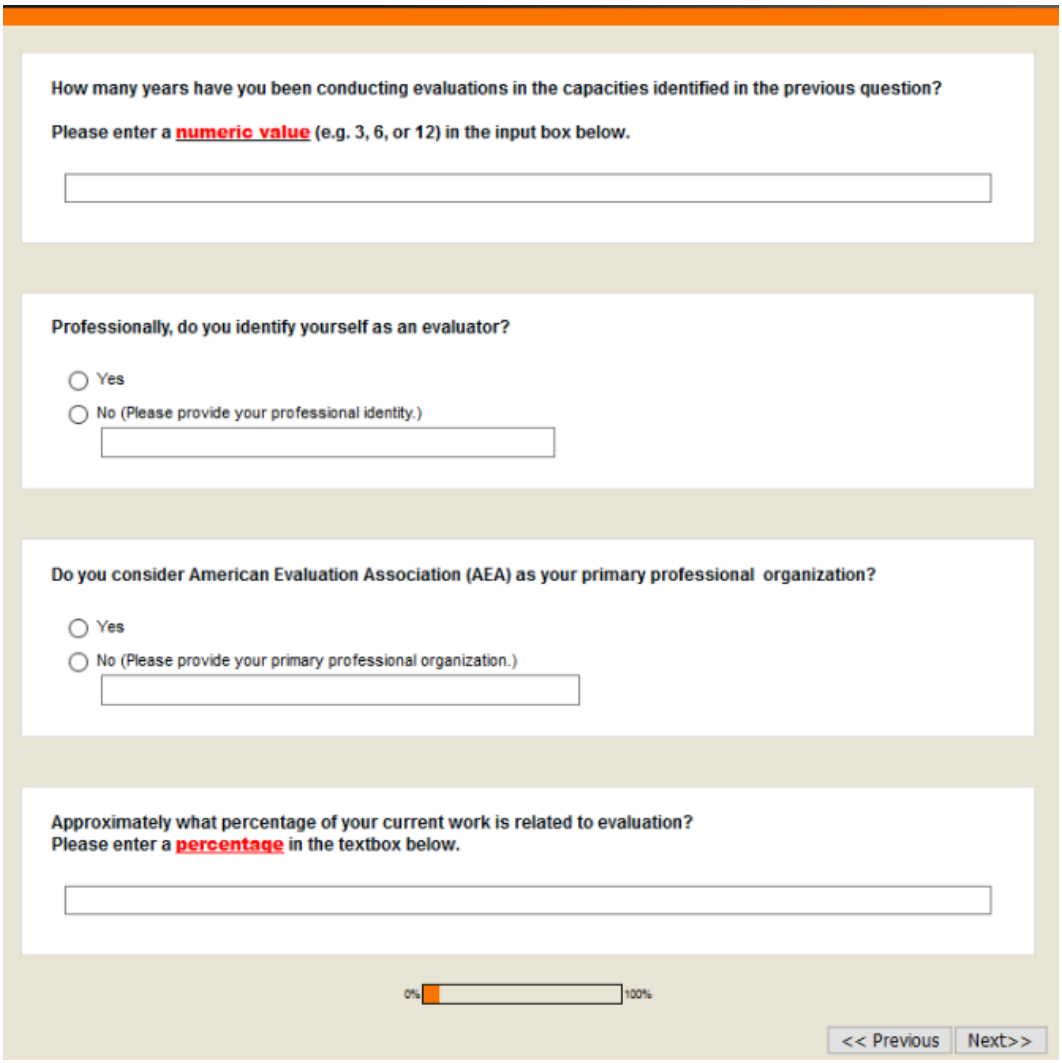
If you have any questions or would like additional information about this research, please contact the researcher, Jie Zhang, directly at jzhang08@syr.edu. You can also contact my dissertation research adviser, Dr. Nick L. Smith at nlsmith@syr.edu. If you have any questions, concerns, or complaints about your rights as a research participant, and you wish to address your concerns to someone other than the investigator, or you cannot reach the investigator, please contact the Syracuse University Office of Research Integrity and Protections at orip@syr.edu or 315-443-3013.

IRB Approval

The Institutional Review Board (IRB) approval from Syracuse University can be downloaded here.


Statement of Consent (Download the Statement of Consent.)
I verify that I am 18 years of age or older and agree to participate in this research. I understand the above statement about the research and grant the researcher permission to use my responses provided in this survey for research and publication purposes.
Yes
No
0%
100%
Next>>


SYRACUSE UNIVERSITY
Dissertation Study: A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice
In the past Ten years, have you conducted evaluations (by yourself or in a team) in any of the following capacities?
• designing evaluation
• implementing evaluation
• reporting evaluation results
• managing/supervising evaluation projects
• consulting on evaluations
Yes
No
0%
100%
<< Previous
Next>>



How many years have you been conducting evaluations in the capacities identified in the previous question?

Please enter a numeric value (e.g. 3, 6, or 12) in the input box below.

Professionally, do you identify yourself as an evaluator?

- Yes
- No (Please provide your professional identity.)

Do you consider American Evaluation Association (AEA) as your primary professional organization?

- Yes
- No (Please provide your primary professional organization.)

Approximately what percentage of your current work is related to evaluation?
Please enter a percentage in the textbox below.

0% 100%

<< Previous Next>>

SYRACUSE UNIVERSITY

Dissertation Study: A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice

Essential Evaluator Competencies

This section is consisted of questions about a set of evaluator competencies. In this context, evaluator competencies are defined as knowledge, abilities, skills, and attitudes that are essential for evaluators to conduct professional evaluations effectively.

Please share your perceptions on these essential evaluator competencies by performing two ratings. First, rate the importance of these competencies to your evaluation practice; then, provide a self-assessment of your level of competencies as objectively as possibly.

The following elaborations are provided to anchor your responses:

Importance Rating:

7 - Extremely Important: Important in practically all my evaluations
6 - Very Important: Important in most of my evaluations
5 - Somewhat Important: Important in some of my evaluations
4 - Neutral: Generally neither important nor unimportant in my evaluations
3 - Somewhat Unimportant: Not important in many of my evaluations
2 - Very Unimportant: Not important in most of my evaluations
1 - Extremely Unimportant: Not important in any of my evaluations

Level of Competency Rating:

5 - Expert: Extensive successful experience in a wide variety of applications
4 - Proficient: Considerable successful experience in most applications
3 - Intermediate: Moderately successful experience in many application
2 - Advanced Beginner: Some successful experience in a few applications
1 - Beginner: Little experience with limited success in applications

The competency of

	How important are these competencies to your evaluation practice?	What are your self-assessed levels on these competencies?
applying professional evaluation standards		
acting ethically and strive for integrity and honesty when conducting evaluations.		
conveying personal evaluation approaches and skills to potential clients.		
respecting evaluation clients, respondents, program participants, and other stakeholders.		
considering the general and public welfare in evaluation practice.		
contributing to the knowledge base of evaluation.		

0% 100%

<< Previous Next>>

SYRACUSE UNIVERSITY

Dissertation Study: A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice

The following elaborations are provided to anchor your responses:

Importance Rating:

7 - Extremely Important: Important in practically all my evaluations
6 - Very Important: Important in most of my evaluations
5 - Somewhat Important: Important in some of my evaluations
4 - Neutral: Generally neither important nor unimportant in my evaluations
3 - Somewhat Unimportant: Not important in many of my evaluations
2 - Very Unimportant: Not important in most of my evaluations
1 - Extremely Unimportant: Not important in any of my evaluations

Level of Competency Rating:

5 - Expert: Extensive successful experience in a wide variety of applications
4 - Proficient: Considerable successful experience in most applications
3 - Intermediate: Moderately successful experience in many application
2 - Advanced Beginner: Some successful experience in a few applications
1 - Beginner: Little experience with limited success in applications

The competency of

	How important are these competencies to your evaluation practice?	What are your self-assessed levels on these competencies?
building a knowledge base of evaluation (terms, concepts, theories, assumptions).		
being knowledgeable about quantitative methods.		
being knowledgeable about qualitative methods.		
being knowledgeable about mixed methods.		
conducting literature reviews.		
specifying program theory.		
being able to frame evaluation questions.		
developing evaluation design.		
identifying a variety of data sources when conducting evaluations.		
collecting data when conducting evaluations.		
assessing validity of data.		
assessing reliability of data.		
analyzing data.		
interpreting data.		
making judgments on programs being evaluated.		
developing recommendations.		
providing rationales for decisions throughout the evaluation.		
reporting evaluation design, procedures and results.		
noting strengths and limitations of the evaluation.		
conducting meta-evaluations.		

0% 100%

<< Previous Next>>

The competency of

	How important are these competencies to your evaluation practice?	What are your self-assessed levels on these competencies?
describing the program being evaluated		
determining program evaluability		
identifying the interests of relevant stakeholders		
serving the information needs of intended users		
addressing conflicts when conducting evaluations		
examining the organizational context of the evaluation		
analyzing the political considerations relevant to the evaluation		
attending to issues of evaluation use		
attending to issues of organizational change		
respecting the uniqueness of the evaluation site and client		
remaining open to input from others		
modifying the study as needed		

The competency of

	How important are these competencies to your evaluation practice?	What are your self-assessed levels on these competencies?
responding to request for proposals.		
negotiating with clients before the evaluation begins.		
writing formal agreements.		
communicating with clients throughout the evaluation process.		
budgeting an evaluation.		
justifying cost given information needs.		
identifying needed resources for evaluation, such as information, expertise, personnel, and instruments.		
use appropriating technology when conducting evaluations.		
supervising others involved in conducting the evaluation.		
training others involved in conducting the evaluation.		
managing the evaluation process in a non-disruptive manner.		
presenting work in a timely manner.		

<< Previous Next>>

The competency of

	How important is this competency to your evaluation practice?	What are your self-assessed levels on these competencies?
being aware of my own knowledge, skills, and dispositions as an evaluator.		
reflecting on personal evaluation practice (competencies and areas for growth).		
pursuing professional development in evaluation.		
pursuing professional development in relevant content areas		
building professional relationships to enhance evaluation practice		

The competency of

	How important is this competency to your evaluation practice?	What are your self-assessed levels on these competencies?
using effective written communication skills when conducting evaluations.		
usinging verbal/listening communication skills when conducting evaluations.		
performing negotiation effectively when conducting evaluations.		
resolving conflicts during evaluations		
facilitating constructive interpersonal interaction (teamwork, group facilitation, processing).		
demonstrating cross-cultural awareness/knowledge.		

0% 100%

<< Previous Next>>

In the past Ten years, approximately how many evaluations have you conducted (by yourself or in a team) in any of the following capacities?

- designed evaluation
- implemented evaluation
- reported evaluation results
- managed/supervised evaluation projects
- consulted on evaluations

Please enter ONLY a numeric value (e.g. 3, 6, or 12) in the input box below.

0% 100%

<< Previous Next>>

SYRACUSE UNIVERSITY

Dissertation Study: A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice

In this section, you will be asked to respond to a set of questions, such as evaluation purposes, evaluator roles, and methods, regarding the most recent **3 to 10** evaluation studies (**or as many as you can keep track of**) you have conducted.

Each evaluation is unique in many aspects, context, stakeholders, designs, or purposes. However, there exists a commonality or a pattern of how each evaluator approaches and conducts evaluation studies. The aim of the study is to uncover this commonality or pattern.

When responding to these questions, please think about the **general pattern** of how you have conducted your evaluations. For instance, in the past 5 evaluation studies, I "Always" assumed the role of an facilitator, and my evaluations "Rarely" had the purpose of influencing decision makers.

The following scale is used to anchor your responses:

Always: About 91% to 100% of the time
Often: About 61% to 90% of the time
Sometimes: About 31% to 60% of the time
Rarely: About 5% and 30% of the time
Never: Less than 5% of the time

Considering evaluations you completed, how frequently did your evaluations serve the following purposes?

	Always	Often	Sometime	Rarely	Never
To measure program effects	○	○	○	○	○
To improve program performance	○	○	○	○	○
To influence decision makers	○	○	○	○	○
To judge program value	○	○	○	○	○
To provide information to clients that they can use	○	○	○	○	○
To explain how program works	○	○	○	○	○
To identify solutions to social problems	○	○	○	○	○
To meet the needs of disadvantaged program clients	○	○	○	○	○
To build social science theory	○	○	○	○	○

How frequently did the following factors influence your decisions to conduct evaluations?

	Always	Often	Sometimes	Rarely	Never
Whether it can be shown how the results of evaluation would be used to change the program	○	○	○	○	○
Evaluator's interest in the program being evaluated	○	○	○	○	○
Evaluator's interest in basic research question addressable in the evaluation	○	○	○	○	○
Managers/supervisors decided to conduct the evaluation	○	○	○	○	○
The evaluations were conducted because clients paid all the expenses	○	○	○	○	○
Whether a good evaluation can be done within the budget	○	○	○	○	○
Evaluator's interest in publishing in the area	○	○	○	○	○
Whether the fiscal benefits of the evaluation would exceed its costs	○	○	○	○	○
Whether the money to be spent on evaluation could better be spent on something else	○	○	○	○	○

<< Previous Next>>

When conducting evaluations, how frequently did you assume following roles?

	Always	Often	Sometimes	Rarely	Never
A methodological expert	○	○	○	○	○
An educator of the clients	○	○	○	○	○
A facilitator of local change	○	○	○	○	○
A judge of the program	○	○	○	○	○
A sheperd for the public good	○	○	○	○	○
Part of the program team	○	○	○	○	○
An achiever working with the program manager	○	○	○	○	○
A resource for program stakeholders	○	○	○	○	○

How frequently did the following sources influence the questions that you asked in your evaluations?

	Always	Often	Sometimes	Rarely	Never
Information needs of clients who paid for the evaluation	○	○	○	○	○
Past research/evaluations	○	○	○	○	○
Evaluator's own experience about which questions are usually most important	○	○	○	○	○
Information needs of program manager	○	○	○	○	○
Information needs of program staff	○	○	○	○	○
Pending decisions on the program being evaluated	○	○	○	○	○
Social science theory	○	○	○	○	○
Pending legislation	○	○	○	○	○
Information needs of program clients	○	○	○	○	○

<< Previous Next>>

How frequently did you gather data about the following issues in your evaluations?

	Always	Often	Sometimes	Rarely	Never
Manner in which the program is actually implemented	○	○	○	○	○
Changes in service recipients brought on by the program	○	○	○	○	○
Explanation of variables that mediate the relationship between program implementation and effects	○	○	○	○	○
Number and characteristics of real and potential service recipients	○	○	○	○	○
Cost/fiscal benefits of the program	○	○	○	○	○
Changes in other people/institutions that interact with the program client	○	○	○	○	○

When judging program effectiveness, how frequently did you use the following as your evaluation criteria?

	Always	Often	Sometimes	Rarely	Never
Program goals	○	○	○	○	○
Criteria used in past evaluations of the program or similar programs	○	○	○	○	○
Criteria in relevant program regulations/legislation	○	○	○	○	○
Criteria suggested by relevant social science theory	○	○	○	○	○
Criteria selected by program managers	○	○	○	○	○
Criteria selected by the client who is paying for the evaluation	○	○	○	○	○
Criteria selected by program staff	○	○	○	○	○
Criteria selected by program clients	○	○	○	○	○
Unintended side effects	○	○	○	○	○
The needs of the disadvantaged	○	○	○	○	○

<< Previous Next>>

How frequently did you devote resources (time, personnel, money) to the following methods or activities in your evaluations?

	Always	Often	Sometimes	Rarely	Never
Inspecting program records	○	○	○	○	○
Onsite observation	○	○	○	○	○
Survey	○	○	○	○	○
Interviews with stakeholders	○	○	○	○	○
Program monitoring (e.g. Management Information system)	○	○	○	○	○
Client needs assessment	○	○	○	○	○
Constructing program theory/Theory of Change	○	○	○	○	○
Quasi-experimental design	○	○	○	○	○
Participant observation	○	○	○	○	○
Achievement tests	○	○	○	○	○
Conducting a Meta-evaluation	○	○	○	○	○
Casual modeling (e.g. Path analysis/Structural Equation Modeling)	○	○	○	○	○
Randomized experiments	○	○	○	○	○
Conducting Meta-analysis	○	○	○	○	○

How frequently did you engage in the following activities to facilitate use of evaluation results?

	Always	Often	Sometimes	Rarely	Never
Disseminate a written report of results	○	○	○	○	○
Translate results into action recommendations	○	○	○	○	○
Provide oral briefings to clients	○	○	○	○	○
Keep in frequent contact with users during the conduct of the evaluation	○	○	○	○	○
Provide interim results to clients during the evaluation	○	○	○	○	○
Ask the clients how potential evaluative information would be used to make changes	○	○	○	○	○
Identify potential users in order to include their questions in the evaluation	○	○	○	○	○
Publish results in books or journals	○	○	○	○	○
Make evaluation results available to the public in the media	○	○	○	○	○

<< Previous Next>>

What's the highest degree that you have achieved?

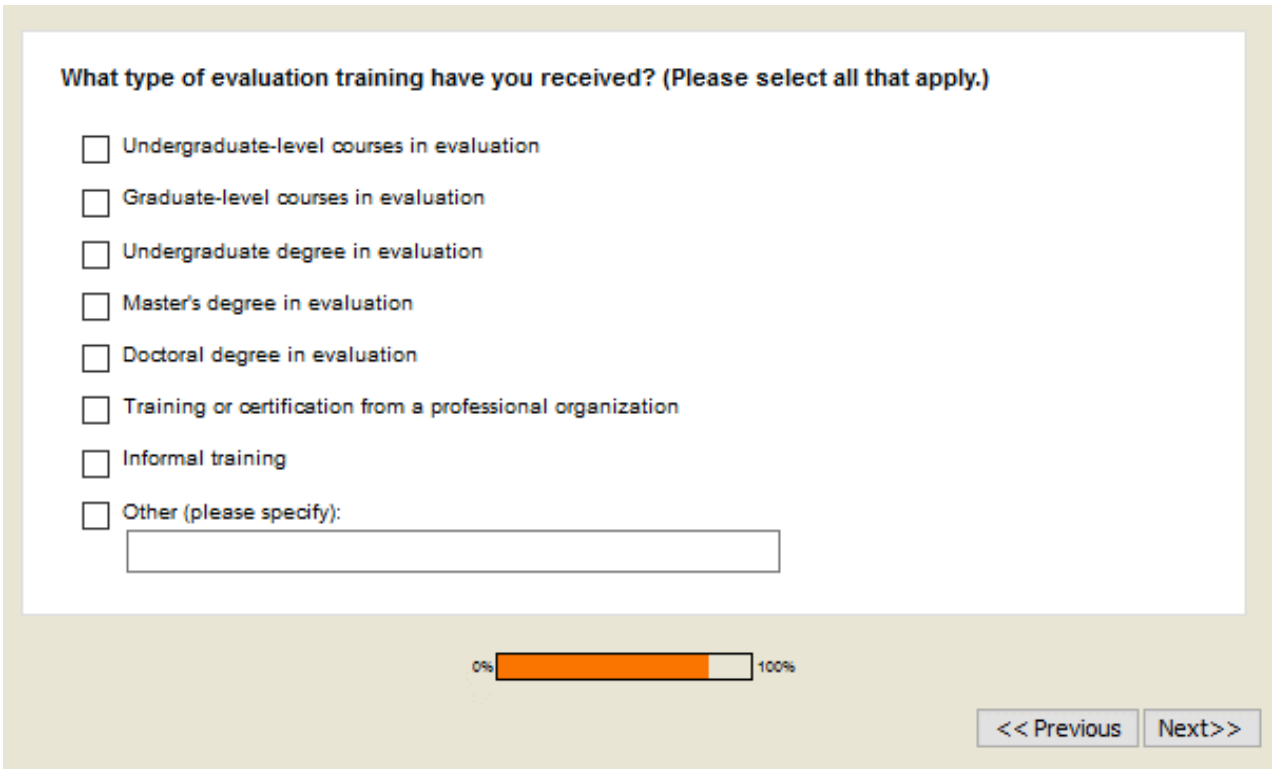
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ Doctorate degree
- ◯ Other: (Please provide your degree information)

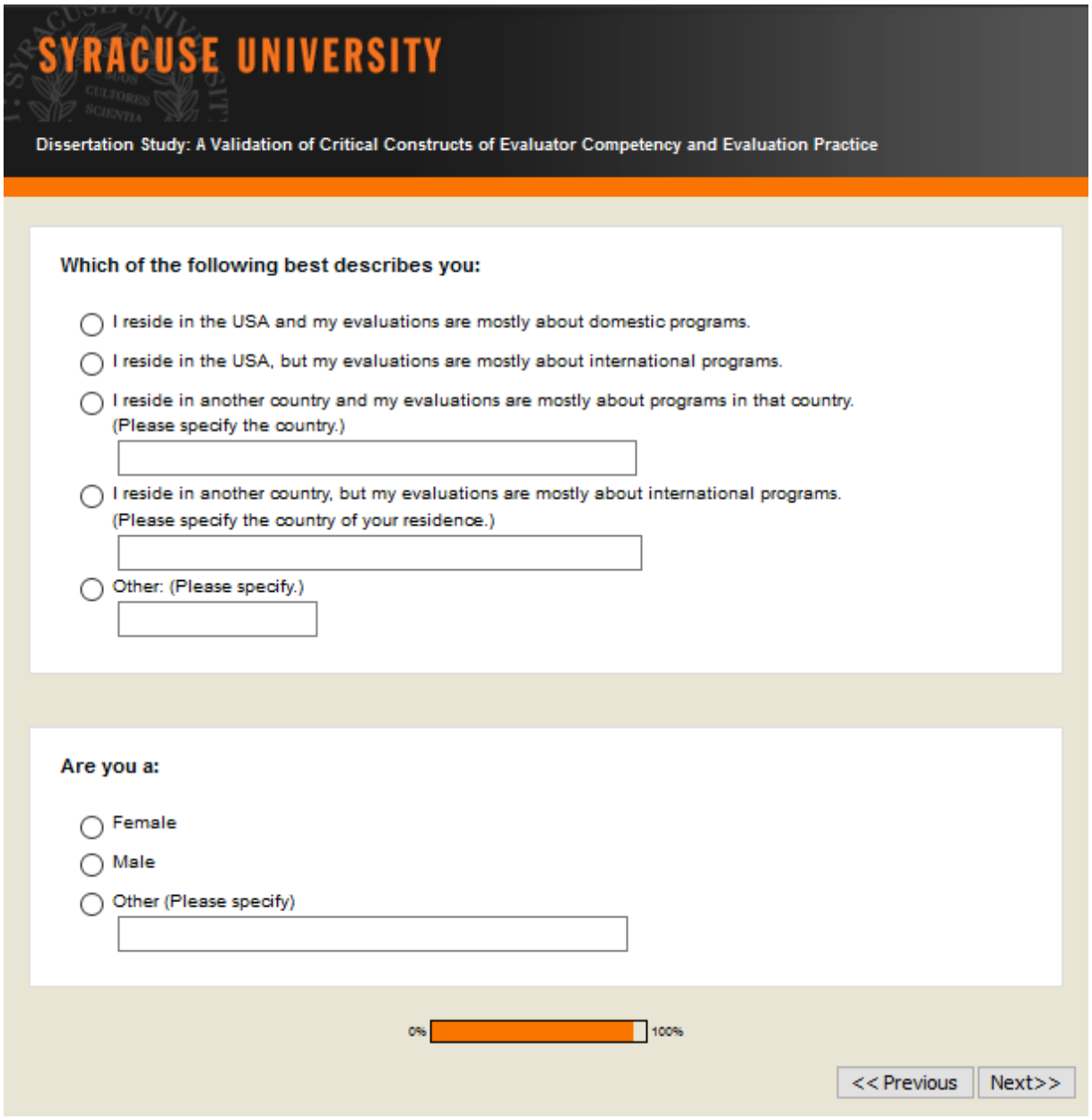
In what field did you receive your highest degree?

- ◯ Business & Economics
- ◯ Education including educational Psychology
- ◯ Evaluation
- ◯ Psychology
- ◯ Health/Public Health
- ◯ Public Policy/Public Administration
- ◯ Social work
- ◯ Sociology
- ◯ Other (please specify):

Which of the following describes your job setting?

- ◯ College/University
- ◯ Company in business and industry
- ◯ Federal government
- ◯ Local or State government
- ◯ Independent Consulting
- ◯ For-profit research, evaluation, and/or consulting firm
- ◯ Non-profit organization/foundation
- ◯ Student involved in evaluation
- ◯ Other (Please specify)


What type of evaluation training have you received? (Please select all that apply.)
Undergraduate-level courses in evaluation
Graduate-level courses in evaluation
Undergraduate degree in evaluation
Master's degree in evaluation
Doctoral degree in evaluation
Training or certification from a professional organization
Informal training
Other (please specify):
0%
100%
<< Previous
Next>>


SYRACUSE UNIVERSITY
Dissertation Study: A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice
Which of the following best describes you:
I reside in the USA and my evaluations are mostly about domestic programs.
I reside in the USA, but my evaluations are mostly about international programs.
I reside in another country and my evaluations are mostly about programs in that country.
(Please specify the country.)
I reside in another country, but my evaluations are mostly about international programs.
(Please specify the country of your residence.)
Other: (Please specify.)
Are you a:
Female
Male
Other (Please specify)
0%
100%
<< Previous
Next>>


SYRACUSE UNIVERSITY
Dissertation Study: A Validation of Critical Constructs of Evaluator Competency and Evaluation Practice
Thank you very much for participating in my dissertation study!
If you have any suggestions or comments regarding my study, I sincerely welcome you to contact me at jzhang08@syr.edu or call/text me at 571-982-6685.
0%
100%

REFERENCES

Alkin, M. C., & Ellett, F. S. (1990). Development of evaluation models. In H. J. Walberg & G. D. Haertel (Eds.), *The international encyclopedia of educational evaluation* (pp. 15-21). Oxford, Angleterre ; Toronto: Pergamon.

Alkin, M. C., & House, E. (1992). Evaluation of programs. In Alkin, M. (Ed.), *Encyclopedia of educational research*. New York: Macmillan.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Altschuld, J. W. (1999). The certification of evaluators: highlights from a report submitted to the board of directors of the American Evaluation Association. *American Journal of Evaluation, 20*(3), 481-493.

Anderson, S. B., & Ball, S. (1978). *The profession and practice of program evaluation* (1st ed.). San Francisco: Jossey-Bass Publishers.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397-438.

Azzam, T. (2007). *Evaluator contextual responsiveness.* Unpublished Doctoral Dissertation, University of California, Los Angeles, United States -- California.

Babbie, E. R. (2007). *The practice of social research*. Belmont, CA: Wadsworth.

Bagozzi, R. P. (2007). On the meaning of formative measurement and how it differs from reflective measurement: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods, 12,* 2, 229-237.

Barela, E. M. (2005). *How school district evaluators make sense of their practice.* Unpublished Doctoral Dissertation, University of California, Los Angeles, United States -- California.

Barrett, G. V., & Depinet, R. L. (1991). A reconsideration of testing for competence rather than for intelligence. *American Psychologist, 46*(10), 1012-1024.

Becker, H., & Kirkhart, K. (1981). The standards: implications for professional licensure and accreditation. *American Journal of Evaluation, 2*(2), 153-157.

Benkofske, M. T. (1996). *An examination of how evaluators and evaluation clients choose data collection methods: The factors and decision-making process.* Unpublished Doctoral Dissertation, The University of Nebraska - Lincoln, United States -- Nebraska.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural equation modeling. *Sociological Methods & Research, 16,* 78-117.

Boyatzis, R. E. (1982). *The competent manager: A model for effective performance*. New York: Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Brown, R. D., & Dinnel, D. (1992). Exploratory studies of the usefulness of a developmental approach for supervising evaluation students. *Evaluation Review, 16,* 1, 23-39.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111-150.

Brzezinski, E., & Ahn, U. (1973). *Program to operationalize a new training pattern for training evaluation personnel in education: Final report: Part A—Report on development of self-assessment of evaluation skills.* Columbus: Ohio State University.

Byrne, B. M. (1989). A *Primer of LISREL: Basic application and programming for confirmatory factor analytic models.* New York: Springer-Verlag.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: basic concepts, applications, and programming*. New York: Routledge.

Camstra, A., & Boomsma, A. (1992). Cross-Validation in regression and covariance structure analysis – an overview. *Sociological Methods & Research, 21,* 1, 89-115.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carr-Saunders, A. M., & Wilson, P. A. (1933). *The professions*. Oxford: The Clarendon Press.

Carroll, J. S., & Johnson, E. J. (1990). *Decision research: a field guide*. Newbury Park: Sage Publications.

Chan, D. (2009). So why ask me? Are self-report data really that bad? In Charles E. Lance and Robert J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*. New York, NY: Routledge.

Cheetham, G., & Chivers, G. (1996). Towards a holistic model of professional competence. *Journal of European Industrial Training, 20,* 5, 20-30.

Cheetham, G., & Chivers, G. E. (2005). *Professions, competence and informal learning*. Cheltenham, UK; Northampton, MA: Edward Elgar.

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

Christie, C. A., & Fleischer, D. N. (2010). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation, 31*(3), 326-346.

Christie, C. A., & Rose, M. (2003). Learning about evaluation through dialogue: lessons from an informal discussion group. *American Journal of Evaluation, 24*(2), 235-243.

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61(*12), 1250-1262.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, N.J.: L. Erlbaum Associates.

Coryn, L. S. C., Noakes, L. A., Westine, C. D., & Schroter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32*(2), 199-226.

Coryn, L. S. C., Ozeki, Satoshi, Wilson, Lyssa N., Greenman, Gregory D., Schröter, Daniela C., Hobson, Kristin A., . . . Vo, Anne T. (2016). Does Research on Evaluation Matter? Findings From a Survey of American Evaluation Association Members and Prominent Evaluation Theorists and Scholars. *American Journal of Evaluation, 37*(2), 159-173. doi: 10.1177/1098214015611245

Costello, A. B. & Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10,* 7, 1-9.

Davis, B. G. (1986). Overview of the teaching of evaluation across the disciplines. *New Directions for Program Evaluation, 1986*(29), 5-14.

Demarteau, M. (2002). A theoretical framework and grid for analysis of programme-evaluation practices. *Evaluation, 8*(4), 454-473.

DeVellis, R. F. (2003). *Scale development: theory and applications* (2nd ed.). Thousand Oaks, Calif.: Sage Publications Inc.

Dewey, J. D., Montrosse, B. E., Schroter, D. C., Sullins, C. D., & Mattox, J. R., II. (2008). Evaluator competencies: what's taught versus what's sought. *American Journal of Evaluation, 29*(3), 268-287.

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management, 17,* 4, 263-282.

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14, 20.

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: developing practical knowledge. In I. Shaw, J. C. Greene & M. M. Mark (Eds.), *Handbook of evaluation : policies, programs and practices* (pp. 56-75). London; Thousand Oaks, Calif.: SAGE.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: the power of human intuition and expertise in the era of computer*. New York: Free Press.

Dunn, T. J., Baguley, T., & Brunsden, V. (August 01, 2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 3, 399-412.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 2, 155-174.

Eraut, M. (1994). *Developing professional knowledge and competence*. Washington, D.C.: Falmer Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 3, 272-299.

Fitzpatrick, J. L., Christie, C. A., & Mark, M. M. (2009). *Evaluation in action: interviews with expert evaluators*. Los Angeles: Sage Publications.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use: results from a survey of U.S. American Evaluation Association members. *American Journal of Evaluation, 30*(2), 158-175.

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

Fowler, F. J. (2002). *Survey research methods*. Thousands Oaks (Calif.): Sage Publications.

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology*. Los Angeles; London: SAGE.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction*. Los Angeles: SAGE.

Gable, R. K. (1986). *Instrument development in the affective domain*. Boston, Mass.: Kluwer-Nijhoff.

Galport, N., & Azzam, T. (2017). Evaluator training needs and competencies: A gap analysis. *American Journal of Evaluation, 38,* 1, 80-100.

Galport, M., & Galport, N. (2015). Methodological trends in research on evaluation. In Paul R. Brandon (Ed.), *Research on Evaluation*. *New Directions for Evaluation, 2015,* 148, 17-29.

Ghere, G., King, J. A., Stevahn, L., & Minnema, J. (2006). A professional development unit for reflecting on program evaluator competencies. *American Journal of Evaluation, 27*(1), 108-123.

Goremucheche, R. (2017). Towards achieving consensus on essential competencies for evaluation practice in South Africa (final draft report). Retrieved from https://ucarecdn.com/f7b8f147-c5ba-49a1-9e70-93163345cd04/

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, N.J: L. Erlbaum Associates.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6,* 4, 430-450.

Gonzi, A., Hager, P., & Athanasou, J. (1993). *The development of competency-based assessment strategies for the professions, National Office of Overseas Skills recognition research paper No. 8,* Canberra, Australian Government Publishing Service.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis with readings* (4th ed.). New Jersey: Prentice Hall.

Hair, J. F., Hult, G. Tomas M, Ringle, C. M., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM).* 2nd ed. Los Angeles: SAGE.

Hauer, D. M., & Slee, E. J. (1989). A review of comprehensive examinations in selected evaluation training programs. *Evaluation Practice, 10,* 4, 20-25.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-85.

House, E. R. (1994). The future perfect of evaluation. *American Journal of Evaluation, 15*(3), 239-247.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12,* 2, 205-218.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling,* 6, 1, 1-55.

Ingle, M. D., & Klauss, R. (1980). Competency-based program evaluation: A contingency approach. *Evaluation and Program Planning,* 3, 277–287.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q Hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(1), 128-141.

Jamieson, V. E. (2009). *Evaluation practice and technology.* Unpublished Doctoral Dissertation, The Claremont Graduate University, California, United States

Johnson, T. J. (1972). *Professions and power*. London: Macmillan.

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

Jones, S. C., & Worthen, B. R. (1999). AEA members' opinions concerning evaluator certification. *American Journal of Evaluation, 20*(3), 495-506.

Jöeskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70,* 351, 631-639.

Jöreskog, K. G., & Sörbom, D. (1989). LISREL 7: A guide to the program and application. Chicago: SPSS.

Kaesbauer, S. A. M. (2012). *Teaching evaluator competencies: an examination of doctoral programs.* Unpublished Dissertation, The University of Tennessee, Knoxville.

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: Principles, advances, and applications. *The Counseling Psychologist, 34*(5), 684-718.

Kaslow, N. J., Grus, C. L., Campbell, L. F., Fouad, N. A., Hatcher, R. L., & Rodolfa, E. R. (2009). Competency Assessment Toolkit for professional psychology. *Training and Education in Professional Psychology*, 3(Suppl.), S27-S45.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Thousand Oaks, CA: Sage Publications.

Kelley, K., & Lai, K. (2011). Accuracy in parameter estimation for the Root Mean Square Error of Approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research, 46*(1), 1-32.

King, J. A., Stevahn, L., Ghere, G., & Minnema, J. (2001). Toward a taxonomy of essential evaluator competencies. *American Journal of Evaluation, 22*(2), 229-247.

Kirkhart, K. E. (1981). Defining evaluator competencies: new light on an old issue. *American Journal of Evaluation, 2*(2), 188-192.

Klemp, G. O. (1980). *The assessment of occupational competence*. Washington, DC: Report to the National Institute of Education.

Kline, R. B. (2016). *Methodology in the social sciences. Principles and practice of structural equation modeling (4th ed.).* New York, NY, US: Guilford Press.

Kundin, D. (2008). Everyday approaches to evaluation: A study of how evaluators make practice decisions. Dissertation Abstracts International, 69/01(134). (UMI No. AAT 3299404)

Kundin, D. M. (2010). A Conceptual framework for how evaluators make everyday practice decisions. *American Journal of Evaluation, 31,* 3, 347-362.

Le Deist, F. D., & Winterton, J. (2005). What is competence? *Human Resources Development International, 8*(1), 27-46.

Lewis, N. R., Harrison, G. M., Ah, Sam, A. F., & Brandon, P. R. (2015). Evaluators' Perspectives on Research on Evaluation: Evaluators' Perspectives on Research on Evaluation. In Paul R. Brandon (Ed.), *Research on evaluation. New Directions for Evaluation, 2015,* 148, 89-102.

Light, R. J. (1995). The future for evaluation. *American Journal of Evaluation, 15*(3), 249-253.

Love, A. J. (1994). Should evaluators be certified? *New Directions for Program Evaluation, 1994*(62), 29-40.

Lucia, A. D., & Lepsinger, R. (1999). *The art and science of competency models: pinpointing critical success factors in organizations.* San Francisco, Calif.: Jossey-Bass/Pfeiffer.

Lynch, K. B. (1988). Evaluation practices of educational programs reviewed by the Joint Dissemination Review panel, 1980-1983. *Evaluation Review, 12*(3), 253-275.

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114(*3), 533-541.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods Psychological Methods, 1*(2), 130-149.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*(4), 611-637.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample Size in Factor Analysis. *Psychological Methods, 4*(1), 84-99.

Maicher, B., Kuji-Shikatani, K., & Buchanan, H. (2009). *A professional designation for evaluators: How we got there*. Paper presented at the DEG Conference.

Mark, M. M. (2001). Evaluation's future: furor, futile, or fertile? *American Journal of Evaluation, 22*(3), 457-479.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." *The American Psychologist, 28*(1), 1-14.

McKechnie, L., & Pettigrew, K. E. (2002). Surveying the use of theory in Library and Information Science research: a disciplinary perspective. *Library Trends, 50*(3), 406-417.

Mertens, D. M. (1994). Training evaluators: Unique skills and knowledge. *New Directions for Program Evaluation, 1994*(62), 17.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: The American Council on Education/Macmillan.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2014). *Applied multivariate research: Design and interpretation*. Los Angeles: SAGE.

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation, 31*(3), 390-399.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation, 27*(3), 296-319.

Mills, G. E., & Gay, L. R. (2016). *Educational research: Competencies for analysis and applications*. Boston, MA, US: Pearson.

Morell, J. A., & Flaherty, E. W. (1978). The development of evaluation as a profession: Current status and some predictions. *Evaluation and Program Planning, 1*(1), 11-17.

Morin, A. J. S., & Maïano, C. (2011). Cross-validation of the short form of the physical self-inventory (PSI-S) using exploratory structural equation modeling (ESEM). *Psychology of Sport and Exercise, 12*(5), 540-554.

Muthén, L. K., & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599-620.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide: Statistical analysis with Latent variables*. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures : issues and applications*. Thousand Oaks, Calif.: Sage Publications.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287-293.

Noar, S. (2003). The Role of Structural Equation Modeling in Scale Development. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(4), 622-647.

Nunnally, J. C. (1978). *Psychometric theory* (2d ed.). New York: McGraw-Hill.

Owen, J. M. (2007). *Program evaluation: Forms and approaches*. New York: Guilford Press.

Parry, S. B. (1998). Just what is a competency? (And why should you care?). *Training, 35*(6), 58-64.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: the use of factor analysis for instrument development in health care research*. Thousand Oaks, Calif.: Sage Pub.

Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://www.quantpsy.org/rmsea/rmsea.htm.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research, 48*, 1, 28-56.

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation, 29*(4), 443-459.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Reigeluth, C. M., & Carr-Chellman, A. A. (2009). *Instructional-design theories and models. Vol. 3, Building a common knowledge base*. New York; London: Routledge.

Richey, R., Fields, D. C., Foxon, M., Roberts, R. C., Spannaus, T., & Spector, J. M. (2001). *Instructional design competencies: the standards* (3 ed.). Syracuse, N.Y.: Eric Clearinghouse on Information & Technology, Syracuse University.

Riffe, D., Lacy, S., & Fico, F. (2005). *Analyzing media messages: Using quantitative content analysis in research*. Mahwah, N.J.: Lawrence Erlbaum.

Roberts, N., Thatcher, J. B., & Grover, V. (2010). Advancing operations management theory using exploratory structural equation modeling techniques. *International Journal of Production Research, 48*(15), 4329-4353.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study. *Social Work Research, 27*(2), 94-104.

Russ-Eft, D. (1995). Defining Competencies: A critique. *Human Resource Development Quarterly, 6*(4), 329-336.

Russ-Eft, D. F. (2008). *Evaluator competencies: Standards for the practice of evaluation in organizations*. San Francisco: Jossey-Bass.

Sanders, J. R. (1986). The teaching of evaluation in education. *New Directions for Program Evaluation, 1986*(29), 15.

Satorra, A. and Bentler, P. M. 1988. "Scaling corrections for chi-square statistics in covariance structure analysis." In *ASA 1988 Proceedings of the Business and Economic Statistics Section*, 308–313. Alexandria, VA: American Statistical Association.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 1, 83-90.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schwandt, T. A. (1997). Evaluation as practical hermeneutics. *Evaluation, 3*(1), 69-83.

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

Schwandt, T. A. (2005). The centrality of practice to evaluation. *American Journal of Evaluation, 26*(1), 95-105.

Schwandt, T. A. (2008). The relevance of practical knowledge traditions to evaluation practice. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 29-40). New York: Guilford Press.

Scriven, M. (1991). *Evaluation thesaurus*. Newbury Park, Calif.: Sage Publications.

Scriven, M. (1995). The logic of evaluation and evaluation practice. *New Directions for Evaluation, 1995*(68), 49-70.

Scriven, M. (1996). Types of Evaluation and Types of Evaluator. *American Journal of Evaluation, 17*(2), 151-161.

Sechrest, L. (1980). Evaluation researchers: Disciplinary training and identity. *New Directions for Program Evaluation, 1980*(8), 1-18.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2015). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.

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

Shadish, W. R., & Epstein, R. (1987). Patterns of program evaluation practice among members of the Evaluation Research Society and Evaluation Network. *Evaluation Review, 11*(5), 555-590.

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *American Journal of Evaluation, 18*(1), 195-208.

Smith, M. F. (1999). Should AEA begin a process for restricting membership in the profession of evaluation? *American Journal of Evaluation, 20*(3), 521-531.

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *American Journal of Evaluation, 14*(3), 237-242.

Smith, N. L. (2008). Fundamental issues in evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 1-23). New York: Guilford Press.

Smith, N. L. (2010). Characterizing the Evaluand in Evaluating Theory. *American Journal of Evaluation, 31*(3), 383-389.

Smith, N. L., & Brandon, P. R. (2008). *Fundamental issues in evaluation*. New York: Guilford Press.

Smith, N. L. (2015). Using Action Design Research to Research and Develop Evaluation Practice. In Paul R. Brandon (Ed.), *Research on evaluation. New Directions for Evaluation,* 148, 57-72.

Spencer, L. M., McClelland, D. C., & Spencer, S. M. (1994). *Competency assessment methods: history and state of the art*. Boston, MA: Hay/McBer Research Press.

Spencer, L. M., & Spencer, S. M. (1993). *Competence at work: Models for superior performance*. New York: Wiley.

Stake, R. E. (2004). *Standards-based & responsive evaluation*. Thousand Oaks, Calif.: Sage.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation, 26*(1), 43-59.

Steiger, J. H. (1996). Coming full circle in the history of factor indeterminacy. *Multivariate Behavioral Research, 31*(4), 617-630.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: a practical guide to their development and use* (4th ed.). Oxford; New York: Oxford University Press.

Stufflebeam, D. L. (2001). *Evaluation models*. San Francisco: Jossey-Bass.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications* (1st ed.). San Francisco: Jossey-Bass.

Stufflebeam, D. L., & Wingate, L. A. (2005). A self-assessment procedure for use in valuation training. *American Journal of Evaluation, 26*(4), 544-561.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson/Allyn & Bacon.

Tourmen, C. (2009). Evaluators' decision making: The relationship between theory, practice, and experience. *American Journal of Evaluation, 30*(1), 5-6.

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

Van Prooijen, J.-W., & van der Kloot, W. A. (2001). Confirmatory analysis of exploratively obtained factor structures. *Educational and Psychological Measurement, 61,* 5, 777-792.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester, West Sussex UK: Wiley.

Wehipeihana, N., Bailey, R., Davidson, E. J., & McKegg, K. (2014). Evaluator competencies: The Aotearoa New Zealand experience. *Canadian Journal of Program Evaluation*, 28, 3, 49-70

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). New York: Guilford Press.

Wilcox, Y. (2012). An initial study to develop instruments and validate the Essential Competencies for Program Evaluators (ECPE). Unpublished Dissertation, University of Minnesota.

Wilcox, Y., & King, J. A. (2014). A professional grounding and history of the development and formal use of evaluator competencies. *Canadian Journal of Program Evaluation, 28,* 3, 1-28.

Williams, J. E. (1989). A numerically developed taxonomy of evaluation theory and practice. Evaluation Review, 13(1), 18-31.

Witkin, B. R. (1994). Needs assessment since 1981: The state of the practice. American Journal of Evaluation, 15(1), 17-27.

Wolf, E. J., Harrington, K. M., Miller, M. W., & Clark, S. L. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73,* 6, 913-934.

Worthen, B. R. (1975). Competencies for educational research and evaluation. *Educational Researcher, 4*(1), 13-16.

Worthen, B. R. (1994). Is evaluation a mature profession that warrants the preparation of evaluation professionals? *New Directions for Program Evaluation, 1994*(62), 3.

Worthen, B. R. (1999). Critical challenges confronting certification of evaluators. *American Journal of Evaluation, 20*(3), 533-555.

Worthen, B. R. (2001). Whither evaluation? That all depends. *American Journal of Evaluation, 22*(3), 409-418.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: alternative approaches and practical guidelines.* New York: Longman.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, 17, 2, 253-269.

Vita

Jie Zhang

CORE COMPETENCIES

Research Design: ▪ Quantitative & Qualitative Data Analysis ▪ Survey Design ▪ Sampling Techniques ▪ Research Instrument/Questionnaire Development ▪ Data Collections	**Instructional Design:** ▪ Curriculum design & development ▪ Learning assessment design ▪ Instructional content authoring & management ▪ Instructional design theory & learning theory ▪ Design & delivery of professional training
Quantitative & Psychometric Analysis: ▪ Latent trait measurement models: Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), Item Response Theory (IRT) ▪ Longitudinal/Multilevel Analysis ▪ Psychometrics: IRT, Differential Item Functioning (DIF) ▪ Structural Equation Modeling (SEM)	**Program Evaluation:** ▪ Program Evaluation Design & Implementation ▪ Development of Logic Model/Program Theory ▪ Evaluation Approaches & Models ▪ Educational Evaluation Standards
Analytics: ▪ Quantitative Analysis: Mplus, R, SAS, SPSS ▪ Qualitative Analysis: Nvivo, QDA Miner ▪ Graphics & Reporting: R lattice & ggplot2, Tableau ▪ Survey System: Qualtrics	**Project Management:** ▪ Effective communication management ▪ Lead organizational change ▪ Strategic negotiation ▪ Cross-cultural communication ▪ Creative & strategic problem-solving

EDUCATION

Ph.D. Instructional Design, Development & Evaluation (IDD&E), Syracuse University, 2019
M.S. Instructional Design, Development & Evaluation, Syracuse University, 2014
M.S. Applied Statistics, Syracuse University, 2013

SELECTED PUBLICATIONS & PRESENTATIONS

Zhang, J., & Smith, N. L. (2008). *An analysis of how evaluators are empirically testing evaluation methods*. Paper presented at 2008 American Evaluation Association (AEA) International Conference (November 5-8, 2008), Denver, CO.

Zhang, J. (2010). *Constructing a Measure for Evaluator Competencies: Exploratory and Confirmatory Factor Analyses Approach.* Paper presented at 2010 American Evaluation Association (AEA) International Conference (November 11-13, 2010), San Antonio, TX.

Topchyan, R. & Zhang, J. (November 24, 2014). Validation of Virtual Learning Team Competencies for Individual Students in a Distance Education Setting. *American Journal of Distance Education*, 28, 4, 264-279. DOI: 10.1080/08923647.2014.958909

Long-Tolbert S., Zhang J. (2017) An Exploratory Study of Consumer Price Mastery Index: A Self-Efficacy Perspective (An Abstract). In: Stieler M. (Eds) *Creating Marketing Magic and Innovative Future Marketing Trends. Developments in Marketing Science: Proceedings of the Academy of Marketing Science*. Springer, Cham

Zhang, J. & Dempsey, P. R. (2018). Exploration and confirmation of a reflective-thinking scale to measure transformative learning in online courses. *Assessment and Evaluation in Higher Education*. DOI: 10.1080/02602938.2018.1520194

Dempsey, P. R. & Zhang, J. (Forthcoming). Re-examining the Construct Validity and Causal Relationships of Teaching, Cognitive and Social Presence in Community of Inquiry Framework. *Online Learning Journal*.

Zhang, J., & Dempsey, P. *Re-examining the Construct Validity and Causal Relationships of Community of Inquiry Framework*. Paper presented at 2018 American Educational Research Association (AERA) Annual Conference (April 13-17, 2018), New York NY.

Dempsey, P. R. & Zhang, J. (2018). Assessing Transformative Learning in the Community of Inquiry Framework: An Online Application. Under review.