Professionalization of Program Evaluation:

A Comparative Mixed Methods Case Study of Canada and the United States

A Dissertation

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Sandra Ayoo

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Jean A. King

John Matthew LaVelle

May 2020

ACKNOWLEDGMENTS

This dissertation was written and completed as a team effort. I want to thank my committee members who believed in me and supported me all the way. I am so grateful to my advisors, Jean A. King and John M. LaVelle for your guidance on many iterations of this work. I give special thanks to Jean, who was a teacher, mentor, motivator, and provided emotional support throughout my graduate career. Jean brought me into her spaces and gave me opportunities to connect with her network locally and internationally as part of the AEA evaluator competencies taskforce and to present my work internationally.

I sincerely thank my dissertation chair, David Johnson, for continuously reminding me that "this is your dissertation" and to do what I want. You always made time for me and helped me find myself and create a niche in this scholarly work. A special thank you goes to Ernest Davenport Jr., my advisor for minor in quantitative methods in education (QME). You asked critical questions on each item and provided resources and oversight to improve the psychometric qualities of my instrument.

Outside of my committee, I am indebted to Michael Rodriguez for teaching me item writing and always making time to provide guidance and answers to my questions. Michael provided invaluable comments on the raw items of the instrument and guidance on many iterations of the measurement instrument. I also want to thank the panel of experts who gave their time and invaluable inputs in item selection and feedback. Finally, I want to thank Benoît Gauthier and Gail Barrington for their generous support and helping me to recruit interview participants in Canada.

DEDICATIONS

This dissertation is dedicated to my father, Tarasiso Ogwang. You believed in educating girls and gave me an equal opportunity with my brothers to go to school. It is also dedicated to my mother, Jemimah Ochen, who went to be with the Lord in 1995. You were the inspiration and strength of the family. I also dedicate this dissertation to my daughter, Gloria A. Noki, whose continuous encouragement, prayers and support helped me carry through the highs and lows of adult learning. This work is for you.

ABSTRACT

Professionalization of evaluation means different things to different people across the globe. This study explored the professionalization of program evaluation in Canada and the United States of America (US) using sociological models of professionalism as the guiding framework for assessing the level of maturity of program evaluation as a field of professional practice. Five concepts identified by the model were used to develop a professionalism assessment tool for program evaluators (PAToPE) to measure the behavioral attributes of professionalism: professional autonomy, expertise, ethical dispositions, innovation and research, and credentialing. Data for the study were collected using a sequential mixed methods approach starting with interviews of 27 evaluators and a test of the measurement instrument in an online survey to a random sample of 1,000 American Evaluation Association (AEA) members and 573 Canadian Evaluation Society (CES) members from the 2017 membership directory.

The results of the study suggest that credentialing is the major difference between Canada and the US and that Canada is perceived to be more advanced in the professionalization of evaluation than the US. The empirical information also suggests that program evaluators demonstrate professionalism differently by country, place of work, gender, credentialed evaluator status, level of skills, level of education, age, and years of evaluation practice. The study concludes with implications and recommendations for professional associations, evaluation practice, and future studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

The United Nations (UN) made a historic declaration that 2015 was the international year of evaluation or what is commonly referred to as “EvalYear 2015.” This declaration came at a time when the Millennium Development Goals (MDGs) came to an end and were replaced by the newly developed Sustainable Development Goals (SDGs). Regarding monitoring and evaluation, the United Nations resolution stated that:

> Our Governments have the primary responsibility for follow-up and review, at the national, regional and global levels, in relation to the progress made in implementing the Goals and targets over the coming 15 years. To support accountability to our citizens, we will provide for systematic follow-up and review at the various levels … They will be rigorous and based on evidence, informed by country-led evaluations and data which is high-quality, accessible, timely, reliable and disaggregated by income, sex, age, race, ethnicity, migration status, disability, and geographic location and other characteristics relevant in national contexts. (United Nations General Assembly A/RES/70/1, 2015: pp. 31-32)

The world’s leaders vowed to eliminate extreme poverty by 2030 and pledged that “no-one will be left behind” (United Nations General Assembly A/RES/70/1, 2015, p. 1). The SDGs have 17 goals and 169 targets to be achieved by 2030.

The UN’s historic declaration that made evaluation visible to the world is important for two reasons. First, the UN profiled program evaluation globally as an indispensable discipline needed for the successful implementation of the SDGs because the MDGs were highly criticized by the member states for not paying sufficient attention to evaluative evidence on achievements of the goals and learnings from implementation challenges (IOCE website, n.d.). According to the UN (2015), program evaluation is expected to make three primary contributions towards successful implementation of the SDGs: (1) to provide information for decision making (p. 2), (2) to provide accountability to citizens (p. 11), and (3) to ensure improvement in policies, strategies, and programs (p. 32) so that no one is left behind. The UN member states are required to engage in three fundamental roles to achieve all 17 goals and 169 targets of the SDGs by 2030: (1) develop their own national policies, (2) develop national development strategies and programs consistent with international rules and commitment, and (3) establish robust national monitoring and evaluation systems to ensure systematic follow-up, reviews, and annual report on progress and obstacles (United Nations General Assembly A/RES/70/1, 2015).

At the global level, the UN has developed a set of approximately 300 global indicators for all targets developed by chief statisticians from member states (United Nations website, n.d.). The country-led reviews will feed into the UN Secretary-General's report to the high-level political forum that will meet regularly to discuss progress toward achievement of the goals. The UN member states are implementing the SDGs in partnership with civil society, the private sector, the United Nations system, and other actors (p. 27). The SDGs provide opportunities for evaluators to practice their craft in different contexts around the world. This global development reiterates LaVelle and Donaldson's (2015) conclusion that the demand for evaluators outweighs the supply.

Second, the UN declaration has led to an inward reflection by evaluators worldwide with a call to professionalize program evaluation to meet the global demand for evaluation. This process is led by the International Organization for Cooperation in Evaluation (IOCE[1]) and EvalPartners in collaboration with the Voluntary Organizations for Professional Evaluation (VOPEs). I will return to this professionalization movement when discussing the context of this study.

Statement of the Problem

Although the UN SDGs have institutionalized demand for 'high-quality' program evaluation in government, the private sector, civil society, and the United Nations, concern remains how the field of evaluation will meet this demand (Stockmann, 2016). The problem addressed in this study is five-fold. First, despite the huge global demand for 'high-quality' evaluators, there is no entity responsible for regulating the field to supply quality evaluators, similar to the American Medical Association (AMA) for producing doctors in the United States or the 50 state Departments of Education or the General Teaching Council of England for producing teachers in the US and England, respectively. Voluntary Organizations for Professional Evaluation (VOPEs), formally

---

[1] The International Organization for Cooperation in Evaluation (IOCE) is an alliance of regional, national, and international evaluation organizations (associations, societies, and networks) from around the world that collaborate to: (a) Build evaluation leadership and capacity, especially in developing countries; (b) Foster the cross-fertilization of evaluation theory and practice around the world; (c) Address international challenges in evaluation; and (d) Assist the evaluation profession to take a more global approach to contributing to the identification and solution of world problems (AEA website, n.d.).

organized societies or associations or, in some cases, informal networks have no mechanism for supplying high-quality evaluators to meet this demand.

Second, there is currently disagreement among evaluators on what constitutes basic evaluator competencies needed to conduct high quality and technically sound evaluations, accreditation of pre-service programs, on credentialing (Shackman, 2015; Stevahn, King, Ghere, & Minnema, 2005), on the very idea of professionalizing program evaluation (Jacob & Boisvert, 2010), and a lack of unified sets of ethical codes and guiding principles for evaluators (IOCE website, n.d.). There are considerable variations in what constitutes knowledge, skills, and competencies to practice evaluation and in training programs in universities and non-university settings (Stevahn et al., 2005; LaVelle & Donaldson, 2015). As such, anyone can practice evaluation. Some scholars have argued that the advice or the work done by unqualified program evaluators may actually be useless or, worse still, misleading in the absence of certification or licensure programs (Altschuld & Engle, 2015; LaVelle & Donaldson, 2015; Picciotto, 2011; Worthen, 2003).

Stevahn, King, Ghere, and Minnema (2005) conceded more than a decade ago that evaluator competencies are an important step in professionalizing the field because evaluator competencies could be foundational in university and non-university-based evaluation training programs, could provide topics in discussions towards licensure or credentialing of evaluators, and could increase the potential for program accreditation (pp. 45-47). They listed five reasons why competencies matter: (1) regardless of skills, anyone can claim to be an evaluator since there is no standardized licensing or credentialing; (2) program directors who set out to hire an evaluator do not know how to differentiate between a qualified and a non-qualified evaluator; (3) individuals interested in becoming evaluators may struggle to determine what they need to learn or where they should study; (4) evaluation trainers base their curricula only on perceived needs or personal preferences; and (5) there is a lack of theory-based descriptive research in the field.

Altschuld and Engle (2015) and Picciotto (2011) made the same arguments many years later about the lack of exclusion of the unqualified from practicing program evaluation. Furthermore, the field of evaluation is fraught with challenges of quality assurance in the absence of requirements for certification or licensure, accreditation of pre-service preparation programs, and exclusion of unqualified people from practicing (Worthen, 1994, 2003).

Third, there is a lack of funding for empirical research on evaluation, particularly on professionalizing the field of evaluation. Ofir (2016) laments that theory-based innovative evaluation research lacks funding similar to Michael Scriven's Faster Forward Fund. Jacob and Boisvert (2010) concur that debates on the professionalization of evaluation in the United States (US) consist of articles or chapters in books presenting the authors' opinions and positions without actual research on the process of professionalization such as state of Louisiana's credentialing system for educational evaluators initiated in the 1980s.

Fourth, professionalization means different things to different people. VOPEs are taking different paths to professionalize the field, such as the voluntary credentialing program in Canada and the voluntary evaluator peer review (VEPR) system in Europe. In fact, Peersman and Rogers (2017) identified 41 approaches used for professionalization in program evaluation and other fields in their recommendations of pathways to the Australasian Evaluation Society (AES) to advance professionalization. It is not surprising that the IOCE Task Force on professionalization conceded that, "The current Task Force does not think that it is either possible or even advisable to aim for a unified approach to professionalization at this point; too much still needs to be defined and discussed before considering this possibility" (IOCE website, n.d.).

Professionalization of evaluation has been a grand challenge characterized by the fundamental and intractable problems of accreditation, credentialing, and certification for over two decades since 1994 when Worthen concluded that evaluation was not a profession because it falls short in three of the nine criteria[2] for judging professions. Worthen (2003) questioned how evaluation can be considered a profession if it does not require its practitioners to exhibit competence in some body of "specialized knowledge" and if its practitioners do not conform to the "technical or ethical standards" usual for professional bodies. The lack of requirements for certification or licensure to practice and nonexistent accreditation of pre-service preparation programs potentially lowers the prestige and status of evaluation practitioners.

Finally, EvalPartners and IOCE are leading a global movement to professionalize evaluation using professionalism as their organizing framework for professionalization

2 These criteria include a need for evaluation specialists; stable career opportunities; content knowledge and skills unique to evaluation; preparation programs for evaluators; certification and licensure; professional associations; exclusion of unqualified; accreditation of pre-service preparation programs; and standards and ethics codes.

efforts, but it is not clear what the term *professionalism* means in evaluation (Schwandt, 2017) or how to empirically assess it. In fact, existing professionalism measurement instruments do not speak to the needs and practice of program evaluation (e.g., the Professionalism Assessment Tool for pharmacists by Kelley, Stanke, Rabi, Kuba, and Janke, 2011).

Study Context: Global Evaluation Agenda 2016-2020 (EvalAgenda2020)

Professionalization of evaluation has become high on the agenda of the International Organization for Cooperation in Evaluation (IOCE), the only partnership of evaluation associations worldwide. The aim of the IOCE is to assist VOPEs in pursuing professionalization initiatives that are currently underway or planned (IOCE website, n.d.). This initiative is in response to several demands for quality control and standardization of evaluation training and practice, including the rising global demand for high-quality evaluation services, the internationalization of evaluation as many evaluators are practicing evaluation across national borders, the aspirations of VOPEs to professionalize the field, and the push for establishment of national evaluation policies[3] through parliamentarians' movement for evaluation initiated by the EvalPartners network[4] following the UN declaration of EvalYear 2015. The central concepts running through the Global Evaluation Agenda are competencies, capacities, standards, and training.

3 According to the study on "mapping status of national evaluation policies" (EvalAgenda, 2016) only 20 countries have established national evaluation policies (the study was conducted by the Parliamentarians' Forum for Development Evaluation (PFDE) with support from EvalPartners.
4 This was comprised of the International Development Evaluation Association (IDEAS), the International Organization for Cooperation in Evaluation (IOCE), the United Nations Evaluation Group (UNEG), the Independent Evaluation Office of UN Women, and the Global Evaluation Facility (GEF).

The IOCE professionalization task force noted that professionalization is not a quick fix, but "…a gradual, long term, context-dependent process with many facets: (i) improved access to quality education and training; (ii) dissemination of evaluation knowledge and good practices; (iii) harmonization of ethical guidelines and guiding principles for evaluators; (iv) agreed evaluator capabilities or competencies frameworks; and (v) legitimate ways of recognizing the fundamental knowledge, skills and dispositions needed to carry out work to an adequate standard of quality" (IOCE website, n.d.). In this regard, EvalAgenda2020 developed three strategies shown in Figure 1 to promote professionalization.

As shown in Figure 1.1, these strategies are inter-related; building individual evaluators' capacities, evaluation knowledge creation, and dissemination, and frameworks for evaluation quality depend on evaluation standards, competencies, accreditation of pre-service programs, and credentialing to identify capable evaluators and research-based theory to inform practice. At the core of these strategies is the quest to have evaluation recognized externally as a profession. The Global Evaluation Agenda recognizes that program evaluation is the "new kid on the block" among the social sciences because evaluation has not yet met all the prerequisites of a professional occupation (p. 21). As such, EvalAgenda2020 is responding to the need for certification or verification of competencies to practice as a prerequisite for evaluators to secure the franchise to operate in the public domain, similar to other professions.

Figure 1.1. Three inter-related strategies from the Global Evaluation Agenda

The IOCE is mandated to develop general guiding principles for professionalization that VOPEs can use when implementing voluntary designation, credentialing, and peer review systems (IOCE website, n.d.). Although the Japanese Evaluation Society was the first VOPE to create a professional designation framework, the Canadian Evaluation Society (CES) was the first VOPE to launch the groundbreaking voluntary Credentialed Evaluator (CE) designation program. The Canadian CE designation program requires evaluators to demonstrate competencies in at least 70 percent of five domains: 1) professional practice; 2) technical practice; 3) situational practice; 4) management practice, and 5) interpersonal practice. A more recent development is the Voluntary Evaluator Peer Review (VEPR) pilot projects initiated by the European Evaluation Society (EES) and the UK Evaluation Society (UKES) in 2014 with the support of EvalPartners. The VEPR framework is focused on evaluator knowledge, practice, and dispositions. Other evaluation associations are following suit (EvalAgenda2020 et al., 2016).

Canada has made a tremendous effort to professionalize evaluation practice through credentialing and maintaining a public registry or roster of credentialed evaluators for each province. It is anticipated that Canada may be the first country to start an evaluator licensing program. As such, Canada is more experienced; the USA is older, but slower.

Purpose of the Study and Research Questions

The purpose of this study was to investigate the extent to which and ways that program evaluation is becoming a profession in Canada and the US. The status of program evaluation as a field of professional practice was measured by the degree of professionalism in these countries. This study addressed three research questions:

1) How do program evaluators in North America describe professionalism in their everyday practice of evaluation?
2) How is professionalism in evaluation measured?
3) To what extent and in what ways are program evaluators in Canada and the US demonstrating professionalism in their practice of evaluation?

Significance of the Study

The vision of IOCE is to provide VOPEs with support in their professionalization efforts using professionalism as a framework for professionalization. However, as stated earlier, evaluation scholars decry the glaring absence of vigorous discussion of what professionalism means in the current debate on professionalization of evaluation (Schwandt, 2017) and the lack of theory-based research on evaluation (Stevahn et al., 2005; Christie, 2003; Stufflebeam, 2001, Scriven, 1996).

My study contributes to addressing these gaps in two ways. First, I used definitions and theoretical frameworks from the sociology of professions to develop a professionalism assessment tool for program evaluators (PAToPE), and I validated it in assessing evaluator professionalism in Canada and the US. Second, this study provides insights into what professionalism means in the current debate of the professionalization of evaluation. The study has implications for IOCE, VOPEs, trainers of evaluators, and evaluation practitioners. I shall discuss the implications of the study in the final chapter.

Before discussing the state of the literature on the professionalization of evaluation, I want to define key terms and concepts that inform this study: competencies, professions, professionalization, and professionalism.

Definitions of Terms

Richey, Fields, and Foxton (2001) define *competencies* as “a . . . set of related knowledge, skills, and attitudes that enable an individual to effectively perform the activities of a given occupation or job function to the standards expected in employment” (as cited in King & Stevahn, 2015, p. 23). This definition posits competencies as a set of general knowledge, skills, and dispositions necessary for an individual to perform high-quality program evaluation in any discipline or profession since evaluation, like statistics, is a trans-discipline that serves other disciplines and professions (Worthen, 2003;

Scriven, 1991). An example of such a set of competencies is the Essential Competencies for Program Evaluators (ECPE) published over a decade ago. The ECPE are being used in different countries to create context-specific sets of competencies for different purposes such as establishing procedures for credentialing evaluators in Canada, Japan, and Thailand (King et al., 2015; Podems, 2014).

Professionalization refers to the adoption of formal or informal rules and procedures that structure a field of activity and contribute to the autonomy of the professional practice (Donaldson & Christie, 2006). Van der Krogt (2015) defines professionalization from the collective power approach as "[t]he process by which a socially significant occupation organizes itself to ensure its practitioners perform their services well and thereby earn a larger share of societal respect and reward" (Professionals and their work blog book).

There is no agreed upon definition of *profession* in the sociology of professions literature (Morell & Flaherty, 1978; Modarresi, Newman, & Abolafia, 2001), but sociologists and historians (Abbott, 2014, 1988; Carr-Saunders & Wilson, 1934; Freeman, 1992; Evetts, 2003) have identified professional attributes that are important in distinguishing professions from other occupations: organized bodies of experts; elaborate systems of instruction and training; possession and enforcement of a code of ethics or behavior administered by their members; clear standards and processes for entry and licensure; and operation from a recognized and defined knowledge base.

Evetts (2003) gives a broad definition of professions as the knowledge-based category of occupations that usually follow a period of tertiary education and vocational training and experience. Evetts argue that professions assert exclusive jurisdictions through establishing forms of certification or licensure for its members to practice the craft as a mechanism of market closure. In this dissertation, I have adopted Starr's (2008) definition of profession as "an occupation that regulates itself through the systematic, required training and collegial discipline; that has a base in technical, specialized knowledge; and that has a service than profit orientation, enshrined in its code of ethics" (p. 15).

I will rely on Altschuld and Engle's definitions of *credentialing, certification, licensure*, and *accreditation* to avoid confusion since these terms are inconsistently used in literature. Altschuld and Engle (2015) define *credentialing* as a process that "specifies that a person has the requisite knowledge, skills, and practical experiences to be deemed worthy of a credential or designation of such status"; *certification* as attesting "to a person's competence on specific skills and knowledge often determined by a test given right after preparation"; *licensure* as "the right to practice in a field as conferred through a governing body or under the aegis of the government" (p. 6); and *accreditation* as "the formal review of a preparation or educational program to see that it meets standards of a profession or discipline" (p. 6). *Certification* and *licensure* lead to the control of who can practice evaluation and who cannot, that is, the *exclusion of the unqualified*. Fein's (2012) definition clarifies what a credential is. As he noted, "A *credential* is a formal indicator that someone has demonstrated a specified level of proficiency in some combination of occupation-related knowledge, skills, and abilities" (Fein, 2012, p. 10), like the Credentialed Evaluators (CE) designation in Canada.

Finally, Freidson (1970) defines *professionalism* as “a set of attributes said to be characteristic of professionals” (p. 70). He observed that professionalism is a collection of personal values or attitudes such as “commitment to one’s work as a career so that one’s work becomes part of one’s identity” (Freidson, 1970, p. 70). Picciotto (2011) asserts that evaluation provides a useful test of Freidson’s (1970, 1984, 1994, 2001) logic of professionalism because it embodies “[c]ivic values, ethical guidelines, and a shared understanding of the knowledge, skills, and dispositions that underlie quality work…” (p. 173). I shall use Picciotto’s (2011) set of indicators for assessing professionalism in this study: ethical dispositions; expertise; professional autonomy, and credentials in addition to research and innovation since they are “…[w]ell-documented list of traits that can inform the strategy of any occupational association intent on exercising more influence in society through collective means” (Freidson, 1994, p. 170).

Chapter Summary

This chapter discussed how the UN elevated program evaluation as a field of professional practice necessary for monitoring and reporting the achievement of the new sustainable development goals launched in 2015 and how the field of evaluation is not yet ready to meet the UN's call for professional evaluators. In summary, the chapter outlined the problem this study addressed; the global context to compare professionalization processes in Canada and the US; its significance in the ongoing search for answers to the professionalization of the field; and definition of terms. Chapter Two describes the theoretical concepts underpinning this study and the state of the literature on professionalization by comparing the histories of the medical profession and program evaluation. I shall introduce the theoretical model for my study in Chapter Two.

CHAPTER 2

Literature Review

Introduction

This chapter begins with a review of the literature on the professionalization of evaluation. This research is situated in the sociological studies of professions. In 1994 Eliot Freidson, an internationally renowned American sociologist, argued that what distinguishes professions from occupations is the degree of professionalism, which can be empirically established by studying the behaviors of professionals. I studied professionalism behaviors in a sample of evaluators in Canada and the US using five traits of professionalism: Professional Autonomy, Ethics, Expertise, Credentialing, and Innovation and Research in order to understand the level of maturity of program evaluation in North America. I will return to the topic of professionalism later when discussing the bodies of literature and the results of my study, but briefly, I present the literature review in three sections.

In the first section, I compare the histories of program evaluation and medicine in the US, particularly the role of the American Medical Association in professionalizing the medical field. Occasionally, I refer to the history of medicine in England to give context and also use the legal profession as an example to give a sense of where program evaluation is in the process of becoming a profession. Starr (2008) asserts that medicine is the standard – the model used for judging other professions in the sociology of professions literature since it has both the paradigmatic and the exceptional case emulated by other professions. Second, I critique the current research on the professionalization of evaluation, and finally present my conceptual framework, the “purposeful professionalization model, and a summary of the chapter.

The History of Professionalization of Medicine in England and the United States

The history of medicine can be traced back to developments in England and its spread to the British colonies, including the United States (Shafer, 1936). The movement towards the professionalization of medicine in Britain dates to the mid-nineteenth century with the establishment of medical schools and publication of the first issue of *The Provincial Medical and Surgical Journal* in 1840 (Waddington, 1990). But prior to journals was the establishment of medical schools.

Medical Schools. The formal study of medicine in the US started with the College of Philadelphia founded by Benjamin Franklin in 1762, which later became the College of Philadelphia Medical School in 1765 (now the University of Pennsylvania). Three other medical schools were established in the 1700s: Kings College (now Columbia University) in 1768, Harvard Medical School (commonly called Massachusetts Medical College) in 1783, and the Medical College of New Hampshire (Dartmouth College) in 1798. The number of US medical schools mushroomed between 1820 and 1830. By the 1850s, medical schools had been established in every large city throughout the country (Shafer, 1936), and there were 42 schools in the United States compared to only three in France at the time (Starr, 2008). The number of US medical schools continued to grow, reaching a total of 160 by 1890 (Starr, 2008). Journals were an integral part of these schools.

Medical Journals. The *Provincial Medical and Surgical Journal* is the forerunner to the *British Medical Journal*. But the development of medical journals in the US was gradual. Shafer (1936) writes that the only medical magazine in the US in 1797 was the *American Medical Repository*, but by 1850 the profession had books and 24 magazines or periodicals that circulated throughout the country. The magazines covered diverse medical subjects, including general medicine, medical bibliography, dentistry, and insanity. Prior to this date, the profession in the US had relied upon sympathetic philosophical and scientific journals for publication, but by the late nineteenth century the profession had its own medical journals that served the entire nation (Shafer, 1936).

Licensure and Professional Autonomy. Waddington (1990) writes that early in the history of the professionalization of medicine was medical practitioners' concern to acquire professional autonomy. This started with complaints from doctors of overcrowding in the profession (Starr, 2008; Waddington, 1990; Shafer, 1936). Overcrowding was caused by a rapid increase in the number of people qualifying from the medical schools in the US and from the College of Surgeons and the Society of Apothecaries in England in the 1820s and the 1830s. The profession also experienced high competition from unqualified practitioners. Starr (2008) writes that between 1870 and 1910, the rate of growth of doctors was 153 percent, faster than the population growth rate of 138 percent – in 40 years.

Waddington (1990) advances two arguments presented by the doctors in England for licensure. The first was that medical education was an investment and that unqualified practitioners denied the qualified to reap the legitimate return on investment. The second reason was the harm patients might suffer from the unqualified practitioners. Starr (2008) noted that while doctors in the US framed their arguments as science versus quackery, the "irregulars" or the unqualified framed licensure as free competition versus monopoly. The irregulars in the US also argued that the public had good sense and ought to be free to make their own choice.

As Waddington (1990) noted, these widespread demands within the medical practice in England led to the passing of the *1858 Medical Act*, which paved way for the establishment of the General Council of Medical Education and Registration. The General Medical Council (GMC) was tasked with the responsibility of self-regulation of medical practice on behalf of the state, to oversee medical education, and to maintain a register of qualified practitioners (Waddington, 1990). The Act imposed a legal boundary that constrained the practice of the unqualified by excluding them from government service, and it also conferred a competitive advantage on registered practitioners as "the officially recognized healers" (p. 689). The passage of the Act restricted entry into the profession as early as the 1870s. The GMC had the power to enforce codes of ethics since they were backed by law.

Unlike England where licensing led to the exclusion of the unqualified from practice, licensure of doctors in the US grew incrementally. In the early nineteenth century, the general public perceived the licensing of doctors to be an expression of favor rather than competence, so it was treated with suspicion (Starr, 2008). The medical societies and licensing boards were perceived to be closed corporations, like banks and monopolies. As a result, state legislatures rescinded all earlier medical licensing laws entirely in the 1830s and 1840s during the Jacksonian period (Starr, 2008). Medical societies were also constrained for lack of resources.

Nevertheless, licensing of US doctors became successful in the late nineteenth to early twentieth centuries as a result of the reorganization of the American Medical Association (AMA). Unlike England where the GMC was established by the *1858 Medical Act*, AMA grew out of state medical societies. Despite a rough beginning, AMA provided a central body to judge education, to evaluate innovations in the various phases of medicine, and to give the profession a sense of concerted action (Shafer, 1936). Shafer (1936) argues that AMA established general examination boards conferring certificates from the Association. The initial licensing of doctors required only diplomas and made exceptions for long-established practitioners, but gradually all candidates were required to present an acceptable diploma and to pass an independent state examination (Starr, 2008). By 1901 no jurisdiction was without some form of licensing statues.

Enforcement of Codes of Ethics and Minimum Standards. Another major development in the professionalization of medical practice in England was the enforcement of a *code of medical ethics* in the nineteenth century. Waddington (1990) writes that the development of the modern code of medical ethics in 1803 entailed the establishment of a system of social control in which the professional activities of doctors were regulated by the actions and sentiments of their professional colleagues. According to Waddington (1990), in addition to enforcing the code of medical ethics that regulated intra-professional relationships due to widespread conflicts among medical practitioners, the General Medical Council had powers to discipline practitioners and *enforce minimum* *standards* of professional behavior. Although the association was crippled with financial constraints and internal weaknesses in the earlier decades, AMA succeeded in keeping unqualified practitioners out of the few medical positions in the federal government (Starr, 2008).

Standardization of Medical Education. Finally, one of the major developments in the medical profession was the standardization and centralization of medical education. Practitioners in the US and England received medical training through an apprenticeship system in the early years except for a few who went to medical schools (Waddington, 1990; Starr, 2008; Shafer, 1936). The quality of training received by an apprentice varied from one master to another depending on the master's library and personal commitment (Starr, 2008). The apprenticeship system in England lost ground to hospital medical schools and later to universities in the nineteenth century.

Waddington (1990) writes that the reforms in England started with the amassing of medical students in emerging regional and national centers of medical education where they were intensively socialized to foster a sense of professional community. As such, many students passed through the hands of a relatively small number of teachers who subjected them to broadly similar influences (Waddington, 1990). Branch (2010) agreed that the socialization of medical students begins in medical school and continues throughout a doctor's career – a process that contributed to the framing of their professional identity formation.

Standardization of medical education in the US followed a somewhat different path. Foremost, the efforts of local reformers were frustrated by competition for students.

Medical schools were hosted in universities, but operated independently as private entities without connections with hospitals. The professors did not receive salaries – they shared student fees from what remained after paying their due to the universities. Second, local reformers could not rely on the state for protection after the massive repeal of licensing laws in most states. Instead, they turned inward to rely on their own system of regulation through AMA as the leader of a national approach to reform.

Starr (2008) observes that three factors triggered educational reform in the US: money, leadership, and the birth of AMA. First was the availability of money such as the endowment of $7 million from the wealthy Baltimore merchant Johns Hopkins in 1870 to build a hospital and university – this was the single largest endowment in the country's history at that time. The second reason was the emergence of new leadership from a generation of younger college educators with a passion for research and the utility of a college education. Among this new blood were notably President Charles Eliot of Harvard and Daniel Coit Gilman of Johns Hopkins (Starr, 2008; Shafer, 1936).

Eliot overhauled the entire medical education system at Harvard. He abolished the system of division of student fees to professors, reformed curriculum and standards of admission, and assimilated the medical school into the university by placing the school's finances under the control of the Harvard Corporation (Starr, 2008; Shafer, 1936). Starr (2008) describes how the success of reform at the Harvard Medical School triggered reforms in other schools and the formation of the National Association of American Medical Colleges (AAMC) by the reformed schools. The most significant reforms happened at Johns Hopkins University under the leadership of Gilman. Gilman established a model medical school in 1893 rooted in basic science and hospital medicine that drew academic medicine away from private practice. This was the model that eventually governed all institutions in the country (Starr, 2008).

Coming of Age of AMA. Finally, AMA's coming of age made consolidation of the reforms and standardization of education possible. The main reason for this success was the evaluation and grading of medical schools in 1906. Representatives of the AMA Education Council visited all 160 medical schools in the country and graded them into Class A, B, and C according to the record of their graduates on state licensing examinations (Starr, 2008). Furthermore, in 1908 the Council invited the Carnegie Foundation for the Advancement of Teaching to evaluate the curriculum, facilities, faculty, and requirements for admission using external evaluators under the leadership of Abraham Flexner.

The Flexner Report was published in 1910, and Flexner recommended closure of 129 schools, although 70 survived after lobbying by state legislators since some states would have been without a medical school (Starr, 2008). The AMA Education Council eventually became a national accrediting agency for medical schools, and the state licensing boards formed the Federation of State Medical Boards. Starr (2008) observes that the decisions of the Federation of State Medical Boards and the AMA Council on Medical Education had the force of law even though they were not set by legislation.

In summary, the history of professionalizing medicine started as early as the mid-nineteenth century with the creation of the first medical schools and medical societies; the publication of medical journals; the autonomy of practitioners; exclusion of the unqualified; enforcement of code of medical ethics and minimum professional standards through legislation in England; the establishment of General Medical Council and AMA; and centralization and standardization of medical education.

History of Professionalizing Program Evaluation in the United States

As noted in Chapter 1, program evaluation is the “new kid on the block” among the social sciences (EvalAgenda2020, 2016) or what Scriven (1996) referred to 20 years ago as "… a very young discipline although it is a very old practice – an ancient practice" (p. 395). House (1990) writes that program evaluation has transformed from a small, sideline activity conducted by part-time academics into a “professionalized minor industry, replete with its own journals, awards, conventions, organizations, and standards” and that evaluation has its own entry in the Dictionary of Modern Thought (House, 1990, p. 24). This old behavioral practice has evolved in the past 560 years into a field of professional practice with more than 138 professional associations of evaluators and VOPEs of various capacities existing in more than 110 countries (EvalAgenda2020, 2016; Rugh & Segone, 2013).

It is interesting to note that program evaluation originates from the practice of educational assessments of programs and student performances. Madaus, Scriven, and Stufflebeam (2012) developed a comprehensive outline of the history of evaluation in the US. I shall use this historical framework since it is frequently cited in the literature and a reliable framework. These scholars discuss the history of program evaluation by using nine stages of the development from a period well before 1900 to the present time. The developmental stages or “ages” they describe are as follows:

(1) Age of Originality – 1444-1700s

(2) Age of Reform – 1792-1900

(3) Age of Efficiency and Testing – 1901 until 1930

(4) Tylerian Age – 1931 to 1945

(5) Age of Innocence/Ignorance – 1946 to about 1957

(6) Age of Development – 1958 to 1972

(7) Age of Professionalization –1973 to 1983

(8) Age of Expansion and Integration – 1983 to 2000

The Age of Originality (1444 – 1700s) represents a period in history when educators were held accountable for their services to the students as well as the society through "Payment-by-Results (PBR)" schemes originally developed in Italy in 1444 (Sou, 2008, p. 1). Schoolmasters' salaries were linked with the students' performance on tests in reading, spelling, writing, and arithmetic. In the 1800s, PBR was widespread in the United Kingdom and its colonies and subsequently in the United States. The PBR scheme was abandoned in the UK by 1920 and in the US by the 1930s.

The Age of Reform: 1792 – 1900 is characterized by the evolution of evaluation as testing of student performance and assessment of educational programs (Madaus et al., 2012). The first documented formal use of evaluation was in 1792 when William Farish used quantitative marks to assess student performance in Britain (Hogan, 2007; Hoskins, 1968). This was the initial step in the development of psychometrics and questions designed to measure factual technical competence in subject areas. Evaluation was used to reform education in Great Britain during this period. Smith (1987) noted that the earliest documented formal use of evaluation in the United States occurred in 1815 when the army developed a system of policies for “uniformity of manufacturers’ ordinance” (as cited in Hogan, 2007, p. 2).

As Madaus et al. (2012) observed, the first formal educational evaluation in the US took place in Boston, Massachusetts in 1845 where pupil test scores were used to evaluate school or instructional programs. Joseph Rice, a journalist and educational reformer concerned about methods of teaching spelling in the US, used experimental research design to study student learning outcomes. Madaus et al. (2012) noted that Rice’s experimental design approach to evaluation was later advanced by Lindquist and Campbell in the 1960s and 1970s. During the Age of Reform, evaluations of educational programs were sponsored by governments both in the UK and the US. In the UK, the government appointed commissions to conduct program evaluation while in the US evaluation was performed by presidential commissions (Sou, 2008). As observed by scholars in the history of evaluation (Madaus et al., 2012; Sou, 2008), these inquiries that were conducted by commissions and a number of associations dedicated to social inquiries constituted an empirical approach to program evaluation.

The Age of Efficiency and Testing: 1901 – 1930 experienced the growth of program evaluation as the objective measurement of the efficiency of school systems and the use of tests to make inferences about district effectiveness (Madaus et al., 2012). During this age, Fredrick Taylor introduced scientific management, which emphasized systemization, standardization, and efficiency through evaluation. By 1915, large educational systems prescribed objectives that were the basis of Objective-Referenced

Assessments (ORA) used to improve the efficiency of educational systems. Tests with normative data were developed to judge whether educational programs were above or below the general standard for the systems (Sou, 2008). Program evaluation was done based on Norm-Reference Assessments (NRA). The concept of external evaluators was also introduced at this age. In fact, Sou (2008) reports that a few local people invited outside experts to expose defects and propose remedies.

Age of Tyler: 1930 to 1945 witnessed a major development of criterion-referenced testing by Ralph W. Tyler. According to Madaus et al. (2012), Tyler coined the term "educational evaluation" (p. 8), and he conceptualized evaluation as a comparison of intended outcomes with actual outcomes. Tyler applied his conceptualization during the early and mid-1930s to help improve programs and tests. From 1932 – 1940, Tyler directed the Eight Year Study to assess outcomes of programs in 15 progressive high schools and 15 traditional high schools in the US (Hogan, 2007) with funding from the Carnegie Corporation (Sou, 2008). Tyler's work emphasized measuring attainment of educational objectives, a shift from focusing on individual students in relation to test norms (Russ-Eft & Preskill, 2009; King, 2003). As such, the Tylerian approach led to the development of criterion-referenced testing as a replacement for norm-referenced testing, therefore laying the foundation of program evaluation by the mid-1940s (Russ-Eft & Preskill, 2009; King, 2003).

Age of Innocence/Ignorance: 1946-1957 did not see much growth in evaluation because American society gave insignificant attention to the accountability of national funds spent on education (Madaus et al., 2012; Hogan, 2007). Nevertheless, this period showed the rapid adoption of Tyler's view of evaluation. The Tylerian approach was used to train teachers in test development and the development of many national NRAs that were purchased by schools for evaluation (Sou, 2008). Furthermore, the 1950s saw the expansion of the standardized testing business with the expansion of technology. Madaus et al. (2012) labeled this the period as the Age of Innocence or Ignorance because the work in evaluation had no social purpose despite the establishment of the American Educational Testing Service in 1947. These scholars were concerned with the lack of quality assurance since evaluation was not concerned with identifying stakeholders' needs and critically examining society's response to those needs. In this age, professional organizations like the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME) were concerned with the quality of assessments.

Age of Development: 1958 – 1972 is when the field of evaluation witnessed historical events that advanced program evaluation in the US (Madaus et al., 2012). The first was the federal funding of evaluation to measure the success of the new curricula mandated in 1958 when Congress enacted the National Defense Education Act (NDEA) following the national crisis sparked by Russia's launch of Sputnik I (Hogan, 2007). The second event was the passage of the Elementary and Secondary Education Act (ESEA) of 1965 that stimulated large-scale evaluation studies of federal government grants. The Act was passed to provide compensatory education to disadvantaged children. Since 1965, evaluation has become part and parcel of every federal grant (Russ-Eft & Preskill, 2009; King, 2003), positioning evaluation as an emerging profession or a near profession

(Mertens, 1994; Worthen, 1994; Shadish, Cook, & Leviton, 1991; House, 1990). In this period, universities began offering courses in evaluation methodology, and several evaluation journals began, along with the formation of professional associations and the development of evaluation standards (Russ-Eft & Preskill, 2009; Hogan, 2007).

According to House (1990), before 1965 evaluation meant testing or curriculum evaluation mainly by measurement people, but after 1965 it became a multidisciplinary field with separate traditions and multiple histories. He further observes that the methodology, philosophy, and politics of evaluation changed during this period. The methodological conflicts between the scientific randomized trials led by Campbell and Stanley and qualitative approaches led by evaluators like Robert Stake and Barry MacDonald ultimately culminated into a rapprochement in which most theorists advocated for mixed methods (pp. 24-25). The Age of Development also saw a reconceptualization of evaluation due to a perceived lack of relevance and utility of evaluation findings and joint efforts by AERA, APA, and NCME to develop the 1966 edition of the Standards for Educational and Psychological Test and Manual (Sou, 2008).

Age of Professionalization: 1973 – 1983 was a period in which the field of educational evaluation crystalized as a practice, a field of professional practice distinct from research and testing (Madaus et al., 2012). These scholars noted that evaluators faced an identity crisis on whether to be researchers, testers, administrators, teachers, philosophers, or iconoclasts. Evaluations generated confusion, anxiety, and animosity since anyone could carry the title of evaluator without adherence to any standards. APA recognized the need for separate standards dealing with program evaluation and revised its 1966 edition of Standards for Educational and Psychological Tests in 1974 to address this concern (Sou, 2008). Other initiatives included the issuance of a comprehensive set of standards for judging evaluations of educational programs, projects, and materials by the Joint Committee on Standards for Educational Evaluation after their appointment in 1981 (Madaus et al., 2012).

Age of Expansion and Integration: 1984 – 2000 in Stufflebeam's (2000) account characterizes a period of a growing professional consciousness among evaluation practitioners and theorists (as cited in Sou, 2008). However, program evaluation struggled under the Reagan administration with cutbacks in funding for evaluation and social initiatives in the 1980s (Weiss, 1998, as cited in Hogan, 2007). This led to decreased opportunities for external program evaluators, but the role of internal evaluators arose in this quiet decade when large bureaucracies started to develop their own evaluation offices and hired internal evaluation staffs (House, 1990).

There were methodological wars between evaluators affiliated with positivistic or quantitative approaches versus phenomenological or qualitative approaches due to ideological differences (House, 1990; Madaus et al., 2012; Sou, 2008). The Reagan administration re-emphasized standardized achievement testing for purposes of centralized accountability in which evaluation was used to judge programs and personnel for purposes of holding them accountable (House, 1990). Other developments in this period included the merger of the Evaluation Research Society and the Evaluation Network to become the American Evaluation Association in 1986. The Evaluation Research Society was established in 1976 to serve government-based evaluators and some university evaluators while the Evaluation Network served the needs of university professors and school-based evaluators (Russ-Eft & Preskill, 2009). AEA has grown from just over 3,000 members in 2001 to approximately 7,000 today (AEA website, n.d.), an increase of 42.86% in almost two decades.

The Age of Expansion and Integration has continued into the 2000s with current discussions on credentialing, certification, and accreditation that have become major barriers to the professionalization of program evaluation in the US (Picciotto, 2011). As noted earlier, the professionalization of evaluation has been a grand challenge, but this challenge has caught the attention of evaluators through the global movement epitomized in the 2020 Global Evaluation Agenda.

Brief History of the Professionalization of Evaluation in Canada

Like the United States where program evaluation grew out of federal government funding, the growth of program evaluation in Canada is linked to the federal government Policy on Evaluation. The federal government has had five evaluation policies over four decades; first introduced in 1977, four decades ago, and changed four times over the past 40 years – in 1991, 2001, 2009 and 2016. Gauthier, Lahey, and Jacob (2019 in press) observed that evaluation in Canada is driven by the federal government sector in its "evolution, structure, and practice" (p. 4). Each policy is accompanied by standards of practice and change in mandate, focus, and resourcing. These authors noted that the evolution of evaluation in Canada is simultaneous with the development of results-based management, the model for evaluation institutionalized by the Treasury Board of Canada Secretariat (TBS), like the Office of Management and Budget (OMB) in the US.

Like the US where the growth of program evaluation mirrored the pattern of federal funding – and was affected by both political and economic factors – the growth of evaluation in Canada reflected each governing party priorities. For example, Gauthier et al. (2019) noted that budget cuts by the Conservative Party government in the mid-1980s and the mid-1990s led to a scaling back of evaluation activity and a reduction in the evaluation role for the TBS. The 2016 "Policy on Results" has moved away from the mandatory and prescriptive or "*straitjacket*" evaluations requirements based on a set of predefined questions to a more relaxed approach to allow for evaluation scoping and timing to be based on needs of departments (Gauthier et al., 2019 in press).

While the role of the TBS is to set the rules and provide oversight for federal government evaluation, the individual government departments are responsible for measuring the performance of their programs and policies. Government evaluation in Canada is linked with the use of performance measurement and monitoring as a tool to track progress and performance (Lahey et al., forthcoming as cited in Gauthier et al., 2019).

Unlike the federal government evaluation that is well-developed in Canada, the evaluation in the not-for-profit (NFP)/Voluntary sector in Canada started to take preeminence in the 2000s. While both government and non-governmental evaluation in Canada have focused mainly on accountability in line with the 2006 Federal Accountability Act, other emerging purposes of evaluation in government are to promote learning by evaluating the relevance, effectiveness, impacts, and efficiency of programs and policies. Furthermore, evaluation in government has moved away from internal evaluators acting as “managers of external evaluations” to evaluations conducted predominantly by internal evaluators (Jacob, 2006 as cited in Gauthier et al., 2019).

Aside from setting rules and giving oversight for federal evaluations, the TBS established a Centre of Excellence for Evaluation (CEE) – replaced by “Results Division” in the TBS in the 2016 policy – with a mandate to assume a leadership role in the advancement of evaluation practice in the country (Gauthier et al., 2019). Not only did the CEE set competency profiles for federal evaluators, develop an Internship Program for new evaluators, a training and development curriculum for in-career evaluators, and a capability to monitor the health of evaluation government-wide, it also played a role in the establishment of the Canadian Consortium of Universities for Evaluation Education (CUEE), established in 2008 (Gauthier et al., in press). CUEE was created as a voluntary, collaborative partnership among universities, government, and the CES with the goal to “…[I]ncrease access to high quality, graduate-level education in evaluation that enables its graduates to acquire the required competencies to practice as evaluators and eventually to meet the academic requirements for a Credentialed Evaluator (CE) designation, awarded by the CES” (CUEE website, n.d.).

The CES is the oldest professional association in existence, serving a membership of approximately 1800. CES launched the historic Credentialed Evaluator (CE) professional designation in 2010, an initiative geared to change the future of evaluation in Canada and possibly elsewhere. The launch of the CE Professional Designation Program (PDP) arose in part from the demand by the federal government. For example, CEE’s inquiry papers on evaluation showed that senior managers questioned the rigor and quality of evaluation, so, as such, CEE seriously considered setting up a credentialing program in an attempt to raise the professionalism of evaluators employed by the federal government (Gauthier, Borys, Kishchuk, & Roy, 2006).

The push from the government for high-quality evaluation by qualified evaluation professionals contributed to two processes by CES National Council over 2005-2007; (a) a fact-finding on evaluator credentialing, and (b) a member-wide consultation on the best possible option. Cousins et al. (2009) narrate that the Council engaged a consortium of prominent and leading members of the Canadian evaluation community to conduct the fact-finding in 2006. Figure 2.1 summarizes the historical moments in the development of the CES Credentialed Evaluator designation program. According to Kishchuk, Gauthier, Barrington, Cummings, and Farina (personal communication, May 28, 2019), the events leading to the development of the CE program spanned nearly three decades.

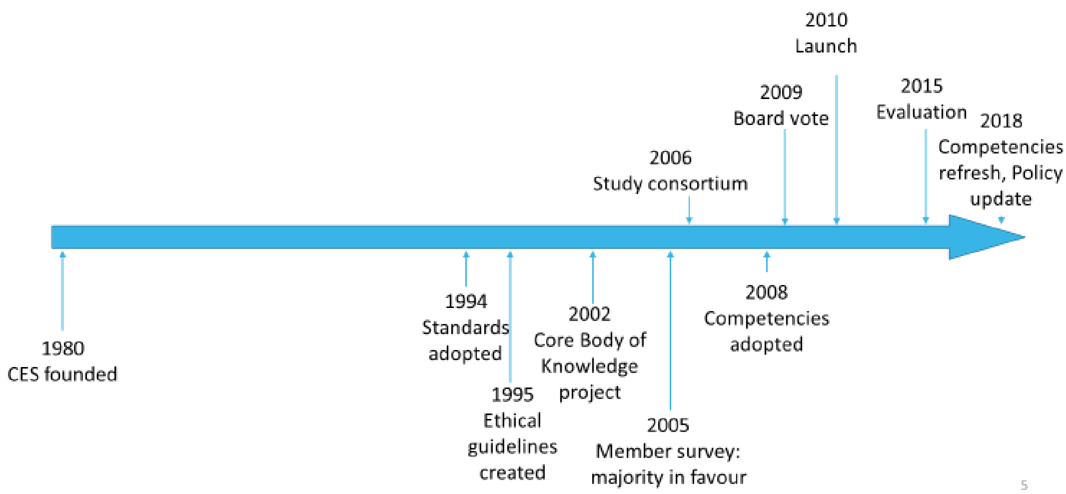

Figure 2.1. A Brief Historical Pathway of Professional Development in Canada since 1980s-2018 (Kishchuk et al., 2019)

The consortium recommended a three-tiered system of an exam-based certification professional designation program that would lead to (a) paid association members subscribing to CES objectives, evaluation standards, and ethical guidelines, (b) Credentialed Evaluator (CE) or entry-level designation awarded to members who applied after successfully completing an accredited program or its equivalent, and (c) Certified Professional Evaluator (CPE), a professional designation awarded to members who meet all of the prescribed requirements including successful completion of standardized examinations to test core knowledge competencies in the longer term (Cousins et al., 2009).

After further consultations with membership, the evaluation community at large, partner organizations, and employers in 2007, the CES Council opted to implement an intermediate voluntary CE professional designation system as the most appropriate pathway at the time. A special group of CES volunteers developed the credentialing option between 2007 and 2009 (Buchanan, 2015) including the establishment and approval of 49 competencies in 2008 (Maicher & Frank, 2015). The CES CE designation program is also anchored on the standards of evaluation practice and ethical guidelines. The credentialing program was launched in 2010 (Kuji-Shikatani, Thompson, & Mathew, 2015). Overall, the CE program attracted 40% of CES members, and Figure 2.2 shows an uptake punctuated by the 2012 Fast Track process, with an upward trend after the lowest season in 2013. An average of 74 new applications is received every year.

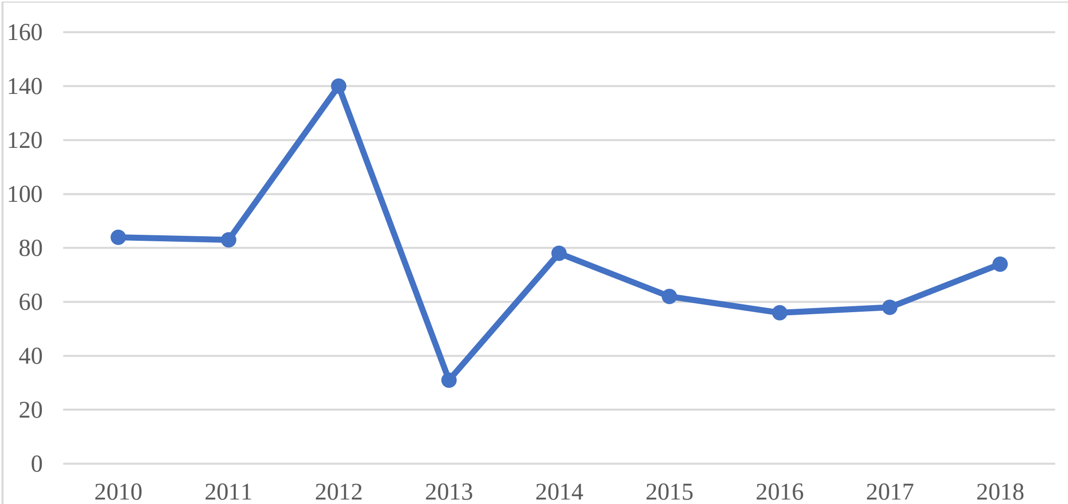

Figure 2.2. New CE Registrations per Year 2010-2018 (Kishchuk et al., 2019)

The National Council has not yet embarked on the provision for additional, higher level tier recommended by the Consortium in 2007. However, the evaluators of the CE recommended CES to pay greater attention to commissioners and employers of evaluation to create awareness about the value of CE and to understand the alignment between courses offered to the CES competencies. Kishchuk et al. (2019) noted that more employers are beginning to award more points in the requests for proposals (RFPs), thus the association has noted interests in CE designation by the non-evaluation practitioners like management consultants and auditors. Furthermore, the CUEE has put together a 77-page inventory of available courses offered by universities in Canada (Hunter & McDavid, 2018). As shown in Figure 2.3, CEs are present in all the ten provinces and the three territories of Canada. The national capital and Ontario have the highest numbers of CEs, followed by British Columbia and Alberta.

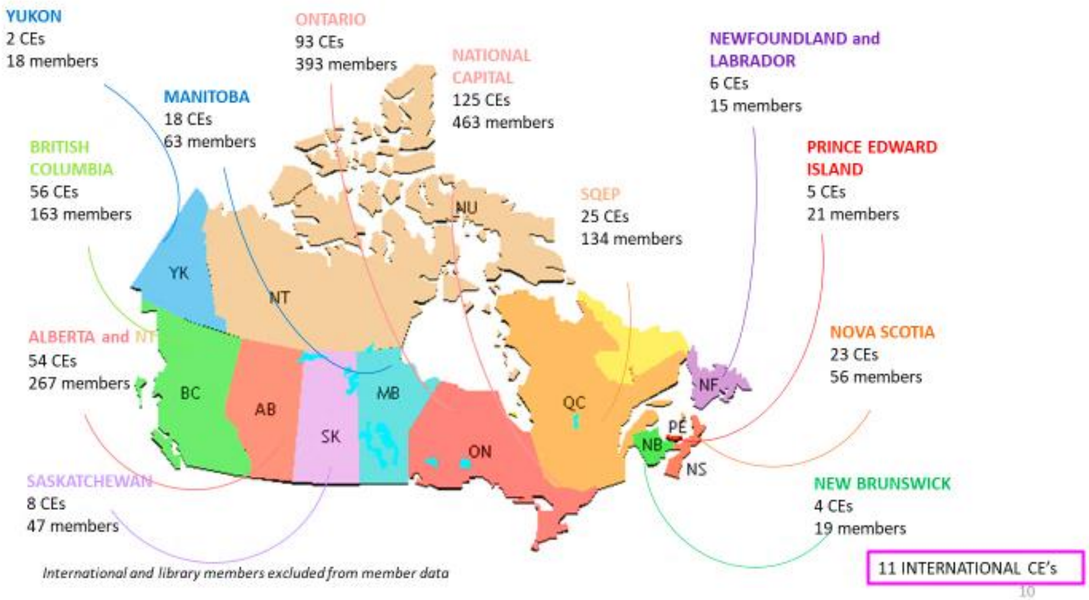

Figure 2.3. CE by Province and CES Membership (Kishchuk et al., 2019)

Comparison of the Histories of Medicine and Program Evaluation

A historical comparison of medicine and program evaluation in Figure 2.4 shows that program evaluation is presently at a place where medicine was in the 1800s. It is interesting to note that program evaluation is experiencing similar growth challenges to medicine although this is manifesting differently for evaluation. First, the professional formation of medicine was a long and enduring process. Figure 2.4 suggests that it took approximately 250 years for medicine to consolidate and standardize the curriculum for educating doctors (1700 – 1950s) and approximately 100 years to consolidate professional practice (1850s – 1950s). By comparison, it has taken program evaluation 560 years to develop the first sets of competencies that could lead to the standardization of the education of evaluators (1444 – 2004), and it might take several decades for evaluation to consolidate its professional practice unless current controversies on licensure, accreditation, and exclusion of the unqualified are resolved sooner.

Another striking difference between the two fields is the need to evaluate training schools and reform curriculum to standardize and centralize pre-service education. In the case of medicine, the initial assessment by AMA – 200 years since the formation of medical schools – paved way for the extensive Flexner evaluation, which opened opportunities for huge financial investments in medical care (Starr, 2008). According to Starr (2008), the reason for the rise of medicine to cultural authority and occupational control was scientific and technological changes, specifically, the rise of medical authority to the improved therapeutic competence of physicians related to research.

This distinction is different and largely missing in the professionalization of program evaluation in the USA. As observed in Chapter 1, evaluation scholars noted considerable variations in competencies required for the training programs in universities and non-university settings as well as diversity in the curricula used for pre-service training programs (Stevahn et al., 2005; Davies & MacKay, 2014; LaVelle & Donaldson, 2015). AEA and CES have recently adopted evaluator competencies that could be used for standardization of the curriculum and education of evaluators. Evaluation scholars are calling for action to improve the quality, consistency, and integrity of evaluator education (Gullickson, King, LaVelle, & Clinton, 2019).

Second, the medical field progressed significantly from breakthroughs in finances and leadership. In comparison, two major shortcomings in the professionalization of evaluation are the absence of a national professional association like the AMA or the

GMC to regulate the field and solicit funds for research on evaluation – like endowments from Johns Hopkins or Rockefeller Research Funds. Starr (2008) noted that research in medicine is considered a "public good," and he describes how philanthropists like Rockefeller promoted a model of medical education closely wedded to research (p. 122).

On the contrary, evaluation education closely wedded to research is unheard of and research on evaluation or evaluation reports are not branded as public goods. Fullerton and Maloney (2017) argue that evaluation must both be "non-excludable" and "non-rivalrous" for it to become a public good (p. 2). Their definition implies that the use of evaluation report by one organization should neither reduce the availability of this report to others nor exclude other organizations from using that report – like the use of fresh clean air or streetlights.

The OECD/DAC network decries the lack of positioning of evaluation as a public good, saying, "The evaluation profession needs to understand this terminology and be able to respond to requests to evaluate from this perspective, if required by politics, policies or public debates" (Van den Berg, 2015, p. 1). As I will show in the presentation and discussions of the results of this study, program evaluation has the predisposing characteristics of public service, but framing evaluation as a public good in and of itself is desirable for the growth of the field. American Evaluation Association's foundational documents posits evaluation as a public good. In fact, the 2017 AEA conference Presidential Strand focused on how evaluation can support the public good through principles-based approaches (AEA website, n.d.). Similarly, the Australasian Evaluation

Society (AES) discussed the notion of evaluation for the public good or as a public good in line with their 2017 conference theme of "Evaluation Capital" (AES website, n.d.).

Third, while AMA or the GMC had the mandate to enforce codes of ethics and a minimum standard of practice, this is not true of evaluation associations. As noted earlier, enforcement of codes of ethics enabled the medical profession to establish a system of social control that promoted intra-professional accountability and responsibility – the profession turned inward to rely on its own system of regulation. Furthermore, the governing bodies had powers to discipline practitioners and *enforce minimum standards* of professional behavior.

An examination of the history of program evaluation indicates that evaluators began to experience a sense of professional consciousness in the 21st century. As shown in Figure 2.4, evaluation practitioners and theorists started engaging in the development of the codes of ethics in the 1950s to date and more recently in the development and approval of evaluator competencies in the 2000s. I shall return to this point in Chapters 4 and 5 when presenting the findings an implications of this study, however, there is currently a lack of enforcement of codes of ethics and standards by professional associations.

Finally, medicine experienced resistance to professionalization, especially from unqualified practitioners. As noted earlier, although qualified doctors wanted to reap the legitimate returns on their investment in extensive schooling, unqualified practitioners appealed to the freedom of choice by the public and likened licensure to a corporate monopoly to gain the support of states to rescind licensure acts. Interestingly, opponents of professionalization of evaluation raise similar arguments for the freedom of choice by evaluation users. I shall return to this distinction late in Chapter 5, but Table 2.2 at the end of this chapter provides a summary of arguments for and against the professionalization of evaluation.

Sociology of Professions Approaches to Studying Professions

This study draws upon four bodies of literature in the sociology of professions[5] theory to understand the development of evaluation in the US and Canada and to explore how program evaluation is becoming a field of professional practice: (1) the trait approach, (2) the process approach; (3) the power approach; and (4) professionalism. The trait approach is the oldest, originating from the works of Carr-Saunders and Wilson (1933). The process approach is associated with the works of Caplow (1954) (as cited in Wilensky, 1964). The power approach is the newest, attributed to works of Freidson (1970, 1988, 1994) and Larson (1977), and finally, the professionalism approach associated with Freidson (1970; 1994). These approaches are interrelated. For example, Ritzer (1975) and Abbott (1991) argue that the power approach is the motor force behind the drive toward professionalization, and it is one of the attributes in the trait approach. Similarly, Freidson (1970, 1994, 2001) discusses professionalism as an approach that borrows greatly from the trait and power approaches in his detailed study of the history of medicine and how the medical profession affects other fields.

[5] The sociology of professions is a branch of the sociology of work concerned with the analysis of expert occupations (Smelser & Baltes, 2001).

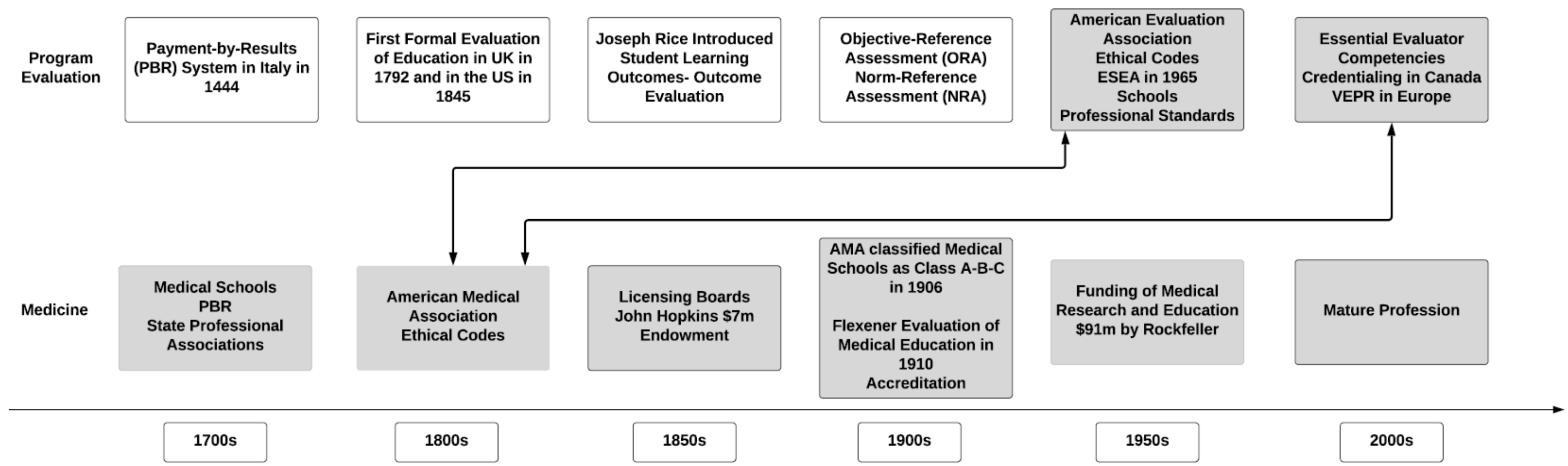

Figure 2.4. A Historical Comparison of the Growth of Medicine and Program Evaluation in the US since the 1700s – 2000s

The sociology of professions literature provides foundational insights for understanding how occupations become professions, a helpful framing for the present discussion of the professionalization of program evaluation. Sciulli (2005) writes that the sociology of professions went through three distinguishable stages. First, Alexander Carr-Saunders and Paul Wilson (1933) pioneered the subfield. Second, Anglo-American sociologists undertook to identify the essential qualities of professions and their associations for 40 years. Finally, Anglo-American sociologists (Johnson, 1972; Larson, 1977; Starr, 1982; Freidson, 1989) critical of functionalism or the trait approach adopted a revisionist posture that led to “power,” “process,” and “professionalism” approaches (Sciulli, 2005). I have used all four approaches in this study.

Approach One: The Trait or Structural Approach. As noted above, the trait approach originated from the works of Carr-Saunders and Wilson (1933), British professionals who were interested in differentiating between professions and occupations using the models of medicine and law (Abbott, 1988; Gispen, 1988). Figure 2.5 highlights some of the most common traits in the study of professions literature. In fact, Millerson (1964) identified 23 characteristics of the professions highlighted by 21 authors (cited in Picciotto, 2011).

According to the trait theorists, professions are occupations with special professional attributes or traits, including power, prestige or privilege, control of work systems, collegial authority, discretionary judgment, complex schooling and apprenticeship, qualifying examinations, professional associations, codes of ethics, special competencies in esoteric bodies of knowledge, and rational scientific knowledge based on abstract theories and concepts (Goode, 1957; Hall, 1968; Evetts, 2003). These are occupations with scientific objectivity, social responsibility, harmony of the egalitarian professional community, and ethics of financial disinterestedness (Gispen, 1988). In fact, Picciotto (2011) observed that trait theory postulates that an occupation becomes a profession when members of that occupation group “opt to acquire certain traits and respect some behavioral norms, e.g., comply with ethical codes; master a distinct body of knowledge; demonstrate loyalty to colleagues; and achieve agreed quality standards in the conduct of their craft, etc.” (p. 167).

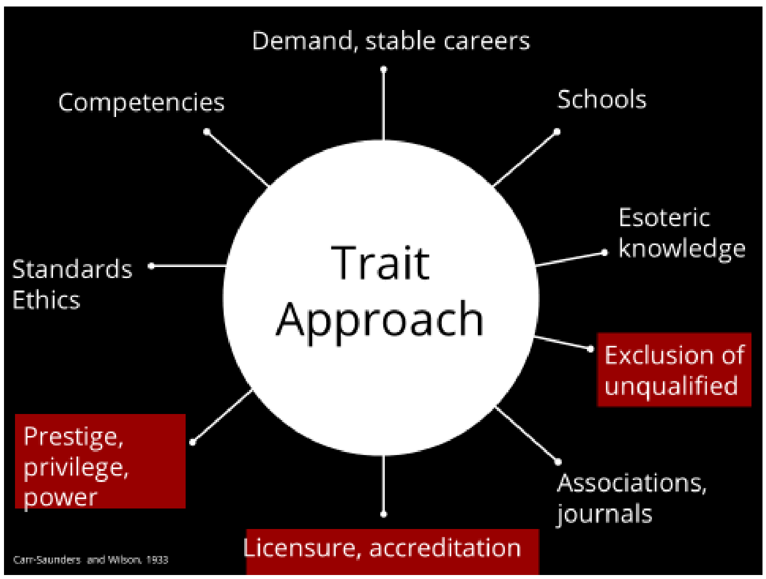

Figure 2.5 The Trait Approach

Note: Program evaluation has not yet achieved traits in red boxes.

Scholars in the trait approach ground their work in functionalist theory, which views professions in terms of their roles in furthering the “survival of the social system as a whole” (Feinberg & Soltis, 1998, p. 25). In this regard, society approves and sanctions the power and authority for professions to control the market for their expertise and the discretionary use of their knowledge and skills as necessary because their special competencies are linked to the central needs and values of the social system and because professions are devoted to the service of the public, above and beyond their material incentives (Gispen, 1988). In return, the professions enforce ethical codes to prevent misuse of authority and power, and they convince the public that the interest of the client is leading. These theorists assume that professionals are altruistic experts motivated by the needs of their clients or the public. They use their expert knowledge to enable their clients and customers to deal with uncertainty.

According to Gispen (1988) and Collins (1990), the trait approach to professionalization is more culturally relevant to Britain and the United States due to the free market economy compared to the bureaucracy that is prevalent in countries in Continental Europe like Germany or France. Collins (1990) argues that the “ideal-type” profession in the Anglo-American context “stresses the freedom of self-employed practitioners to control working conditions,” but the “ideal-type” professions in Continental Europe emphasize “elite administrators possessing their offices by credentials” (p. 15).

I will return to this distinction later in discussing the power approach, but this perspective is useful in understanding why some occupations evolved into professions while others did not. Larson (1977) argues that the existence of the free market economy is a major contributing factor in the emphasis on the autonomy of professions in Anglo-America unlike the European societies with the bureaucratization of the labor force. She argues that unlike physicians who have dominant control of the market, engineers are a subordinate market because they failed to evolve into a real profession. Engineers “sell their labor rather than their services” and are controlled by large corporations (pp. 25-31). She gives the example of the legal profession in the United States, which emerged as an autonomous profession with 75 percent of U.S. lawyers working in private legal practice in 1960 compared to 22-23 percent of German lawyers within the same period (Larson, 1977).

Program evaluators work as “external evaluators” and “internal evaluators,” reflecting both forces of the market economy and bureaucratization of the labor force. Modarresi, Newman, and Abolafia (2001) studied how academic evaluators and other evaluation practitioners experienced professionalism in the US. Their sample of 500 evaluators selected from the American Evaluation Association (AEA) membership directory for the year 1996 revealed that 36.3% evaluators worked in universities, 22% worked in government, 21.2% in the private sector, and 20.6% in non-profits (approximately 57.5% external evaluators). Another study 10 years later (Gauthier, Borys, Kishchuk, & Roy, 2006) revealed a similar pattern in Canada where 42% of evaluators worked as “internal evaluators” and 26% as “external evaluators.” Finally, a more recent survey of 980 AEA members in the United States by Seidling (2015) showed that 41% of evaluators practiced in universities, 57% as external evaluators in various settings, and approximately 28.6%[6] in the government (state, federal, and school districts).

Another important professional attribute of the trait approach is schooling. The trait approach theorists argue that schooling is required for the acquisition of requisite skills, knowledge, and attitudes because expert knowledge is an undisputable ingredient of any contemporary society (Grob & Kieser, 2006). As such, ‘schooling’ or education is

[6] The researcher asked evaluators to check all that applied; hence the percentages add up to more than 100 per cent.

a major event in the professionalization process because by "requiring specified academic education, professionalization reduces uncertainty for clients by assuring them that the professionals have acquired a minimum qualification standard" (Grob & Kieser, 2006, p. 72). According to Pavalko (1987), what matters is the amount of training necessary to master a complex knowledge base, the extent to which the training is specialized, the degree to which it is symbolic and ideational or conceptual (i.e., learning theories, concepts, and principles), and the content of training (i.e., learning of a distinctive set of values, norms, and beliefs peculiar to the occupation). Freidson (1994, 2001) posits that potential applicants are incentivized to shoulder the long, hard educational process because of the protection provided by professions from market risks. The field of medicine, for example, developed specialized training to ensure the specialization of their craft (Starr, 1982).

Waddington (1990) writes that centralization and standardization of medical education in England happened in the nineteenth century with the transition of practitioner training from an apprenticeship system to hospital-based medical schools in regional and national centers. The medical training became standardized because of many students passing through the hands of a relatively small number of teachers; as such, students were subjected to broadly standardized or similar influences (Waddington, 1990). Specialization of medical education in the UK is similar to that in the US.

The situation is not the same for program evaluation. As noted earlier in the comparison of the histories of medicine and evaluation, studies on how evaluators acquire knowledge, skills, and dispositions to perform evaluation show that the field of evaluation lacks uniformity in its educational programs. An empirical study by Dewey, Montrosse, Schroter, Sullins, and Mattox (2008) found that graduate programs are not providing all the necessary evaluation skills reportedly needed by evaluation practitioners. More recent studies by Davies and Mackay (2014) and Christie, Quinones, and Fierro (2014) also found gaps in university teaching programs. Fewer students are reportedly taking evaluation theory-based and practice-based courses (Davies & Mackay, 2014). The current state of evaluation teaching programs is summarized in Table 2.1.

Table 2.1 supports the concerns of evaluation scholars about the need for standardizing evaluation training. Stevahn et al. (2005) reiterated Schulberg's and Perloff's (1979) concerns almost three decades later that the field of evaluation should accredit pre-service training. These scholars argue that evaluator competencies are an important step in professionalizing the field because evaluator competencies are foundational in university and non-university-based evaluation training programs. Patton (2013) predicted that the future of evaluation would be work on essential evaluator competencies, which was number one on his top ten directions of evaluation. King et al. (2015) urge the field of program evaluation to act to revise and use evaluator-specific competencies to standardize evaluation training and practices, a priority that is currently high on EvalAgenda2020 agenda.

Table 2.1
The State of Evaluation Education in the United States

Academic Programs and Faculty AEA Membership	Number of Schools
Doctorate offered	29
Masters offered	47
Certificate/Professional development offered	22
Doctorate, masters, certificate offered	12
Doctorate and masters only	14
Doctorate and certificate only	8
Masters and certificate only	6
AEA Members on faculty	33

Source: AEA website (n.d.)

In summary, a major strength of the trait approach to professionalization is the use of professional attributes as frames of reference in understanding how occupations professionalize. The approach has influenced studies on the professionalization of various occupations such as the ancient professions of physicians and lawyers, which are routinely used as models by scholars of professionalization. The trait scholars of professions compartmentalize the various attributes for judging mature professions to differentiate between an occupation and a profession. However, these attributes are discussed in relative isolation from problems in society and assume relative stability. Problems like social injustice, exclusion of underprivileged groups from practicing, and abuse of power by professionals are treated as residual categories or marks of imperfections of the approach. The trait approach to professionalization leaves many facts unexplained, and it does not provide a testable explanation of the potential sequence of events in professionalization. To answer the questions of the developmental sequence of events, one must leave the functional frame of reference and examine professionalization from the process approach.

Approach Two: Process Approach. The focus of the process approach is on historical stages through which an occupation must go *en route* to becoming a profession. The process approach theorists see traits as a professional *continuum* in which occupations at the professional end of the *continuum* have more of the defining traits or characteristics of professions than those at the other end of this spectrum (Ritzer, 1975, Wilensky, 1964). Wilensky (1964) took the trait model to the next level in which he asserted that professions pass through five phases on the road to professionalization. In fact, he conducted an empirical analysis of the history of eighteen occupations in the

United States to uncover the sequence of development of professions. He concluded that the social events that lead to professionalization occurred in the typical unidirectional linear progression in which one trait leads to another as illustrated in Figure 2.6.

Wilensky's (1964) five stages are: (1) full-time occupation to stake out a jurisdiction; (2) founding of a training school affiliated with established universities; (3) formation of a local association and then a national association; (4) the passing of state-level licensing laws to protect the monopoly of skill and to eliminate the unqualified and unscrupulous; and (5) the establishment of formal ethics codes to reduce internal competition and protect clients. Wilensky's "career model" follows what Van de Ven (2007) calls a "unitary progression in the form U→V→W, where U, V, and W represent qualitatively different patterns, stages, or phases of activities or behaviors" (p. 200). As shown in Figure 2.6, Wilensky's (1964) model shows progression where full-time occupation precedes schooling or university training that also precedes associations because there must, for example, be doctors before one can have an association of them.

According to Max Weber (Weber, 1968, cited in Ritzer, 1975), full-time occupation means that a professionally trained individual is permanently occupied with the trade or discipline. Weber raised two important criteria: individuals are professionally trained (schooling), and they are permanently occupied (demand). In fact, Wilensky (1964) argued that "[s]tandardized training is requisite to entering the occupation" (p. 144). Schooling is required for the acquisition of requisite skills, knowledge, and abilities because expert knowledge is an undisputable ingredient of any contemporary society.

According to Wilensky (1964), occupational associations rarely set up training schools, but it is the schools that would normally promote an effective professional association. Wilensky asserted that associations define essential professional tasks and manage internal conflicts and external competition from neighboring occupations. As discussed in the next section on the power approach, Forsyth and Danisiewicz (1985) argue that professional associations define the service tasks that are essential (of serious importance to clients), exclusive (monopoly on the service task), and complex (discretionary application of specialized body of knowledge). It is important to note that professional associations can establish a monopoly and the closure of the professional job market through licensure (Van der Krogt, 2015). Wilensky (1964) noted that legal protection of the skills monopoly often comes towards the end of professionalization for an occupation to assert an exclusive jurisdiction.

Finally, professional associations establish codes of ethics at the end of the continuum. His claim is that formal codes of ethics are enacted as "…[r]ules to eliminate the un-qualified and unscrupulous, rules to reduce internal competition, and rules to protect clients and emphasize the service ideal" (p. 145). Pavalko (1987) observed that "…[e]thical codes may be written or unwritten and may cover a wide range of work issues including practitioner-client and practitioner-public relations, relations among practitioners, and relations between the practitioner and member of other work groups" (p. 28). Nevertheless, Pavalko (1987) cautioned that the existence of codes of ethics among professionals is not a panacea to honesty, trustworthy, or ethical behavior, but that it reinforces the service ideal and bolsters the occupation's claim to autonomy.

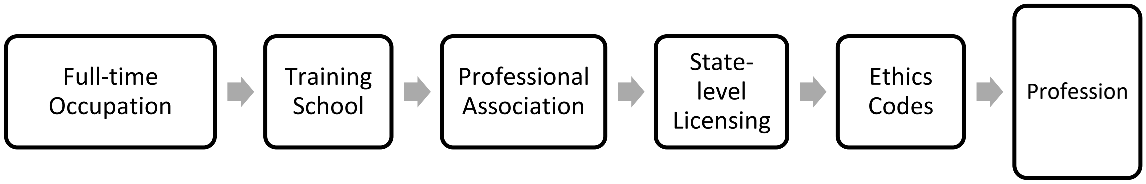

Figure 2.6 Wilensky's "Careers" Model of Professionalization

I used Altschuld's and Engle's (2015) historical timeline to understand the directions of events shaping modern program evaluation in the United States and Canada in relation to accreditation, certification, and credentialing (ACC) between the 1960s and 2014. Altschuld's and Engle's (2015) timeframe does not include events before the passage of the Elementary and Secondary Education Act (ESEA) of 1965 that stimulated large-scale evaluation studies of federal government grants in the US.

Altschuld's and Engle's timeline (2015) in Figure 2.7 indicates that the demand for evaluators and development of a unique content preceded preparation programs and journals and formation of associations, as in Wilensky's model. However, these scholars caution that much would be lost by focusing on singular events in their timeline because

> . . . one event has an impact on another, and they must be seen as interactive; as one event occurs, it collides with others, and their interactions and interplays produce a cumulative effect more than they would individually. (p. 7).

For example, in the second phase of the timeline, preparation programs, journals, and local associations happened collectively within the 1970s.

Altschuld and Engle (2015) observed that unique content was developed in the mid-1960s – 1970s. According to Madaus et al. (2012), the period 1973-1983 characterized a time of growing professional consciousness among evaluation practitioners and theorists. Key developments in this period show that universities began offering courses in evaluation methodology, the emergence of several journals, and the formation of professional associations along with evaluation standards (Hogan, 2007).

The third stage in Altschuld and Engle's (2015) timeline is the development of national professional associations, again like Wilensky's (1964) model. As AEA became more developed, Topical Interest Groups (TIGs[7]) became more noticeable at AEA in the 1980s and early 1990s, and they evolved to become the key players in developing AEA annual conference programs. Furthermore, the membership of AEA exploded and greatly diversified to include international members in the same period (Altschuld & Engle, 2015). Although the period witnessed the increase in the depth of evaluator preparation programs, the programs declined in number possibly due to the definition of what a "training program" entails (Altschuld & Engle, 2015) and the decline in federal funding for evaluator training programs across universities in the US, as pointed out by Russ-Eft and Preskill (2009). According to LaVelle and Donaldson (2015), the evaluation profession has experienced dramatic growth, development, and institutionalization since the 1960s. These scholars argue that in 2011–2012, there were over 35 evaluation-specific certificate programs, 50 evaluation-specific master's degrees, and 40 doctoral programs in the United States.

[7] Topical Interest Groups (TIGs) are a major component of the life of AEA members. Each TIG is defined around a special topic or interest to subgroups of AEA. TIGs coordinate their efforts through AEA and participate actively in AEA's annual conference through reviewing proposals in their area of interest and developing a strand of conference sessions. Members of AEA may join up to five TIGs as part of membership (AEA website, n.d.).

Wilensky's fourth and fifth stages of development of ethics and state licensing differ in Altschuld's and Engle's (2015) timeline. In their fourth stage, Altschuld and Engle (2015) noted that AEA endorsed three major developments to strengthen the quality of evaluation: (1) The Joint Committee on Standards for Educational Evaluation's (JCSEE) program evaluation standards (1981, 1994, 2011); (2) the Guiding Principles for Evaluators (ratified three times by the AEA membership in 1994, 2004 and 2018; and (3) the Cultural Competence Statement (approved by the AEA Membership in 2011).

The major related historical event in Altschuld and Engle's fifth stage in the timeline is the establishment of the Canadian Evaluation Society's evaluator credentialing program. The work on competencies can be traced to earlier writing, including Kirkhart's (1981) article on defining evaluator competencies, Patton's (1990) writing on the challenge of being a profession, and Scriven's (1996) types of evaluation and types of an evaluator. Altschuld and Engle (2015) observed that Stevahn et al. (2005), McGuire and Zorzi (2005), and others began to study and document the skills of an evaluator as part of the earlier work by King et al. (2001) on the Essential Competencies for Program Evaluators (ECPE). Altschuld and Engle (2015) concluded that evaluation is not yet a profession.

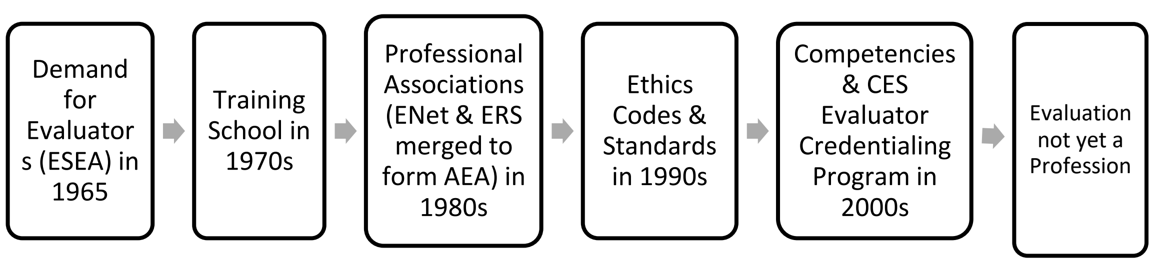

Figure 2.7 Altschuld and Engle's (2015) timeline related to Accreditation, Certification, and Credentialing

Scholars of professionalization are divided on whether events in the development of occupations follow a linear pattern or sequence or not, which criteria or event precedes the other (Abbott, 1988, 1991, 2014), and if there is a single story of professionalization or one of uniqueness (Millerson, 1964). Abbott (1988) found no empirical support for Wilensky's (1964) hypothesis. In his study of 1,932 "professionalization events" happening to the medical communities of 165 American cities, Abbott (1991) asserts that professionalization is complex and has a multilevel character with at least three stories of professionalization: a national story, a state story, and a local story. Abbott's account also shows that professionalization of medicine in the US varied from state to state, although the formation of associations and professional autonomy (licensing and control of workplace) came early in the process.

In the case of England, professionalization of medicine started with publication of medical journals, then autonomy of practitioners, exclusion of the unqualified, and enforcement of code of medical ethics and minimum professional standards enforced through legislation and establishment of the General Medical Council (Waddington, 1990), the counterpart of the American Medical Association (AMA), which was established in 1904.

Approach Three: Power Approach. The power approach is the most recent of the four approaches, and its central arguments relate to professional power and status or prestige (Larson, 1977; Freidson, 1970). According to Larson (1977), abstract knowledge is a form of cultural capital from which professions obtain income, status, and power in the market. She theorizes that "standardization or codification of professional knowledge is the basis on which a professional 'commodity' can be made distinct and recognizable to the potential publics" (p. 40) or what Pavalko (1987) refers to as theory and intellectual technique traits. Pavalko observed that the complex or esoteric knowledge that serves as a basis for an occupation's members' claim to special expertise is not necessarily based on scientific research like medicine, but can be a highly elaborate body of knowledge, such as the practice of law. If knowledge is standardized, the profession can influence the labor market by controlling admissions to universities to reduce the supply of trained professionals.

Like Larson, Freidson (1970) theorizes that professionals have tight control over the labor market based on their advanced training that also gives them autonomy to make independent judgments regarding their work. Freidson (1970) asserts that professionals have occupational control of their own work and their clients because knowledge is power. According to this approach, the power and autonomy of professions are exercised at two levels: by the professionals, and by their professional associations. At the level of the professional, Johnson (1964) argues that the professional has autonomy from clients and autonomy from the employing organization. The professional determines the needs of the client, and the client would be helpless without the help of professionals because of the high level of uncertainty that characterizes these needs (Van der Krogt, 2015; Collins, 1990). Clients must give the professionals the freedom or autonomy to choose what should be done in a situation to solve their problem.

At the level of professional associations, associations have the power to control professional education, the closure of the professional job market through licensure, and the rules of the professional behavior such as codes of ethics and professional standards (Van der Krogt, 2015). Associations have the freedom to establish a monopoly of the labor market through mechanisms like licensure and specialized training and to attain self-governance or freedom from control by the state or outsiders. The professional association of doctors in England, for example, played a vital role in achieving professional autonomy. Similarly, the American Medical Association played a crucial role in attaining the professional autonomy of physicians in the United States in the last quarter of the nineteenth century and the first decade of the twentieth century. Prior to that, doctors in America were not as powerful and authoritative as they are today. Starr (1982) quotes a professional journal that bitterly commented in 1869 that, “In all our American colleges, medicine has ever been and is now, the most despised of all the professions which liberally-educated men are expected to enter” (Starr, 2008, p. 7). The medical profession was weak, had an unregulated schooling system, was divided and financially insecure, lacked standards, and was unable to exclude the unqualified from practice and membership in the associations.

Starr (2008) describes how American medicine realized its authority through a system of standardized medical education, licensure, enforcement of codes of ethics, internal cohesiveness of the association, the commodification of health care, and refusal of salaried arrangements. Through licensing laws, the field restricted the entry of untrained practitioners. Health care was converted into a commodity by removing the care of the sick from the household to the sphere of professional service in the hospitals. The association also resisted the threat of bureaucratic control of doctors by insisting that salaried arrangements violated the integrity of the private doctor-patient relationship (Starr, 2008). As such, the profession sold its services primarily to individual patients rather than organizations, and doctors had the power to set their fees.

Furthermore, the AMA gained the public's trust by setting standards of conduct for medical professionals higher than the minimum rules governing the marketplace. In this way, AMA maintained that medical professionals be judged under those standards only by each other, not by laymen (Starr, 2008). "A professional who yields too much to the demands of clients," Starr writes, "violates an essential article of the professional code: Quacks are practitioners who continue to please their customers but not their colleagues" (p. 23). The departure from the normal market rule to colleagues in the orientation of work is what professionalism demands (Starr, 2008).

Forsyth and Danisiewicz (1985) developed a professionalization theory based on the power approach. They used Hall's (1969) theory of autonomy as involving the feeling that the practitioner ought to be allowed to make decisions without external pressures from clients, from others who are not members of their profession, or from his or her employing organization (as cited in Forsyth & Danisiewicz, 1985). Using Hall's definition of autonomy, their theoretical model examined and explained differences in autonomy using two levels of autonomy: autonomy from clients, and autonomy from the employing organization. In their view, autonomy is an attitude because individuals react to situations as they perceive their work. Forsyth's and Danisiewicz's (1985) interpretation is that attitudinal autonomy among occupational members provides a means to index the professionalization of occupations since power in professional occupations manifests in autonomy from clients and from employing organizations. In this regard, "the most professional occupations are those having members who exhibit high levels of autonomy from clients and autonomy from employing organizations" (p. 61).

As shown in Figure 2.8, Forsyth and Danisiewicz (1985) use Etzioni's (1969) classifications of true or semi-professions to theorize a professionalization model. The model is comprised of three phases. Phase one (called potential) has two components: predisposing characteristics of service tasks, and image building by professional associations. In the second phase of the model (formation), the public evaluates the occupation's claim to professional status and formation of professional autonomy. Phase 3 (stabilization) describes the stabilization and maintenance stages of the profession.

Occupations with unsuccessful public recognition are referred to as mimic professions because members, such as personnel administrators and funeral directors, have neither dimension of autonomy (Forsyth & Danisiewicz, 1985). Occupations with successful public recognition, such as librarianship, are considered semi-professions if professionals have only one dimension of autonomy. True professions like medicine and law are occupations where professionals have the two dimensions of autonomy: from clients, and from employing organizations.

Forsyth and Danisiewicz (1985) tested their model on a sample of 1000 students selected from 21 professional schools in nine US universities, all major doctoral degree-granting institutions. The students were preparing for eight occupations: medicine and law (categorized as true professions); engineering, education, nursing, social work, and business administration (categorized as semi-professions); and librarianship (categorized as a mimic profession). With a response rate of 50%, the results confirmed their hypotheses that students preparing for law and medicine had relatively high attitudinal autonomy on both dimensions while students in the semi-professions exhibited one of the dimensions of attitudinal autonomy, but not both. While students in other semi-professions exhibited one of the two dimensions, students preparing for librarianship did not exhibit either dimension of attitudinal autonomy, meaning that librarianship is not a true profession. Medical students expressed the most autonomy from employing organizations, and law students expressed the most autonomy from clients.

Morell and Flaherty (1978) argued 38 years ago that discerning the autonomy elements of prestige and power in program evaluation is more difficult because of less image building and public appreciation of evaluation service. These authors argued that while autonomy is desirable for program evaluators to assert their authority, the nature of evaluation work requires them to work closely with program staff, administrators, policymakers, and beneficiaries. Furthermore, they pointed out that evaluators work with clientele who know their needs and what they want in evaluation service. They asserted that program evaluation was on the path to professionalization because there was the availability of specialized knowledge, power, prestige, income, and the service ideal. However, Morell and Flaherty (1978) also suggested that the work of evaluators would be strengthened if evaluators could claim authority to insulate themselves from undue interferences and to have control over many areas of their work.

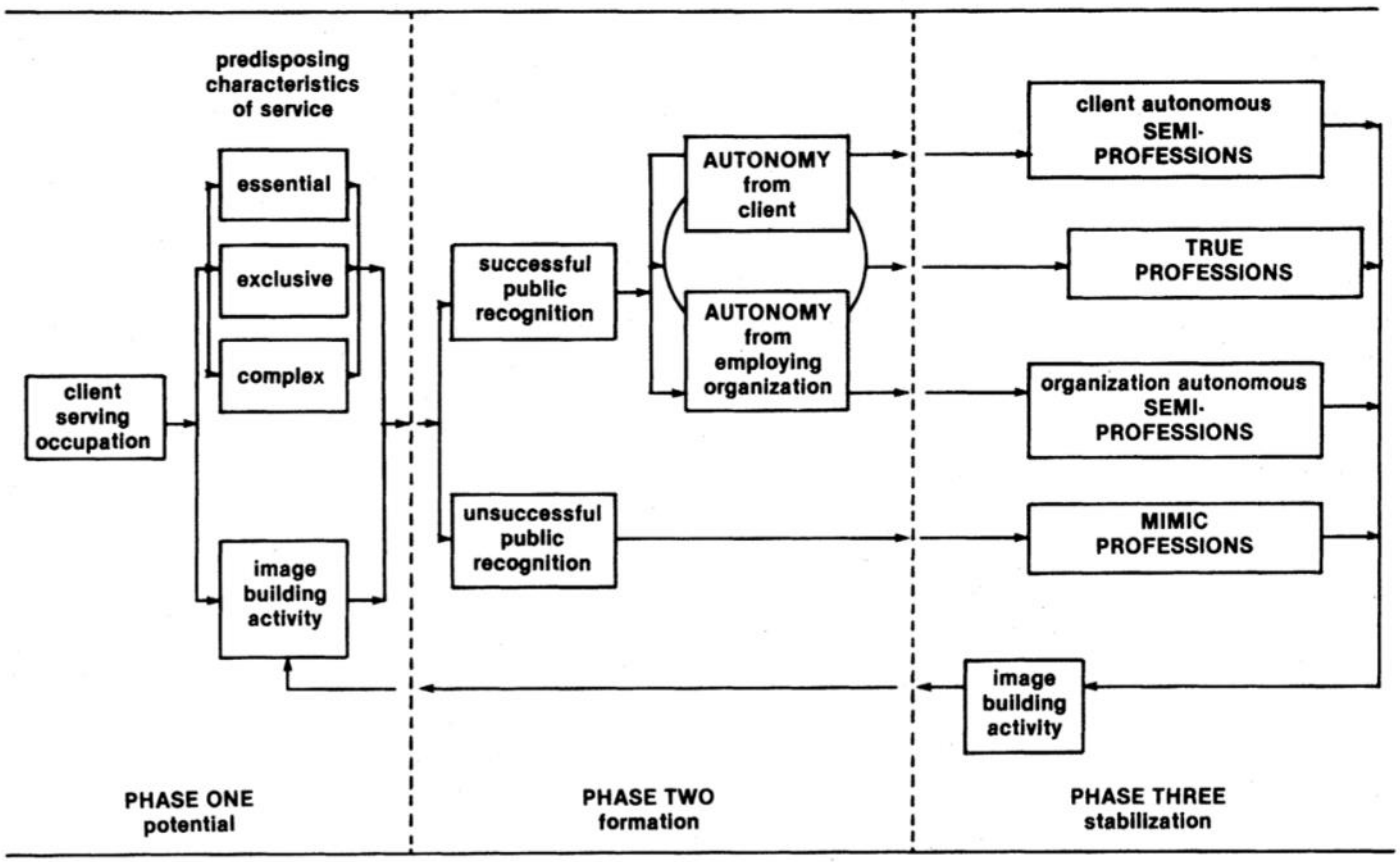

Figure 2.8 Forsyth and Danisiewicz's (1985) Model of Professionalization

Although the power approach provides a useful framework for understanding the phenomenon or the driving force behind professionalization, it assumes that an occupation is at the professional end of the continuum with more characteristics of professions listed in the trait and process approaches. The power approach appears to be useful in evaluating whether an occupation is a profession, semi-profession, or mimic profession. Evaluation scholars have long since concluded that program evaluation is not yet a profession because it does not meet all professional attributes listed in the trait approach (Worthen, 1994, 2003; Rossi & Freeman, 1985). As such, using the power approach in its current form is not helpful in moving the discourse of professionalizing evaluation forward.

Another assumption that appears problematic for the power approach is that of absolute control. Van de Krogt (2015) cautions that there is not absolute autonomy or control since no worker is totally free from external control. Instead, the influence can be very strong, but never total. The notion of autonomy and power is therefore relative. Despite these shortcomings, the power approach introduces useful concepts like image-building, public recognition, and characteristics of the service-task that inform the conceptual framework for this study.

Approach Four: Professionalism. Professionalism is the fourth sociology of professions approach. It originates from the works of Eliot Freidson (1970, 1994) and borrows from the power and trait approaches (Freidson, 1994). As Freidson (1970) observed, professionalism is attitudinal. It is a collection of personal values or attitudes such as "commitment to one's work as a career so that one's work becomes part of one's identity" (Freidson, 1970, p. 70). He defines professionalism as "a set of attributes said to be characteristic of professionals” and argues that what distinguishes occupations from professions is the degree of professionalism – which can be empirically established (p.70). Freidson (1970) theorizes that professionalism comprises service to others, expertise, control of work or autonomy, and innovation and research by faculty members of the profession to improve the profession.

IOCE and EvalPartners are using a professionalism model as the framework for their EvalAgenda2020 strategy. They see professionalism as the outcome of a collective endeavor carried out by an occupational group to improve the relevance, quality, and delivery of its expert services in the public interest (IOCE website, n.d.). Pavalko (1987) noted that the use of the term *professional* connotes competence: “... [t]hat the person knows the trade and is good at the job” (p. 18). He also pointed out that the term *professional* refers to full-time performance of an activity for pay. As such, competence invokes a sense of confidence and trust among potential customers. Interestingly, Irby and Hamstra (2016) observed that there is a lack of agreement on the meaning or definition of professionalism in the medical education community and a lack of a universally accepted method of measuring this attitudinal concept (Gauger, Gruppen, Minter, Colletti, & Stern, 2006).

These authors discussed three dominant frameworks for understanding professionalism, as shown in Figure 2.9. Virtue-based professionalism is the oldest framework. It focuses on the inner habits of the heart and mind like moral character, moral reasoning, and humanistic qualities of caring and compassion. According to Brody and Doukas (2014), professionals apply ethical principles and deeper attitudes requiring practical wisdom to their decisions. These authors argue that inward character translates into the values, attitudes, and beliefs that influence behaviors. This implies that thought precedes action because as a person thinks in their hearts, so they are. Figure 2.9 shows that virtues are assessed by written exams, observations, moral assessment, moral reasoning, and self-assessment by professionals.

The second framework is behavior-based professionalism or characteristics of a professional. It focuses on observable behaviors and milestones, including demonstration and certification of competencies, systems thinking, and diplomacy. Brody and Doukas (2014) argue that behaviors are superficial appearances of conduct. As shown in Figure 2.9, professional behavior can be assessed by performance assessment, written exams, observations, multisource feedback, and self-assessment. The third and final framework is the professional identity framework, which focuses on changing identities and socialization into the community of practice (Irby & Hamstra, 2016). This framework sees professionalism as an aspirational goal that even the best professional must struggle to attain (Brody & Doukas, 2014). Identity-based professionalism can be assessed from aspirational frameworks, multisource feedback, moral reasoning, and self-assessment. Irby and Hamstra (2016) argue that aspirational frameworks are difficult to describe, interpret, and assess.

Self-assessment is common to all three frameworks. I shall return to this distinction when discussing the measurement instrument for assessing professionalism. In fact, Brody and Doukas (2014) also argue that virtue transcends behavior and that professionalism requires integrating a set of values, attitudes, and behaviors into the core of one's being, applying practical wisdom, and having a vision of the sort of practitioner a professional wants to be in their future career.

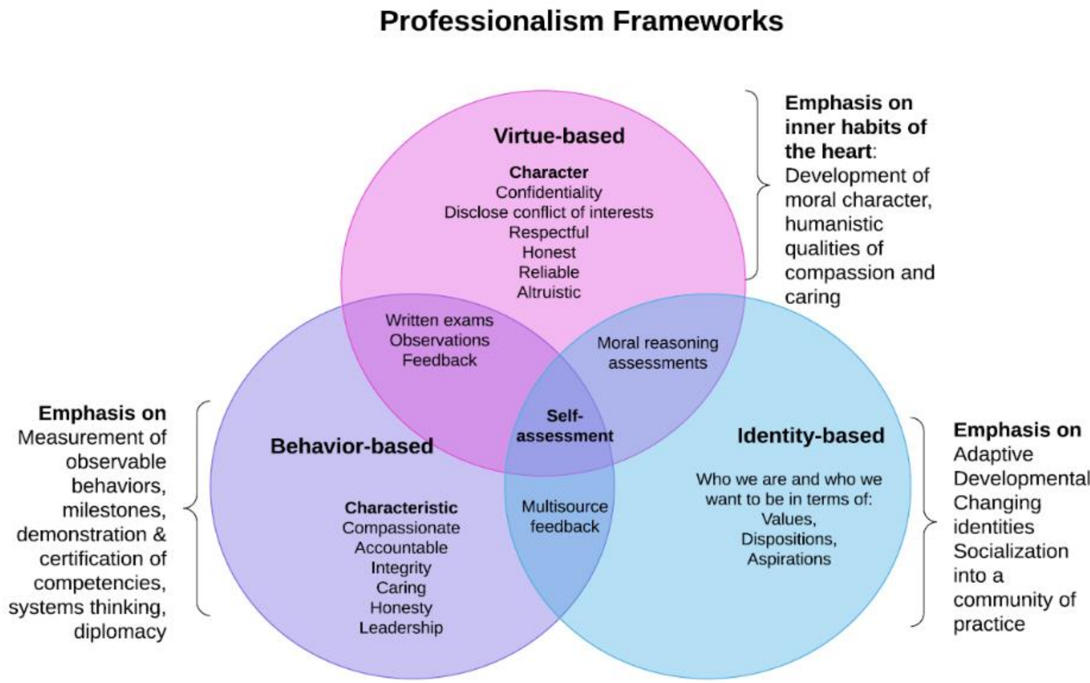

Figure 2.9 Frameworks for defining professionalism in evaluation

Note. The logic of professionalism (Picciotto, 2011; Freidson, 1970, 1986, 1994, 2001)

The main characteristics of professionalism identified can be summarized as prestige and status, professional autonomy, expert knowledge, ethical dispositions, and credentials (Picciotto, 2011; Freidson, 1970, 1994). These concepts are illustrated graphically in Figure 2.10.

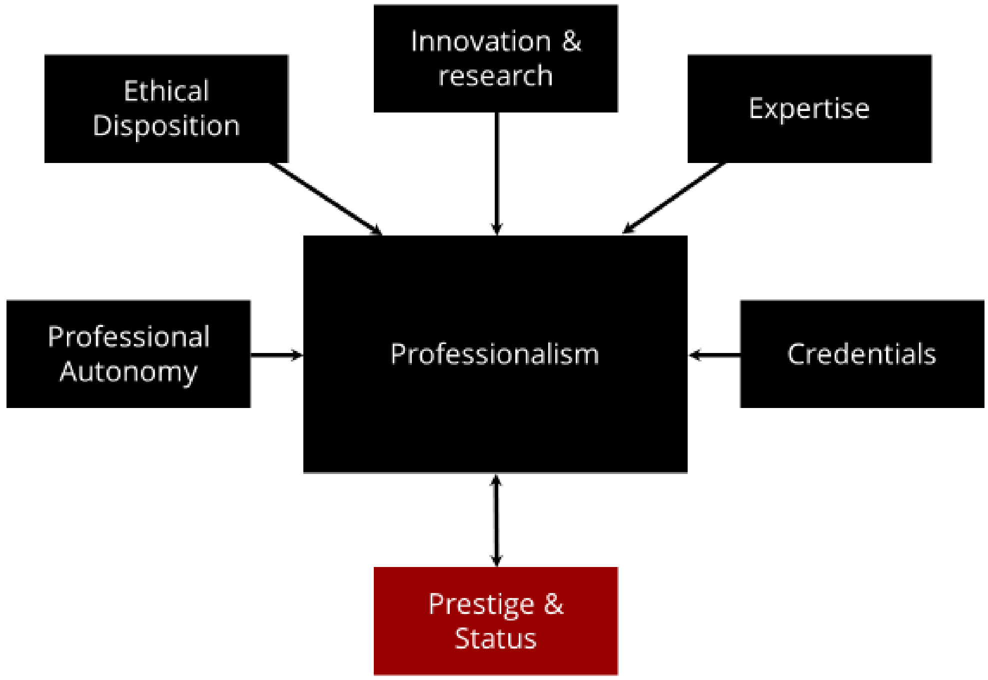

Figure 2.10 The Logic of Professionalism Model

Note: Prestige and status are not included in the model (see text).

Prestige and status. Picciotto (2011) argued that prestige and status are outcome indicators and are accorded a profession as a function of their ethical dispositions, expert knowledge, self-governing, and a guarantee of quality service often sought through accreditation/certification arrangements. He asserts that evaluation has not yet climbed high on the hierarchy of occupational groups because the public is often unclear about the nature of the evaluation discipline and the widespread perceptions of the uneven quality of evaluation work. Furthermore, he argued that the jurisdictional boundary between evaluation and rival disciplines such as auditing and performance assessment are fuzzy (Picciotto, 2011).

In comparison to medicine that gained occupational authority and status through public dependence on its gatekeeping role—such as legitimizing an employee's absence from work by giving a doctor's report—laws prohibiting possession of certain drugs without a doctor's prescription and insurance payments made only for treatment given by physicians (Starr, 2008), evaluation has not yet gained such cultural capital. Freidson (1970) writes that an occupation must control strategic accessories to get stronger sanctions from the state in support of its authority.

In line with this argument, Bickman (1997) noted 22 years ago that unlike accountants who secured the franchise as truth tellers on money matters, evaluators have failed to secure similar franchise as truth tellers of the effectiveness of public action or what House (1994) called being a major social force to legitimize public actions. Unlike physicians who resisted competition from other practitioners, Picciotto (2011) laments that evaluators are not always able to hold out against unqualified competitors from practitioners in other disciplines. Starr (2008) has written a detailed account of how professionalism served the medical profession as a basis of solidarity for resisting forces that threatened "the social and economic positions of an occupational group" (p. 27). The insufficient respectability and recognition of evaluation in the occupational ladder are also informed by the forces of supply and demand. Stockmann (2016) writes that the rising demand for evaluation services in both the public and the private side is not matched by a corresponding supply of evaluation as a specific service.

Professional autonomy as a trait of professionalism relates to an occupation's control of recruitment, quality of training, approval of professional guidelines, enforcement of ethical standards, administrative rules, quality assurance, and disciplinary processes (Freidson, 1970; Picciotto, 2011). According to Freidson (1970), occupational control happens when that occupation gets a delegation of power from the state to "determine the content of the work they do" and "control consumers of their labor" (p. 173).

Evaluation falls short in the professional autonomy attribute since (1) the guidelines and ethical standards are not mandatory or enforced; (2) most suppliers of evaluation services are not regulated; and (3) the evaluation standards or guiding principles and ethical guidelines are not universally accepted or harmonized across sectors and national borders (IOCE website, n.d.; Picciotto, 2011; Love, 1994). Furthermore, training of evaluators is not standardized due to the lack of accreditation of pre-service training programs (Picciotto, 2011; Stevahn et al., 2005; Worthen, 2001). Hung, Altschuld, and Lee (2012) lament the shortage of qualified, competent, or experienced evaluation professionals and the absence of university-based formal evaluation training programs in some countries. As such, program evaluation cannot regulate the supply of professional services, their prices, and fees.

Expertise is another trait of professionalism. It relates to high-quality education, exposure to practice, theoretical knowledge, specialized skills, sound judgment, and the mastery of techniques (Picciotto, 2011; Freidson, 1970). As noted by Pavalko (1987), the term *professional* connotes competence: "… [t]hat the person knows the trade and is good at the job" (p. 18). In fact, Picciotto (2011) argued that professional work is not 'ordinary'; it presumes theoretical and up-to-date knowledge, specialized skills, and sound judgment. He asserted that

> . . . since no one would willingly agree to have one's heart surgery performed by a

> plumber, it makes eminent sense for professional work that involves sensitive and complex tasks that impact on human security to be performed by selected individuals who have undergone specialized education at the tertiary level followed by substantial exposure to expert practice. (pp. 168-169)

Picciotto (2011) concludes that evaluation lacks stability and clout within academic establishments as a trans-discipline.

Expertise is also enhanced by **research and innovation**. Freidson (1994) writes that institutionalization of research and intellectual innovation by members of the profession in universities to develop new knowledge, skills, and ideas helps to advance a discipline. According to Freidson (1994), some members of the profession “will experiment, innovate, and do research to expand their body of knowledge and skills both for its own sake and to find new practical solutions for it” (p. 175). Hence, professionalism is tied to universities as “shelters within which modern intellectuals, scholars, and scientists do their work” (p. 177).

Freidson (1994) endorses research by special members of the profession serving as faculty since practitioners are heavily involved in the day-to-day activity of serving others. Faculty members must be supported economically by the profession to theorize and do research. As discussed earlier, medicine as a disciple advanced because of substantive financial support for research and innovation from wealthy individuals such as Johns Hopkins and foundations like the Rockefeller Foundation (Starr, 1982). EvalAgenda2020 is inviting innovations to institutionalize expert knowledge in evaluation methods and capacity-building, peer-to-peer exchanges, use of technology for communications, and scale-up and involvement of new partners (Stockmann, 2016).

Ethical disposition entails orientation towards the public interest, loyalty to the occupational group, commitment to a life-long career, collegial behavior, occupational solidarity, and responsibility for the quality of one's work (Picciotto, 2011). Like the earlier approaches – trait, process, power -- scholars of professionalism (Freidson, 2001; Picciotto, 2011) see ethical dispositions as measures by the profession to protect the public, especially codes of ethics, disciplinary committees, and peer or collegial review. Picciotto (2011) writes that orientation towards the public interest is widespread among evaluators, and it is an asset nurtured by all evaluation associations. Nevertheless, codes of ethics are not enforced by professional associations since offenders cannot be stripped of their credentials or licenses as evaluators are not certified to practice.

Credentials are associated with a degree from an accredited tertiary education establishment, professional designation, tested performance, and membership in professional associations (Freidson, 1994; Picciotto, 2011). According to Picciotto (2011), while designation is earned by a person qualified to perform a job or task by a professional body acting to safeguard the public interest, credentialing confirms proof of completion of specified training and experience requirements. Aside from Canada that has made progress in rolling out the designation of evaluators through a credentialing system, Picciotto (2011) argues that credentialing in other countries is bogged down with inconclusive debates on competencies and the lack of accreditation of evaluation training programs. He claims that the rationale for credentialing lies in managing the risks faced by users when selecting professional experts through personal contacts, word-of-mouth testimonials, or trial employment.

EvalAgenda2020 rests the case for evaluation professionalism worldwide with a call that it will require generally accepted ethical guidelines adaptable to diverse contexts and focused on commissioners as well as evaluators, access to tertiary evaluation education, mastery of techniques acquired through reflective professional practice, and self-management reinforced by peer review and/or credentialing processes.

Conceptual Framework

Philosophical Assumptions. My worldview for this study is pragmatism. My review of the professionalization literature identified four main conceptual explanations for understanding how occupations like program evaluation are becoming a profession: the trait approach, the process approach, and the power approach, and professionalism. Importantly, there is no single explanation that can capture the entire picture of professionalization of evaluation. In fact, Irby and Hamstra (2016) observed that one framework is not adequate in defining professionalism. My perspective as a researcher is that a combination of constructs drawn from all four approaches can adequately explain the reality of professionalization and professionalism.

The model situates professionalism in the context of four components drawn from the four constructs as shown in Figure 2.11 These are demand, structures, professionalism, and prestige, and status. The units of measurement in the model are professional associations and individual evaluators. I derived the meaning of professionalism from the subjective views of participants and used the collective understanding to empirically measure professionalism to determine the level of maturity of program evaluation in North America.

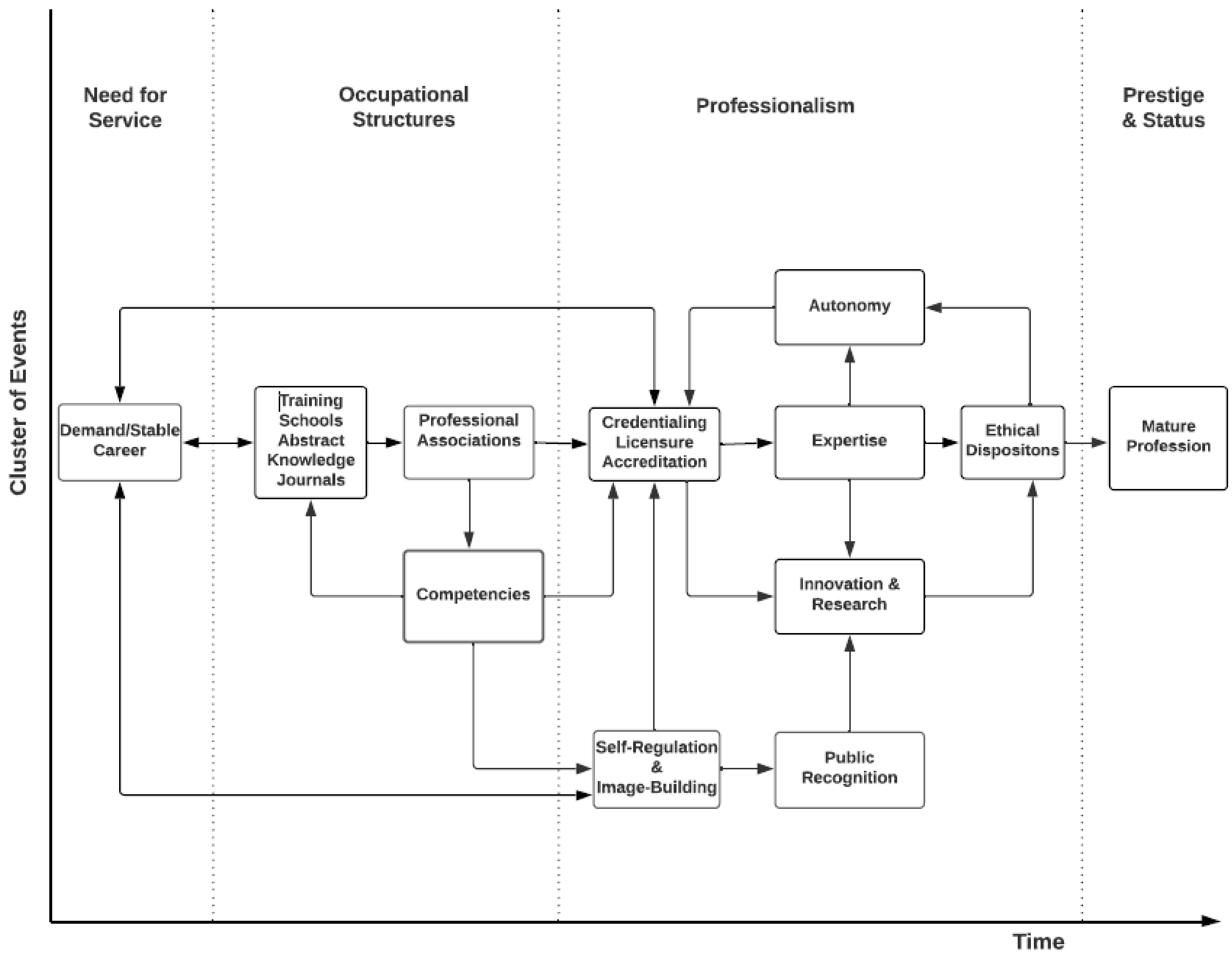

Figure 2.11 Conceptual Framework for Professionalization of Program Evaluation

The sequential development of clusters of events, the concept of service-task or demand, competencies and training linked to professional associations, credentialing linked to self-regulation and image building, public recognition, and sub-sets of professionalism (Autonomy, Credentialing, Innovation, Expertise, and Ethical Dispositions) are useful concepts in understanding the professionalization process for the field of evaluation. The conceptual model for this dissertation explores the theoretical linkages among the demand, structures, certification, and professionalism to explain the phenomenon of evaluation professionalization.

The one-directional arrow lines in the framework depicted in Figure 2.11 represents the sequence of events, and the double arrow lines suggest causation and dialoguing between events because social events are conflictual in nature (Van de Ven, 2007). This distinction provides the potential for a generalizable explanation of occurrences of events and potential areas for VOPEs to influence the process.

The first part of the model is the demand for evaluators and the existence of stable careers. The demand for program evaluators can be assessed from a number of sources: (1) the amount of federal and state funding allocated per year based on legislation in support of evaluation work; (2) the number of evaluators working in government, the private sector, and non-profits; and (3) the perception of employers based on Forsyth's and Danisiewicz's (1985) lens of evaluation service-task, i.e., essential (of serious importance to clients), exclusive (monopoly on the service-task), and complex (discretionary application of specialized body of knowledge). Although the model starts with demand, demand for services is assumed to influence the professionalization process, such as the need for credentialing evaluators, licensure, and accreditation. Hence it becomes a motivating factor for professionalizing the field.

The second part of the model examines occupational structures listed in the trait approach, which are a "must have" for VOPEs to lay the foundation for professionalization. These traits are training schools, professional associations, journals, abstract knowledge, and advanced training. I am theorizing a relationship between image building and demand for evaluators. For example, for certification to take place, there must be a demand for evaluators. Furthermore, the model suggests that professional associations establish evaluator competencies that in turn is related to image building. Altschuld's and Engle's (2015) historical timeline related to understanding accreditation, certification, and credentialing in Canada and the US shows the non-existence of accreditation, certification, and credentialing, although Canada has now established a credentialing body.

Competencies are pivotal in this second stage of the model since accreditation and credentialing/licensure are based on competencies. For example, while the Canadian Evaluation Society's evaluator designation and credentialing system use evaluator competencies to assess applicants, the European Evaluation Society (EES) and the United Kingdom Evaluation Society (UKES) uses the Evaluation Capabilities Framework for assessing evaluators applying to the Voluntary Evaluator Peer Review (VEPR) process. In addition, the model indicates that competencies or capabilities are related to training evaluators and accreditation of pre-service programs, credentialing, and licensure. As observed by Pavalko (1987), professionalism means that the practitioner knows the trade and is good at the job or is competent.

The third part of the model is professionalism, which is the core of my study. This part has four major components. Three parts are related to activities of the professional associations, including (a) certification: credentialing, licensure, and accreditation, (b) image-building and self-regulation, and (c) public recognition. Five traits of professionalism – Autonomy, Credentialing, Expertise, Innovation and Research, and Ethical dispositions – are related to the individual evaluator level. Although certification is one of the traits, it is treated separately in the model because of its contentious nature in the field of evaluation. Setting it apart makes it possible to assess how VOPEs are tackling it. Since the publication of the ECPE in 2005 (Stevahn, King, Ghere, & Minnema, 2005), the field of evaluation has remained stagnant in this stage for over 10 years.

My model shows that events that define professionalism occur both at the professional association level or VOPEs and at the level of individual evaluators. Professionalism is the outcome of a collective endeavor carried out by an occupational group (IOCE website, n.d.). At the professional association level, VOPEs engage in image building and in demonstrating to the state and the public their ability to self-regulate the practice through establishing standards of practice and ethics codes to protect clients and the public. Image building is central to the whole model, and it is a continuous process. For example, for credentialing to happen, VOPEs must engage in continuous image building activities and seek political legitimation through laws.

The commitment of VOPEs and individual evaluators to the protection of the clients and the public good could serve to convince the state to delegate power to VOPEs or the governing entity on condition that the occupation is an organized group, self-regulates, and that it will not abuse its privilege and power (Freidson, 1970). Brody and Doukas (2014) refer to the delegation of power by the state and associated privileges as a social contract between the occupation and society. The occupation promises society to render expert service for the common good of the public and to place the interests of the client or beneficiary foremost.

VOPEs must also institutionalize research and intellectual innovation by members of the profession in universities to develop new knowledge to promote the field. Funding for research is possible when there is public recognition of evaluation and its contribution to society. As mentioned earlier, successful public recognition of evaluation is at the core of the Global Evaluation Agenda. The model relies on the agency of VOPEs and the interests of key stakeholders like government and employers to make evaluation a profession. Without such a focus, professionalization of evaluation will be left to chance and may take an unpredictable amount of time to develop, thus prolonging the maturity of evaluation as a profession.

At the individual level, evaluators must demonstrate expertise and discretionary judgment in serving clients and society. Evaluators must win public trust, a concept that embraces the clients' or beneficiaries' trust in the credibility of the individual evaluator. For example, Brody and Doukas (2014) argue that professionalism requires knowledge and technical competence related to certification of competencies and, more importantly, requires "… the application of virtue to practice (based on virtue, deeper attitudes rather than mere behaviour, and requiring of practical wisdom)" (p. 980). Accreditation, certification, and credentialing are important steps in ascertaining the credibility of the evaluator's technical competence or expertise. The feedback loops between and among these professionalism traits signify relationships between the traits. My professionalism instrument assesses how evaluators perceive and demonstrate the attributes of Autonomy, Ethical Dispositions, Innovation, Expertise, and belief in Credentialing. I assessed the status of the field based on how evaluators perceived and self-reported these professionalism attributes.

The fourth and last part of the model is prestige and status. The status and prestige of the field of program evaluation can be assessed at two levels: mature profession if evaluators rank highly on the attributes of professionalism, or semi-profession if evaluators have low scores on the attributes of professionalism.

Advantages and Disadvantages of Professionalization. Professionalization of program evaluation has evoked strong feelings and positions from evaluation scholars and practitioners in North America and the rest of the world. Proponents for and against are divided on the benefits and risks of professionalizing the field (e.g., Bickman, 1997; Smith, 1999; Altschuld, 1999; Worthen, 1994; Gauthier et al., 2005, 2105). I have summarized the arguments for and against professionalization in Table 2.2. Overall, the viewpoints are legitimate and controversial, but more significantly, they provide ground for caution and much thought when engaging in professionalization activities. For example, the Canadian credentialing process took years of consultations with members of the association and relevant stakeholders (Gauthier et al., 2005, 2105). Despite the concerns about professionalization, the Canadian Credentialed Evaluator (CE) is the most advanced attempt to date in the professionalization of the field.

Table 2.2
Advantages and Concerns for Professionalization of Program Evaluation

Arguments for Professionalization	**Arguments Against Professionalization**
Strengthening the field and establishing some boundaries for evaluation	Credentialing is too controversial, and evaluation is too diverse to support it
Increasing training offerings in evaluation	Restricting or blocking access to the profession – alienate people based on their backgrounds and competencies
Enhancing and improving the status and prestige of evaluation	Credentialing and certification are too costly to evaluators and associations
Improved reputation of the field	Homogenizing evaluation and restricting diversity
Controls entry into the field	Reducing training offerings
Facilitating the selection of evaluators and improve the quality of conducted evaluations	
Protecting the public	
Trust from the consumers of evaluation service	
Avoiding problematic or unethical behavior	
Puts evaluation on par with other fields	
Ensures the minimum level of evaluator competency	
Could lead to higher revenues for evaluators	
Clear professional identity	
Safeguard against poor evaluation	

Note. Sources include Altschuld (1999); Jacob and Boisvert (2010); and Donaldson (2019).

Chapter Summary

In summary, I described and compared the histories of program evaluation and medicine using four approaches in the sociology of professions theory. I also described the theoretical conceptual model that is the foundation for the research questions and methodology discussed in the next chapter. In Chapter Three, I will describe the exploratory sequential mixed-methods case studies research used to answer the research questions. I shall explore the use of structured interviews, online surveys, and document review to collect data for this study.

CHAPTER 3

Methodology

The purpose of this mixed methodology design was to investigate the extent to which program evaluation is becoming a profession in Canada and in the United States (US) through comparing the degree of professionalism in evaluators in North America. The maturity of program evaluation as a field of professional practice is discerned empirically by assessing the behaviors of evaluators on professionalism constructs. Mixed methods design includes both quantitative and qualitative approaches whereby one approach is used to better understand, explain, or build on the results from the other approach (Mertens, 2015; Creswell, 2016).

Tashakkori and Creswell (2007) define mixed methods as “research in which the investigator collects and analyzes data, integrates the findings, and draws references using both qualitative and quantitative approaches or methods in a single study or program of inquiry” (cited in Mertens, 2015, p. 4). According to Mertens (2015), mixed methodology is eclectic. It is useful when the intent is to seek a common understanding through triangulating data from multiple methods. She argues that researchers often use mixed methods when the purpose of the research is complex. Creswell and Plano (2018) observed four defining core characteristics of the mixed methods approach: (a) collect and analyze both qualitative and quantitative data rigorously, (b) integrate (mixes or combines) the two types of data and their results, (c) organize these procedures into specific research designs, and (d) frame these procedures within theory and philosophy.

In this chapter, I shall present my research questions, the phases of the exploratory sequential mixed methods design, study context, participants, sampling plan, data analyses plan, and my positionality as a researcher. In the concluding section, I shall present the limitations of the study and a summary of the chapter.

Research Questions

This study was guided by three research questions:

1) How do program evaluators in Canada and the US describe professionalism in their everyday practice of evaluation?
2) How is professionalism in evaluators assessed?
3) To what extent and in what ways do program evaluators in Canada and the US demonstrate professionalism in their practice of evaluation?

Exploratory Sequential Mixed Methods Design

I used exploratory sequential mixed methods design in this study by exploring the perceptions of practicing evaluators on the meanings of theoretical constructs from the sociology of professions theories. The goal was to develop and test a new professionalism measurement instrument on a larger sample of evaluators in North America. According to Creswell and Plano (2018), the exploratory sequential design prioritizes qualitative data collection and analyses, followed by building from the exploratory results to develop a quantitative instrument, and finally testing the instrument on a larger sample. These authors identified four major steps of the exploratory design presented in the procedural diagram in Figure 3.1, with the key components being theoretical constructs and qualitative data collection, mixing or building a quantitative instrument, quantitative data collection, and integration of study findings to generalize or extend the initial qualitative findings.

Foremost, this study was based on the theoretical framework discussed in the previous chapter, particularly the five theoretical constructs of professionalism: Autonomy, Expertise, Ethics, Innovation and Research, and Credentialing. The theoretical framework is like a map or a travel plan that explains or suggests relationships between concepts, and it provides the critical ingredients or components of the phenomenon being studied (Sinclair, 2007, cited in Moore et al., 2012). As shown in Figure 3.1, I explored the meanings of these constructs by getting the perceptions of practicing evaluators in Canada and the US in the qualitative phase, then developed a quantitative measurement instrument using the qualitative findings.

The purpose of using the exploratory sequential design was to explore the meanings of professionalism constructs with a purposively selected small sample of individuals before engaging a larger number of participants in the quantitative phase (Creswell & Plano, & Leavy, 2017). As shown in Figure 3.1, a key product of the qualitative phase was the generation of a pool of 168 items for the quantitative instrument. The number of items reduced to 59 through a consultative process with experts in evaluation, and this quantitative measure of professionalism was then tested with a sample of 456 members of AEA and CES in the spring of 2018.

The products of the quantitative phase were a refined and validated instrument with 39 items used to measure professionalism in evaluators. Aside from getting a factor structure for professionalism, this final instrument demonstrated reasonable reliability and measures of fit. Finally, the products of the integration phase were analyses of qualitative and quantitative data using joint display tables. Fetters and Molina-Azorin (2017) define integration as “the linking of qualitative and quantitative approaches and dimensions together to create a new whole or more holistic understanding than achieved by either alone" (p. 293). Two years ago, Bazeley (2018) reiterates Morse's (2015) admonition about the risk of missing opportunities for gaining deeper insights when qualitative data used for generating quantitative measures are ignored after the statistical data are in. Morse urged researchers to pay attention to meta-inferencing from both data sets so that appropriate conclusions might be drawn. He wrote,

> Data from the first phase are not usually intended to stand alone, but nor should they be neglected after the survey or other instruments have been designed: if understanding gained from inductive qualitative work has guided the design of quantitative instruments, it will also necessarily guide the interpretation of data from those instruments, even if the second data collection is larger component of the study and statistical reporting is dominant. (cited in Bazeley, 2018, p. 75)

Integration occurred throughout the data collection and analysis. First, I identified quotations from the qualitative data on which the quantitative measure was built, what Creswell and Plano (2018) refer to as the "point of integration in mixing" (p. 87). Aside from discussing the extent to which qualitative and quantitative phases enhanced the validity of the measurement instrument, I also used joint display tables of quantitative and qualitative data (Fetters, in press) to integrate analysis. I provide a joint display table in Chapter 5 to provide inference from confirmation or discordance of the qualitative and quantitative data to better understand what professionalism means in the current discourse of the professionalization of evaluation in North America.

Rationale

The rationale for using a mixed methods approach in this study was to get operational definitions of professionalism constructs from the sociological study of occupations due to the newness of this research in the field of evaluation (Creswell & Plano, 2018; Leavy, 2017). The sequential mixed methods approach was a good fit for my study since I had an explicit theoretical conceptual model. The advantage of a mixed methods approach was its usefulness in collecting a richer and stronger array of evidence that could not be accomplished by a single method alone (Bazeley, 2018; Yin, 2009). In addition, Bazeley (2018) asserts that using mixed methods contributes to better understanding and stronger inferences through "designing better instruments and samples" (p. 12). Finally, the exploratory sequential mixed methods approach enabled me to report not only the more generalized global findings of the study, but also provide examples of individual cases or sub-group practices and contextualized practitioners' perceptions and attitudes specific to Canada and the US (Yin, 2009).

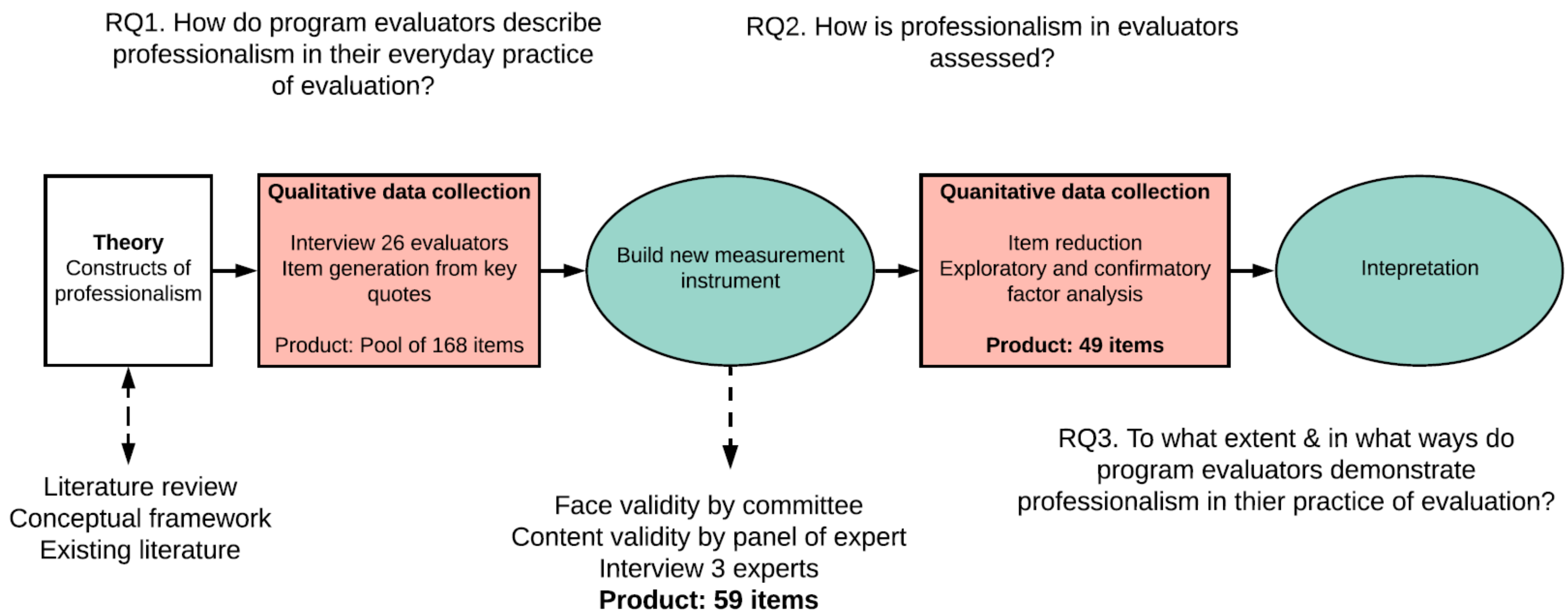

Figure 3.1 Procedural diagram of an exploratory sequential mixed methods study of professionalism in program evaluation.

Note: The figure includes the research questions for each phase of the study.

Exploratory Sequential Design Implementation Plan

This section presents the detailed implementation of the four phases of the exploratory design presented in Figure 3.1 and summarized in Table 3.1.

Phase 1: Qualitative Data Collection and Analysis. As noted earlier, the qualitative data collection and analysis was prioritized in the exploratory sequential mixed methods design, indicated by the notation $QUAL \rightarrow quan,$ signifying that qualitative methods had a greater emphasis in addressing the study's purpose (Creswell & Plano, 2018; Hesse-Biber, 2010). The aim of the qualitative study was to qualitatively explore and understand the meaning of professionalism from the participants' viewpoints.

Rationale

The rationale for using qualitative methods relates to the exploratory nature of this study and to what Schwandt (2017) declared as a "glaring absence of vigorous discussion of what professionalism means," especially in the current debate on professionalization of the field of evaluation (p. 546). Furthermore, the language of the items in the professionalism assessment instruments from the health sciences discussed in the next sub-section on item generation, existing instruments on professionalism does not speak to the needs or context of program evaluators. Patton (2015) observed that quality inquiry illuminates meaning: "Qualitative inquiry studies, documents, analyzes, and interprets how human beings construct and attach meanings to their experiences" (p. 13). In this study, the qualitative phase started with document review.

Table 3.1
Implementation Matrix (Sequential Design)

Phase	Study Aims	Data Collection Procedures	Data Analysis Procedures	Products on Outcomes	Point of Integration
Phase 1: Qualitative data collection and analysis	To obtain evaluators' language of professionalism in their daily practice of program evaluation.	Literature review of existing professionalism measurement instruments In-depth interviews: 26 program evaluators.	Qualitative analysis of themes and key quotes for item development	Items from quotations and key themes for analysis.	Building (using qualitative analysis and data to inform items for measurement instrument)
Item generation	To identify the initial pool of items from interview data and existing instruments	Document review and qualitative analysis	Review of the initial pool of items to ensure that they are clearly stated; conform to format; plausible response options and appropriate wording	Pool of items	Building
Phase 2: Build a new instrument	To develop a new instrument to measure professionalism	Review by the doctoral committee for face validity Review by a panel of experts for content validity	Online quantitative rating of the items using content validity index (CVI) with follow-up interviews with experts	The refined pool of items	Building
Phase 3: Quantitative data collection and analysis	Pilot the instrument with members of the professional associations and conduct factor analysis to identify items that measure the same concepts	Online survey of 456 members of AEA and CES	Exploratory and confirmatory factor analyses	Cluster of items for final statistical analyses	Building
Phase 4: Interpretation	Merge quantitative and qualitative results for comparison		Statistical analyses of survey data and coded qualitative data	Report on key findings	Meta-inference

Study Design. This qualitative study was based on interviews with practicing evaluators in academia, independent consulting, non-profit organizations and government and with the leadership of professional associations, especially people who were registered members of their professional associations since they would be more familiar with the development of the evaluation practice. For instance, the viewpoints of individuals from academia and training institutions were important because they play instrumental roles in the formation of abstract knowledge and the training of evaluators. Similarly, the views of the leadership of professional associations were important because these individuals are confronted with managing and directing change in the profession almost daily. They set goals and targets for professionalization. Finally, including the viewpoints of individual evaluators was important in understanding their opinions and experiences of Professional Autonomy, Ethical dispositions, Innovation and Research, and Credentialing from their practice of evaluation.

I used interviews because my interest was in getting in-depth experiences from individuals and their unique interpretations of professionalism in evaluation. Face-to-face semi-structured interviews yielded detailed insights, and they enabled flexible, in-depth exploration of the topics through probing to get clarifications from respondents. The goal of using this method was to increase the relevance and comprehensiveness of the data. The interviews were conducted from June to August 2017 mainly via Skype with participants across Canada and mostly in-person with Minnesota-based evaluators.

Recruitment. Table 3.2 provides a summary of the characteristics of people interviewed. Overall, the data collection consisted of 27 interviews with evaluators working in government and non-profit organizations, independent consulting, and academic institutions in Canada and the United States, and those who served in various capacities in the leadership of AEA and CES. Evaluators who were eligible to participate were those who worked and resided in Minnesota and in several provinces in Canada. These were people who practiced program evaluation and were willing to give informed consent.

Table 3.2
Summary of the Characteristics of Persons Interviewed

Country	Workplace	Number of Years in Evaluation Practice	Gender
Canada	Internal = 3 External = 5 Academic = 5 Manager = 1	Less than 1 year = 0 1-5 years = 0 6-10 years = 0 11-15 years = 0 16-20 years = 4 21-25 years = 2 26-30 years = 4 Above 30 years = 4	Female = 9 Male = 5
United States	Internal = 1 External = 7 Academic = 4 Manager = 1	Less than 1 year = 0 1-5 years = 2 6-10 years = 3 11-15 years = 2 16-20 years = 3 21-25 years = 1 26-30 years = 0 Above 30 years = 2	Female = 9 Male = 4

Evaluators in Canada were recruited with the help of key leaders in the association while those in the US and residing in Minnesota were recruited with the help of my primary academic advisor. The leaders of the professional association provided the names and email addresses of those they recommended for interviews. The names and email addresses of evaluators residing in Minnesota were obtained from the database of the Minnesota Evaluation Studies Institute at the College of Education and Human Development in the University of Minnesota.

Sampling Strategy and Sample Size. The participants for the study were recruited through purposive sampling. I strategically selected a small sample of information-rich individuals to provide insights and in-depth understanding of professionalism in evaluation (Patton, 2015). The purpose of using purposive sampling was not to represent the population, but rather to understand the phenomenon. I used “saturation or redundancy” strategy (Patton, 2015) to identify 26 practicing evaluators from different professional roles – internal, external, academic, and managers of evaluation to provide a range of opinions about the topic. My target was to recruit 30 participants, 15 each from the US and Canada. According to Benson and Clark (1982), “Usually interviewing 6–10 faculty members/content experts is quite adequate to determine the important relevant expert thinking” (p. 791).

With the saturation or redundancy approach, my interviewing process was interactive and iterative – I conducted preliminary analysis after each interview and stopped getting new participants for interviews after reaching saturation. Trotter (2012) states that the ideal standard for a qualitative sample is to interview to redundancy or saturation. According to Bernard (2011), redundancy is “the process of sequentially conducting interviews until all concepts are repeated multiple times without new concepts or themes emerging” (cited in Trotter, 2012, p. 399). The ideas that came up consistently in the interviews represented the core ideas that were considered important in professionalism in program evaluation in these two countries. These ideas were useful in operationalizing the definitions of the key concepts used in the model to develop a Professionalism Self-Assessment Tool for Program Evaluators (PAToPE).

Semi-Structured Interviews Process. I used semi-structured open-ended interview questions (See Appendix B) to get conceptual clarity on the constructs in my model. I asked two experts to review my interview questions, which I piloted with three practicing evaluators based in Minnesota. The interview protocol had three parts: (1) introduction and understanding of how the interviewee became an evaluator, (2) perceptions of the interviewee on the meanings of key concepts of professionalism, and (3) opinions of the interviewee about the future of program evaluation in Canada and the US. The key questions contained prompts for probing the participant's responses in depth. The prompts included phrases such as "tell me more," "how is that," "in what ways," and "anything else?" Patton (2015) asserts that "probes are used to deepen the responses to a question, increase the richness and depth of responses, and give cues to the interviewee about the level of response that is desired" (p. 465).

The purpose of the opening questions on how long the evaluator served and how they became an evaluator was to build rapport and assist respondents to feel comfortable with the interviewer and the interview process. I then briefly gave theoretical underpinnings of each construct as a way of familiarizing the participants with the concepts and then asked transition questions to move the participants into more in-depth discussion about the topic. The final questions on the interviewee's perceptions on the future of program evaluation in their countries and any final reflections on how the field is evolving in their countries were intended to get additional thoughts from the interviewee.

The interview questions were benchmarked on the indicators of the five constructs. The key informants were asked about their insights, perceptions, experiences, and opinions of these constructs based on their everyday practice of program evaluation. The contents generated from these interviews were organized into five themes of Professional Autonomy, Expertise, Ethical Dispositions, Research and Innovation, and Credentialing, which I used to guide item generation and analysis. The interviews were transcribed verbatim and each interview saved as a separate file.

IRB Process. The Institutional Review Board (IRB) of the University of Minnesota (STUDY00000797) approved this research. Prior to conducting the interviews, participants signed the consent form approved by the University of Minnesota IRB to comply with the codes of research ethics to protect human subjects. The consent form contained information about the purpose of the study, a description of what the participants would be asked, foreseeable risks of harm, and freedom to withdraw at any time. Although I appealed to the participant's altruism to incentivize them to participate, they had the freedom to choose to participate or not to participate in order to promote voluntarism.

Face-to-face interviews. I conducted face-to-face in-person interviews with 13 program evaluators in Minnesota and online interviews with 14 participants in Canada via Skype or Zoom. The order of the questions is provided in Table 3.3, and the duration of the interviews ranged between 45-60 minutes. I audio recorded and transcribed the interviews verbatim to preserve the original meanings.

Table 3.3
Semi-Structured Interview Protocol

Interview Question
Q1. How long have you worked as an evaluator and in what capacity (internal/external, an educator)?
Q2. Tell me about how you became an evaluator. What was the turning point in your career pathway to become an evaluator, if any?
Q3. Think back to the time you first started as an evaluator. What changes do you see in the field of program evaluation today? What has changed?
Q4. Professional autonomy a. To what extent do you feel that you have autonomy in your professional role as an evaluator? b. Did you always have autonomy in your practice, or did it emerge, if at all? c. Are there barriers or challenges limiting your professional autonomy as an evaluator?
Q5. Expertise a. What types of formal training were you involved in to develop expertise in program evaluation? b. What types of continuing education have you engaged in to keep current and/or develop new knowledge and skills? Please explain or give examples. c. What are some of the challenges in keeping up to date with new knowledge and developments in the field, if any?
Q6. Ethical dispositions a. In your opinion, what are examples of ethical standards that an evaluator should have? b. What is an example of the most difficult ethical challenge or ethical dilemma you faced as an evaluator, if any, and how did you resolve it?
Q7. Innovation and research a. Over the course of your career, have you identified a new method or approach for conducting evaluation? (Yes/No). (E.g., use of technology in data collection, data management, engaging stakeholders) b. If yes, please describe ways you think your method was innovative. How might it benefit evaluators or change the quality of evaluations? c. In what ways are you engaging in research to promote the field of evaluation? d. In what ways are you professionally engaged with the field of program evaluation?
Q8. Credentialing a. Are you familiar with CES' current work on credentialing for Canadian evaluators? b. If yes, what did it take to start the credentialing effort in Canada? c. In your opinion, what are the possible pros and cons of credentialing evaluators? d. What influence does credentialing have on (a) evaluators and (b) the clients they serve?
Q9. In thinking about the changes in the field, what do you see as the future of program evaluation in Canada (or the US)?
Q10. Is there anything else you would like to share from your experience or insight about how the field of program evaluation is evolving in Canada?

Data Analysis

My data analysis followed a mix of inductive and deductive approaches illustrated in Figure 3.2. First, I used the deductive approach to organize coding themes for the data based on the five theoretical themes derived from literature: Professional Autonomy, Expertise, Ethics, Innovation and Research, and Credentialing. Second, I used the inductive approach to organize concepts under each of the five themes during the coding process. Corbin and Strauss (2008) refer to coding as "mining" the data or "digging beneath the surface to discover the hidden treasures contained within data" (p. 2) or what Miles and Huberman (1994) refer as to "dissect" a set of field notes, transcribed or synthesized data.

Coding transcripts. I used MAXQDA Version 18.2.0 (VERBI GmbH, 2018) qualitative data analysis software to code and analyze the interview data. I purposefully selected two information-rich interviews, one from each country to get an overview of the data as shown in Figure 3.2. These two transcripts served as the foundation for coding 10 interview transcripts. First, I read each transcript vertically, line-by-line from the beginning to the end (Corbin & Strauss, 2008), and jotted on the margins word summaries of the phrases and paragraphs in the interview data. Some word summaries were derived directly from the data, or what Corbin and Strauss refer to as "In-Vivo Codes" (p. 2).

Second, I read the two transcripts horizontally, making side-by-side comparisons through open coding and axial coding process to make meanings of the codes. For example, the keywords participants used frequently to describe how they developed expertise in program evaluation were "*identifying strengths,*" "*experiential learning,*" "*university courses,*" "*improving skills,*" "*keeping up-to-date,*" "*changing behaviors,*" and "*collegiality.*"

Third, I assigned numbers to each interview, then randomly selected four interviews from the pool of interviews for each country to use for this study. I used the codes from the two analyzed interviews to code the eight interviews, generating a total of 86 codes. The coding process was iterative as some codes merged and others split after going through selected interviews. Fourth, I conducted pattern coding by grouping the word summaries into smaller numbers of categories or analytic units to explore the hierarchy of codes (Miles, Huberman, & Saldaña, 2014). The original word summaries on the margins of the first two interview transcripts were used as parent codes to retrieve and organize the data under each of the five theoretical constructs for analysis. For example, child codes containing words like "*self-taught*" and "*reading on my own*" and so on were coded under the "*experiential learning*" parent code. This process helped me to see patterns in the data and to write analytic memos and assertions from the emerging themes.

Code reduction. I constantly compared and validated the concepts in the data and kept methodological memos of my code reduction process during the focused coding process. As shown in Figure 3.2, the methodological and analytic memoing was an ongoing process that covered the entire process of coding and categorizing themes. As stated by Corbin and Strauss (2008), I labeled each memo with a concept that reflected my interpretation of the data. Furthermore, I used assertions from the memos for the initial qualitative analysis. One example of assertions from the qualitative data was that "*Enforcing the code of ethics by professional associations has a potential to stimulate* *professionalization of evaluation since most evaluators agreed on the importance of ethics in evaluation practice*." As noted earlier, the findings of this qualitative analysis were ultimately integrated with the quantitative findings in interpreting the results of the study. The concepts developed from the memoing process formed the basis for item generation for the quantitative instrument.

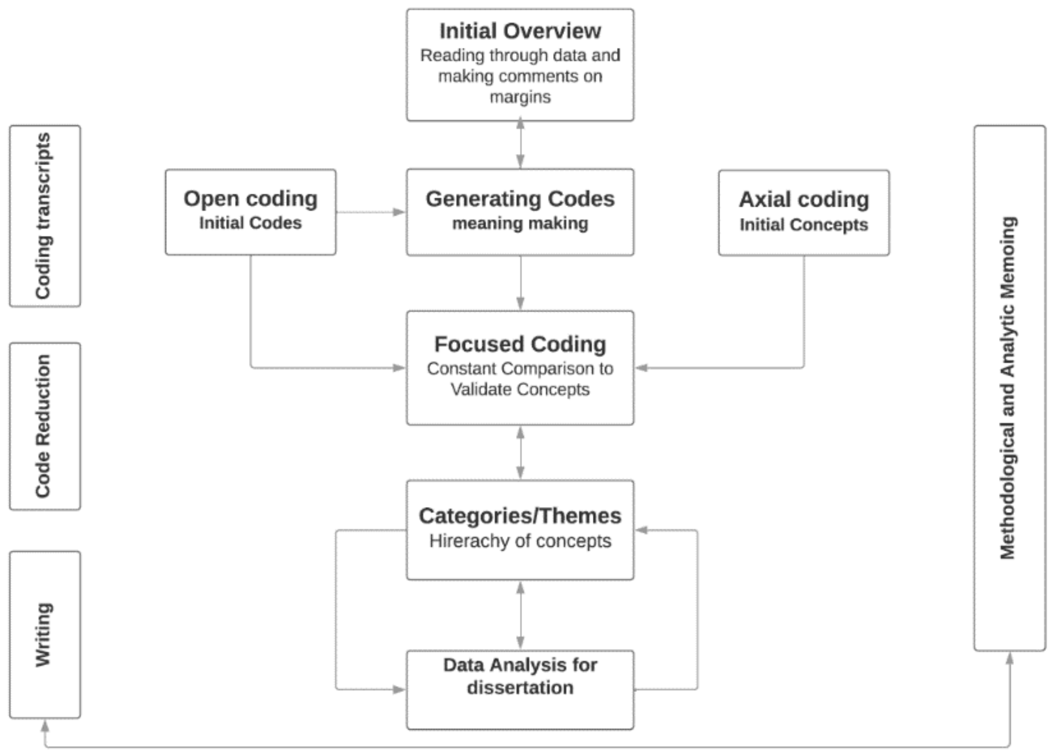

Figure 3.2 Process for analyzing interview data

Phase 2: Build a New Measurement Instrument

As shown in Table 3.1, developing a new measurement instrument for assessing professionalism was the second phase of the study. This phase was the first major point of integration in mixing qualitative and quantitative data through the strategy of building (Creswell & Plano, 2018; Fetters & Molina-Azorin, 2017). In this exploratory study, one database informed the data collection approach of the other. As noted in the previous section, the concepts from the memoing were translated into items for the measurement instrument. The instrument development consisted of four steps: (1) item generation; (2) item reduction; (3) piloting, and (4) quantitative psychometric item analysis, and (5) validation of the instrument (See Figure 3.3).

The purpose of the instrument was to assess evaluator professionalism in Canada and the US with the goal of assessing the level of maturity of the field of program evaluation in Canada and the US. More specifically, the objectives of the instrument were to assess:

1. The extent to which program evaluators are exhibiting professional autonomy in making technical decisions in their work
2. The extent to which program evaluators are demonstrating expertise and adaptability to master evaluation techniques, approaches, and theories in their work
3. The extent to which program evaluators are exhibiting ethical dispositions to guide their day-to-day practice
4. The extent to which program evaluators are aware of their level of competencies and desire proof of their competence on specific skills and knowledge to be deemed worthy of a credential
5. The extent to which program evaluators are demonstrating innovation and research to promote the field of program evaluation

Item Generation

I used existing instruments and the emic views of participants in the qualitative, open-ended, semi-structured interviews to inductively derive meaning and modify operational measures for the concepts being studied as shown in Figure 3.3. I generated the items from both existing instruments in literature and interview data because the use of literature and professional input alone was not enough (Rodenburg, Kuznicki, & Yiu, 2012). Scholars like Zeraatkar, Obeda, Ginsberg, and Hirsh (2017) argue that using items from previously developed tools is advantageous in that these items have already undergone some degree of validation and have demonstrated acceptable psychometric performance.

Literature review of existing instruments on professionalism. I conducted a literature search to identify existing instruments for assessing professionalism in evaluation and identified three instruments that were relevant to my study. Benson and Clark (1982) reason that undertaking a literature search has two benefits: (1) it ensures that an appropriate, reliable, and valid instrument does not already exist, and (2) it aids in formulating an operational definition for the constructs to be measured. These authors argue that "When a construct is operationalized, the components necessary to measure it are spelled out" (p. 791).

My initial search of PubMed and Ovid databases using key terms such as "measurement of professionalism" yielded 446 articles, only 426 of which had abstracts. Checking the abstracts for articles measuring professionalism reduced the search to 15 articles, but only four articles included instruments. Kelley's, Stanke's, Rabi's, Kuba's, and Janke's (2011) professionalism instrument was the more robust and well tested, with a reliability coefficient of at least 0.90 for most of the items. These authors developed and cross-validated a Professionalism Assessment Tool (PAT) to measure behavioral professionalism in pharmacy students. Their 33-item assessment tool included five domains of professionalism developed from the sociology of professions theory.

The PAT instrument was a collaboration of seven institutions,[8] including the University of Minnesota, and piloted in these same institutions. I requested and received the tool from the lead author, Katherine A. Kelley, who is based in the College of Pharmacy, Ohio State University. I also met with Professor Kristin Janke, who is currently a faculty member in the College of Pharmacy at the University of Minnesota. The items in the PAT instrument were tailored to meet the needs for measuring professionalism among pharmacists and were not very relevant to the purpose of my study.

My literature search produced only one professionalism instrument in program evaluation, that developed by Modarresi (1998). Modaressi's dissertation examined the status of program evaluation by assessing both the structural attributes of occupations at the broader level of the profession and attitudinal attributes at the individual level of program evaluators. Her 71-item survey instrument focused on assessing professionalism attributes such as professional autonomy, professional association, professional knowledge, professional training, licensure, the definition of work, and professional identity (Modaressi, 1998). I was able to use a few of Modaressi's items in my behavioral measures of professionalism in evaluators.

Finally, I also used some items from Forsyth's and Danisiewicz's (1985) professional autonomy instrument. Forsyth's and Danisiewicz's (1985) instrument

[8] Purdue University, University of Illinois-Chicago, University of Iowa, University of Michigan, University of Minnesota, The Ohio State University, and the University of Wisconsin.

comprised 22 Likert-type items: 11 items operationalized client autonomy in terms of decisional independence from clients and conviction of one's own knowledge, and 11 items operationalized autonomy from employing organization in terms of submission to organizational administration and loyalty or the person's willingness to bend organization rules.

Items Generated from Interview Data. The primary reason for the qualitative study was to generate items for developing the quantitative instrument. According to Rodenburg, Kuznicki, and Yiu (2012), there are three advantages of developing items from a target group such as practicing evaluators. First, it will directly be reflective of evaluators' experiences and perspectives and therefore have a high degree of face validity. Second, it is more likely to be acceptable to evaluators in terms of their content and experience level. And, finally, it is more likely to ensure that investigators and evaluators have a single and common understanding of the item and domain changes.

Prior to coding the transcripts, I read through each interview transcript and highlighted key quotations used to generate items for the instrument during the preliminary assessment stage. The items were phrases used by practicing evaluators to describe their experiences of the attributes of professionalism. Furthermore, I reviewed this initial pool of items to ensure that they were clearly stated. I shared these quotations with native English speakers to ensure that the items were plausible response options and the wording was appropriate since the items were verbatim quotations from the interviews.

There were five parts to the measurement instrument: Expertise, Autonomy, Ethics, Innovation and Research, and Credentialing. These sections were derived from the sociology professions theory. Overall, my intention was for each domain of the professionalism model to contain at least 8 to 11 items in the final instrument, a total of 40 – 55 items. Benson and Clark (2007) recommend that a pool of items should contain twice as many items as will be needed for the final form of the instrument. As such, I planned to generate at least 80 – 110 items for the five constructs from existing instruments and the interview data. Overall, I generated a pool of 168 items for all five constructs.

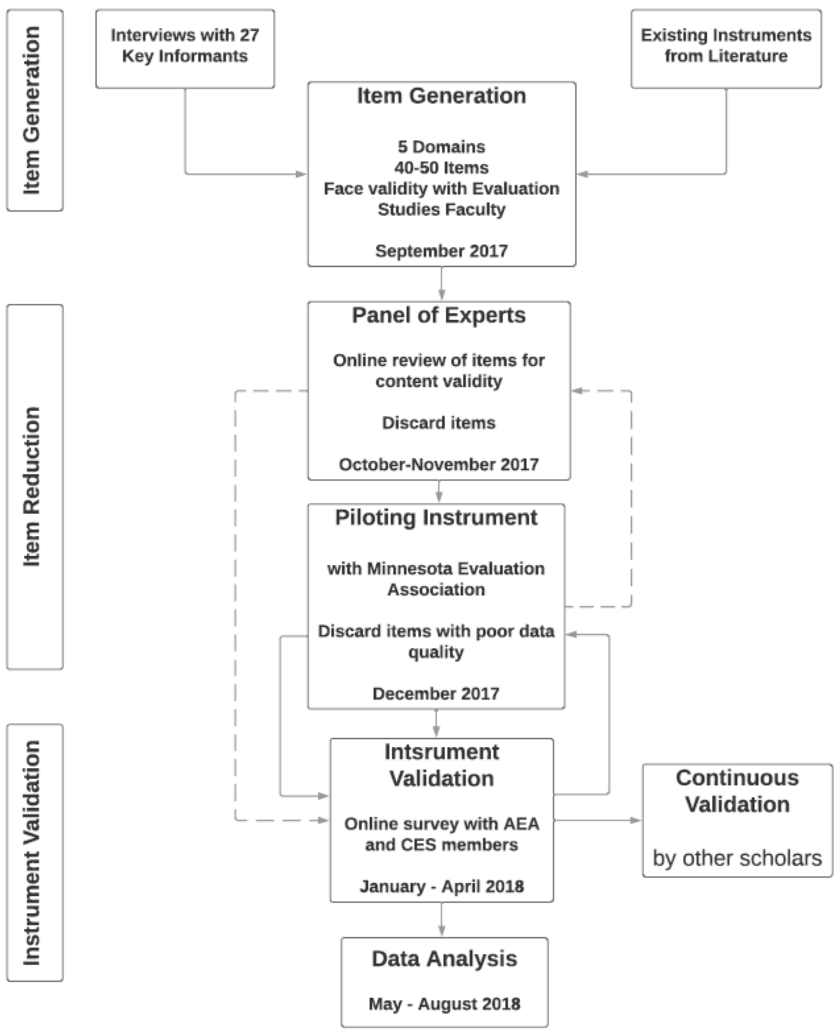

Figure 3.3 Process for developing and validating professionalism PAToPE

Item Reduction

There were two processes used for reducing the number of items: (a) face validation by members of my dissertation committee, and (b) content validation by an expert panel.

Face validity. The initial pool of items was reviewed for face validity by the Evaluation Studies faculty on my dissertation committee. Rodenburg et al. (2012) argue that "face validity" is that characteristic of an item which, "on the face of it," appears to be relevant to the outcome it intends to capture (p. 635). According to Benson and Clark (2007), this initial review is to ensure that (1) the items are clearly stated, (2) the items conform to a selected format, (3) the response options for each item are plausible, and (4) the wording is familiar to the target group.

Content validation by a panel of experts. The next level of validation of the items was by a panel of experts. A panel of eight subject matter experts participated in the review of the first draft of the comprehensive instrument in November 2017 for a period of ten business days, starting November 21 – December 4. The experts provided quantitative ratings on the relevance of each item and qualitative feedback on items they perceived as ambiguous or unclear and comments on the overall instrument. Experts were selected based upon their experience within the practice and teaching of evaluation and their experience with professionalization of the field. Four experts were from Canada, three from the US, and one from Europe. One expert ended their participation mid-way due to time constraints, and another did not rate items in the innovation and research and credentialing domains due to personal preferences. All reviewers had at least graduate-level credentials.

I shared a brief history of the instrument, a conceptual definition of each construct, and 168 items electronically with expert reviewers. The draft instrument was reviewed twice by the panel of subject matter experts. The first review was a quantitative rating of the items using the content validity index (CVI) approach and the second was a qualitative discussion. I distributed the survey items electronically through Qualtrics, the secure study implementation software hosted by the University of Minnesota. Experts were asked to rate the items for relevance to the professional evaluation practice using a four-point scale of *1 = not relevant, 2 = slightly relevant, 3 = somewhat relevant, and 4 = highly relevant*. I used the scale to calculate the Content Validity Index (CVI) for each item. I also asked the experts to suggest possible improvements or changes to the items that they checked "somewhat relevant."

I calculated the item content validity index (I-CVI) by combining the 3-4 responses for each item, then dividing the total by the number of reviewers for the item. The I-CVI is the proportion of experts who rate the item "Somewhat" or "Highly" relevant, that is, the proportion in agreement about relevance (Polit, Beck, & Owen, 2007). I used the I-CVI score of 0.70 as the cutoff for individual items since there were more than seven reviewers, reflecting at least two disagreements or a cutoff point of 0.75 for items with eight reviewers, reflecting at least two disagreements. The items below this cutoff index were flagged for revision or deletion. Items in two constructs (innovation and research, and credentialing) were reviewed by six reviewers, thus with I-CVI of 0.67 as the cutoff point. However, Benson and Clark (1982) cautioned that it is easier to improve a bad item than to develop a new one. Items below the cutoff index were carefully considered before being deleted.

In summary, the number of items of the revised instrument was reduced from 168 to 82 items – 88 items were removed, while others were flagged for further review. The items in the Autonomy domain reduced from 41 to 16, items in the Expertise domain reduced from 39 to 24, items in Innovation and Research reduced from 19 to 10, items in the Credentialing domain reduced from 30 to 10, and items in Ethical Disposition domain reduced from 39 to 19.

Aside from the quantitative rating, the experts provided qualitative feedback about confusing or difficult items, changes that could be made, and suggested additional changes to the instrument (Polit et al., 2007). Finally, I conducted a follow-up interview with three expert reviewers to discuss the CVI ratings and the qualitative feedback. The purpose of the follow-up was to gain feedback as to the meaning behind the quantitative ratings and comments provided in the survey. The reviewers suggested the wording of some items while I also had a chance to conduct a “think-aloud” with two of the reviewers based in Minnesota. This process proved invaluable considering that I did not have a large enough sample to pilot the instrument.

Piloting. I had planned to pilot the revised instrument with members of the Minnesota Evaluation Association (MNEA) and to use the data from the pilot study to conduct psychometric item analysis to establish the reliability of PAToPE instrument as a whole and the suitability of each item. I failed to pilot the instrument because the population of the active members of MNEA was much smaller than expected. Benson and Clark (1982) recommended a large group for the pilot test in order to avoid distortion of the results of the quantitative evaluation, which can be caused if the pilot is administered to only a few subjects. In fact, I needed a sample of at least 820 people for the pilot study based on Kline's (2016) recommendation of at least 10-20 participants per item to do psychometric analyses. I shall return to this distinction when discussing the measurement invariance of the instrument, but it is worth noting that I used the quantitative data from the first validation of the instrument with members of AEA and CES as the baseline to conduct the psychometric analyses. The aim of psychometric validation was to increase confidence in the findings in measuring professionalism and the use of the measurement instrument in further studies (Adams & Wieman, 2011).

Validation

Validation of the instrument is the third major step shown in Figure 3.3. The PAToPE instrument was validated at the beginning of spring 2018 through an online survey to members of AEA and CES. Benson and Clark (1982) argued that the validation of a newly developed instrument is almost never accomplished through one study or by one researcher. I anticipate that my study will generate interest in the professionalization of evaluation and that the instrument will be validated continuously by other researchers for use in their own studies. This first online validation of the instrument was considered the third phase of my mixed methods approach. This phase is discussed below.

Phase 3: Quantitative Data Collection

Rationale

The purpose of the online survey to members of AEA and CES was two-pronged. The first purpose was to validate the instrument and test the reliability and psychometric qualities of the items in the instrument so that items with poor data quality were discarded to achieve the original goal of 40-55 items in the final instrument. Second, I used the data to answer my research questions.

Sampling Strategy and Sample Size

I received a random sample of 1,000 evaluators from AEA and the total population of 573 CES Anglophone members from the 2017 membership directories as the populations of interest. The criteria for selecting AEA members were individuals who practice evaluation in North America. However, the random sample included 90 individuals who were international members of AEA. These responses were excluded from the analysis. The criteria for the selection of the CES sample were individual members in good standing, who agreed to third-party research and surveys, and were only English or Anglophone members. In addition, I advertised my study on the CES website and received 15 responses and sent the survey to all the active and inactive members of the Minnesota Evaluation Association (MNEA). Overall, 269 AEA members responded to the survey, a response rate of 27%. There was a 20% response rate for CES members.

Demographic Characteristics of Participants

The demographic characteristics of the online survey respondents are summarized in Table 3.4. The respondents represented evaluators from various employment backgrounds: internal evaluators working with governments, foundations, and non-profits; external evaluators working in the private sector or self-employed; faculty in academia, and commissioners or managers. Overall, 64% of respondents identified as Americans, 30% Canadians, and 5% as international members of AEA (See details in Appendix D). Participants ranged in age from 34 to over 65 years old. There was some diversity in the survey participants; while nearly four-fifths of the US (77%) and Canadian (78%) participants identified as white and female (71%), a large proportion of participants (90%) had post-graduate education. Nearly two-thirds identified as proficient/skilled evaluators (62%) compared to one-fifth who identified as expert/master evaluators.

Table 3.4
Participant Description

Country	Workplace	Number of Years in Evaluation Practice	Gender	Credentialed Evaluators
Canada	Internal = 41 External = 36 Academic = 16 Manager = 15	Less than 1 year = 0 1-5 years = 28 6-10 years = 24 11-15 years = 19 16-20 years = 14 21-25 years = 12 26-30 years = 10 Above 30 years =10	Female = 80 Male = 35	CE = 44 Male CE =14 Female CE = 28
United States	Internal = 82 External = 67 Academic = 54 Manager = 17	Less than 1 year = 2 1-5 years = 61 6-10 years = 56 11-15 years = 34 16-20 years = 28 21-25 years = 28 26-30 years = 14 Above 30 years =25	Female = 183 Male = 56	CE = 1

Note: This table is based on participants who provided information on these demographic variables. Overall, 387 individuals completed the survey, 248 Americans, 118 Canadians, and 21 international evaluation members of AEA.

Online survey

I administered the instrument as an online survey to members of the professional associations in Canada and the US (n= 456). The survey was administered over a four-week period via email. According to Dillman, Smyth, and Christian (2014), “Surveys that are completely electronic, relying only on e-mail contacts to obtain Internet responses, are the fastest growing form of surveying occurring in the United States, as well as throughout most of the world” (p. 301). These authors argue that the web is especially attractive because of speed, low cost, and economies of scale.

The participants were asked to respond to statements that matched their feeling of professionalism on a four-point Likert scale ranging from strongly disagree to strongly agree as follows: Strongly Disagree = 1, Disagree = 2, Agree = 3, and Strongly Agree = 4 (see Appendix A for the instrument). The instrument is intended to tap the degree to which respondents disagree or agree with statements for the individual evaluator. As such they tap self-perceptions in each of the domains. Evaluators with more of the trait will score higher on the professionalism scale. The higher the score, the more positive the evaluator’s perception of their level or degree of professionalism. The rest of the survey questions are demographic, asking years of practice, disciplinary identity, formal training, evaluation schools attended, current job, and familiarity with evaluation standards, guiding principles, and AEA’s or CES’ statement of cultural competence in evaluation.

Data Analysis

I used semTools and lavaan (v. 0.6-1) packages, the Statistical Package for Social Sciences (SPSS) version 24 (IBM, 2018), and Mplus Version 8.2 software to conduct four types of analyses: (a) descriptive statistics in R and SPSS; (b) reliability assessment in R and SPSS; (c) exploratory and confirmatory factor analyses in SPSS, R and Mplus; and (d) structural equation modeling (SEM) for path analysis and measurement invariance in R and Mplus. First, I cleaned and explored the quantitative data in SPSS to assess missing data, recode the data for analyses, obtain descriptive statistics, and conduct statistical assumptions of normality, randomness, homogeneity of variance, outliers, and independence in order to determine appropriate statistical tests for the data. The Shapiro–Wilk Test in SPSS was used to test the assumption for normality since the sample was small (Leach, Barrett, & Morgan, 2011). I observed if the variables of interest were symmetric around its mean value, had a kurtosis equivalent to zero, and a range of ± 1 skewness. The assumption of randomness of observations was tested using the Runs test in SPSS; outliers were identified using boxplots; homogeneity of variances was tested using Box's M test and Levene's test (Verma & Abdel-Salam, 2019).

Second, I conducted both the exploratory factor analysis (EFA) in SPSS and Mplus to validate the underlying structure of the new scale and confirmatory factor analysis (CFA) in Mplus and R to assess the scale quality of the measurement model. Aside from validating this new instrument, I also wanted to reduce the number of items in the instrument by eliminating those with poor data quality and to identify a good measurement model (Rodenburg et al., 2012; Adams & Wieman, 2011). These authors argued that developing a good instrument involves selecting items that are good indicators of factors and ensuring that factors are well-measured.

I randomly split in half the sample of 456 respondents to conduct the EFA and CFA. Besides evaluating the fit indices, I also used the size of the factor loadings to assess the quality of the model and, more specifically, to identify items that were not contributing to the factor structure. Different fields of study have varying guidelines on how large factor loadings should be. I retained items with a rotated factor loading of at least .40 as the cutoff minimum loading onto the factors (Matsunga, 2010). Furthermore, I used the comparative fit index (CFI), the root means squared error of approximation (RMSEA), and the standardized root mean squared residual (SRMR) to assess model fit for the CFA analyses. The cutoff for the fit indices was CFI $\geq$ 0.95, RMSEA $\leq$ 0 .06, and SRMR $\leq$ 0.08 (Hu & Bentler, 1999).

Third, I examined the reliability of the instrument using Cronbach's coefficient alpha (Cronbach, 1951) in SPSS. Benson and Clark define reliability as "the consistency of the measurement over time or the precision of the measurement," (p. 795). The cutoff for reliability measure was coefficient α = .80 to .90 (DeVellis, 2012; Adams & Wieman, 2011; Benson & Clark, 1982). Although Cronbach's Alpha is the most commonly used coefficient, there are a number of well-documented problems with it (Cortina, 1993; Green, Lissitz, & Mulaik, 1997; Green & Yang, 2009; Rodriquez & Maeda, 2006; Sijtsma, 2009a, 2009b; Terwilliger & Lele, 1979; Cronbach & Shavelson, 2004, as cited in Davenport, Davison, Liou, & Love, 2015), which can often be overcome by using an alternative coefficient, such as Omega (Dunn, Baguley, & Brunsden, 2014; Trizano-Hermosilla & Alvarado, 2016) or coefficients based on SEM (Davenport et al., 2015).

Fourth, I conducted the multigroup CFA in SEM to establish measurement invariance, which means that the survey is measuring Canadians and Americans in the same manner (Hirschfield & von Brachel, 2014) and to get a basis for drawing inferences from the scores of the instrument (Kline, 2016). I used CFI, RMSEA, and SRMR along with Hu and Bentler's (1999) guidelines to assess four types of invariance: (a) configural invariance; (b) weak (pattern or metric) invariance; (c) strong (scalar) invariance; and (4) strict invariance (Wu, Li, & Zumbo, 2007, cited in Kline, 2016). Furthermore, I did a path analysis of the structural model since the measurement model yielded satisfactory fit and reliability indices. I used the professionalism model in my conceptual framework to develop the path of the structural model and examine the correlates between Expertise, Ethics, Autonomy and Credentialing using Mplus and lavaan package in R.

Finally, I conducted univariate and multivariate statistical analyses to see if there were differences in professionalism in evaluators by country, workplace, credentialed evaluator status, and gender. I merged the quantitative and qualitative results for comparison in the fourth and last step of the mixed methods design. The results of these analyses are presented in the next chapters.

Phase 4: Interpretation

The interpretation of the entire analysis was done with both data sets through triangulation and making meta-inferences using joint display matrices to integrate the findings (See Figure 3.2). The triangulation model gives equal weights to qualitative and quantitative evidence. Creswell (2009) observed that in triangulation, the researcher collects both quantitative and qualitative data sequentially and then compares the two databases to determine if there are convergence, differences, or some combination.

Study Sites

The setting was professionalization of program evaluation in North America using evaluator competencies, evaluation standards, and ethical guidelines. Canada is currently the only country using voluntary credentialing while the US is the oldest country where program evaluation was first formalized as a field of professional practice. Canada is more experienced in professionalization of program evaluation, and the US is older in terms of creation of the field of program evaluation. I compared professionalism in the two countries in order to determine the status of the field in each country.

Canada

My theoretical conceptual model shows that professionalization is a complex and sometimes a very contentious process. Despite protracted debates on why or why not to professionalize evaluation (Picciotto, 2011; Jacob & Boidvert, 2010), Canada in 2010 was the only country that started implementing a voluntary credentialing program. Dumaine (2014) pointed out that despite the challenging, difficult, chaotic, and nerve-racking process that led to the adoption of the CE program, CES courageously faced a barrage of initial criticism and opposition from its members to establish the program.

These challenges were like those faced by AEA (Picciotto, 2011), but what makes Canada an interesting case is that its leadership took a risk and started a process that made Canada the first country in the world to begin an evaluator credentialing program. Canada has made a tremendous effort to professionalize evaluation practice through credentialing and maintaining a public registry or roster of Credentialed Evaluators for each province.

This competency-based credentialing system was launched by CES in May 2010 as a service for its members (Kuji-Shikatani, 2015). The CES credentialed evaluator designation is offered to members, who may elect to become credentialed on a voluntary basis. Their portfolio is reviewed by members of the Credentialing Board, which is made up of senior evaluators who have on average 28 years of evaluation experience

(Barrington, Frank, Gauthier, & Hicks, 2015). A member of CES must fulfill three conditions to qualify as a designated Credentialed Evaluator: (a) evidence of a graduate-level degree or certificate related to evaluation; (b) evidence of two years (full-time equivalent) evaluation-related work experience within the last 10 years; and (c) indicators of education and/or experience related to 70% of the competencies in each of the five domains of Competencies (Canadian Evaluation Society, 2010b).

The credentialing process involves submission of an online application by the applicant upon paying a $485 non-refundable fee; the Credentialing Board reviews the application and makes a decision either to award the CE designation or notify the applicant with instructions on how to qualify. Finally, the CES National Council awards CE to applicants who have met all criteria, and the awardee's name is entered into the CE Database. The CES maintains a registry of CE available to the public on the CES website. The CE evaluator is then required to meet 40 hours of ongoing learning and professional development every three years to maintain their status. The Credentialing Board members are experts in evaluation who are recognized as CES Fellows and National Award winners (Kuji-Shikatani, 2015). Applicants with "further preparation needed" notification may appeal to the Vice President of CES within 30 days of notification if they are disgruntled. There are currently 344 Credentialed Evaluators on the CES website, an increase of 34% from the total of 256 CE in 2015.

In addition to launching the CE program, some Canadian universities are now using the CES evaluator competencies to design pre-service training programs. Dumaine (2015) noted that "… an increasing number of CUEversities (sic), curriculum planners for academic programs in evaluation have been able to take into account the three pillars of the CE designation…" (p. 157).

The United States. The US is the oldest country where program evaluation was first formalized as a field of professional practice, but also a country with the most controversial stakes in the professionalization of evaluation (Picciotto, 2011). The US was the site of one of its state's (Louisiana's) credentialing system for educational evaluators in the 1980s, but this died potentially due to the controversies and divisions among the evaluators on professionalizing program evaluation (King, personal communication, April 12, 2017). As pointed out by Picciotto (2011), "Heated doctrinal disputes within the membership of the AEA have blocked progress in the USA" (p. 165). For example, Len Bickman, a former president of AEA, urged members of AEA 20 years ago to take a bold move and professionalize the field. Bickman said, "We need to move ahead in professionalizing evaluation, or else we will just drift into oblivion" (Bickman, 1997, p. 8).

The American Evaluation Association has remained divided over this issue for decades, but fortunately, as King and Stevahn (2015) noted, "The time has come at last for the field of program evaluation in the United States to address head-on an issue that scholars and the leaders of professional evaluation associations have discussed periodically for over 30 years: … competencies ..." Competencies are central to professionalizing the field because they were the foundation on which Canada launched its voluntary evaluator credentialing system. Owing to this, the AEA Board instituted a task force headed by Professor Jean King in 2015 to develop competencies for members of the AEA. It was anticipated, as Stevahn et al. (2005) noted, that "the ECPE offers a pragmatic and more general solution to the challenge of improving evaluation education" (p. 109). In fact, a recent study of AEA by Donaldson (2019) revealed changes in the attitudes of AEA members to favor professionalism. Donaldson writes that,

> Results from in-depth interviews and an online survey demonstrate that AEA members view potential benefits of professionalization to be stakeholder trust, evaluator reputation and, identity, while concerns about a potential negative side effect known as the "narrowing effect" (i.e., some evaluators will be alienated based on their background, competencies, etc.) were expressed by participants. (p. 152)

The history of program evaluation in the US, doctrinal disputes, and renewed interests in professionalization make the US an important case for this study.

Researcher Positionality

Patton and Bondi (2015) encourage researchers to reflect on their positionality and how it affects the conceptualization and execution of their studies. I have inserted my voice as a practicing evaluator and scholar in the design of this study and continued to do so throughout the phases of the study. I am a mid-career professional in international practice and an independent external program evaluation consultant. As such, I understand the research participants and the topic, i.e., the need to professionalize program evaluation.

I started practicing program evaluation as an external evaluation consultant in the private sector in 2001 without the slightest idea or knowledge that program evaluation was an emerging field of professional practice. I did program evaluation consultancies as a part-time and sideline job until 2008 when I practiced fulltime as my sole source of income. The first time I ever heard that program evaluation was a professional field of practice was in March 2012 during the annual spring training organized by the Minnesota Evaluation Studies Institute (MESI) on the University of Minnesota's St. Paul campus. MESI 2012 the turning point in my career pathway and subsequently led to my enrollment in the Ph.D. Program in Evaluation Studies in CEHD. Now I can proudly tell people that I am a program evaluation professional, although not everyone understands what program evaluation is.

My Personal Educational Trajectory

I am an African woman born and raised in Uganda. I am the first college and graduate school professional in my family and in Karamoja, my region of birth. I received undergraduate education from Makerere University in Kampala, the capital city of Uganda, and later got a master's degree from the University of East Anglia in the United Kingdom where I earned my master's degree in Rural Development (Planning).

These two academic qualifications enabled me to get reputable jobs in the public sector and to work as an independent evaluation consultant on public policy, gender, peace, and security in Africa. I have worked for major nonprofits in the area of policy research, program design, and monitoring and evaluation. I was introduced to program evaluation work while supervising evaluation consultants as a program officer. My strength has been in qualitative research, and I wanted to strengthen my quantitative skills by doing a minor in Quantitative Methods in Education (QME). I wanted to create a valid and reliable instrument for my discipline as a content expert. I have always loved critical thinking and analytical work, which helped me to approve consultants' reports prior to them receiving payments. This meant ensuring the quality of the product and usability of the reports. I have since worked as a researcher and program evaluation consultant for nine organizations in Africa.

I understand that professionalizing the field may be threatening to thousands of people who are practicing evaluation without the slightest idea that this is a field of professional practice or the requisite training for practicing evaluation. In discussing the case for a voluntary system for credentialing evaluators, Altschuld (1999) made a point that “grandfathering” or “grandparenting” may be contentious. This is a concept of creating a window for including non-credentialed long-term serving evaluation practitioners in a new system that would require them to be credentialed. My study might be perceived differently by evaluation practitioners, but I was hopeful that it will not limit their participation.

Limitations of the Methodology

As mentioned earlier, my theoretical conceptual framework will be used for generalizing the subsequent results of my study and to limit the scope of the study to the five broad concepts that constitute the ingredients of the phenomenon, which set priorities for data collection. The case studies will contribute to a cumulative body of knowledge that will inform professionalization of evaluation instead of isolated empirical studies.

Creswell and Plano (2018) identified three methodological issues and validity threats in using an exploratory sequential design: (a) the quantitative-designed feature does not build on qualitative findings; (b) the quantitative-designed feature is not rigorously developed, and (c) the quantitative sample consists of the same individuals in the qualitative sample. I addressed these concerns as follows: (a) used the findings from the qualitative study on which the quantitative-designed measurement instrument was built; (b) used systematic procedures to design the quantitatively-designed measurement instrument with good psychometric fit indices and reliability indices; and (c) the quantitative sample of 456 individuals was much larger than the qualitative sample of 26 individuals, thus, not the same individuals. In fact, I split the data in half and used the sample of 232 for EFA and used the other half (n=224) for cross-validation with CFA (Fokkema & Greiff, 2017). Furthermore, the instrument was rigorously developed through four stages and, to address the limitations of mixed methods research, I triangulated data from several sources. I also engaged internal and external reviewers in the measurement development procedures (Creswell, 2013, 2016).

Chapter Summary

The purpose of this chapter was to provide a structure of the methodology for my study. In the first sections, I provided an overview of my study and my research questions. In the second section, I offered a description of the sequential mixed methods approach using Creswell's and Plano's (2018) typology and a detailed process of how the measurement instrument was developed and validated. I also discussed the research context, description of participants, and the IRB process. Finally, in the last section, I discussed my positionality as a researcher and the limitations of this mixed methods case study. This study has generated valid findings that can build scholarship in the areas of professionalization of program evaluation. I created a professionalism assessment tool that can be used by other scholars who are interested in research on evaluation. I now discuss the qualitative findings of my study in the next chapter.

CHAPTER 4

Qualitative Analyses and Findings

Introduction

This chapter presents the results of the qualitative data analyses of each theme derived from the sociology of professions. The chapter has two sections:

1. First, I present the descriptive information of the language program evaluators used in relation to the five themes of professionalism derived from the sociology of professions theory.
2. Second, I present the integration phase of the study in which I developed PAToPE through building on the qualitative findings from interview data (Creswell & Clark, 2018).

This study was guided by three research questions: (a) How do program evaluators in Canada and the US describe professionalism in their everyday practice of evaluation? (b) How is professionalism in evaluators assessed? and (c) To what extent and in what ways do program evaluators in Canada and the US demonstrate professionalism in their practice of evaluation? The first question was addressed with responses from the qualitative interviews while the second with the quantitative responses to the online survey to validate the instrument, and the third question was answered with both the quantitative and qualitative data.

Research Question 1

How do program evaluators describe professionalism in their everyday practice? The purpose of this research question was to explore how program evaluators describe professionalism in their everyday practice of evaluation. I now describe each domain and provide a summary table illustrating descriptors of themes.

Sub-Scale 1: Expertise. Evaluators described how they developed expertise in program evaluation, mastered techniques, and kept up-to-date with new knowledge through various means. Expertise refers to the competence or proficiency in conducting program evaluation encompassed in evaluator competencies. Contrary to the theoretical assertions that link acquisition of expertise in a trade to prolonged specialized education at the tertiary level (Evetts, 2003; Picciotto, 2011), the codes in Table 4.1 are consistent with the participants' responses to a demographic survey question that revealed that 83% of evaluators acquired expertise in evaluation informally, compared to 17% who reportedly acquired it through completing university courses.

Evaluators reported developing expertise through ongoing "*self-assessment*" of their strengths, engaging in experiential learning or "*learning-by-doing,*" participating in "*on-the-job training,*" and learning from thought leaders in the field. This process differed for senior and novice evaluators. Older evaluators who started their careers in evaluation during the formative years of the field reported being "*self-taught*" because there were no evaluation courses at the time. Some of these participants' experiential learning contributed to the "*development of the courses*" in evaluation. Novice evaluators, on the other hand, cited working with "*mentors*" and working as an apprentice alongside expert or more senior evaluators on projects to develop their expertise.

Nearly every evaluator cited improving their skills in evaluation through "*reading on their own*" and facilitating or attending "*professional development*" activities, usually through professional associations. Such activities included pre-conference workshops and short courses offered by professional associations. These professional development activities overlap with the code on keeping up-to-date with new knowledge in the field. Nearly every participant cited that they keep current through formal meetings such as attending professional conferences and informal occasional meetings with other evaluators. Not surprisingly, fewer evaluation practitioners reported reading or submitting journal articles than academic evaluators. In fact, more academic evaluators cited integrating empirical research in their work compared to evaluation practitioners.

This difference distinguishes how different evaluators perceive expertise based on their workplace. Some academic evaluators in the sample argued that people do not become experts by attending professional development activities, but through completing university courses. For example, one interview participant distinguished professional development from university training, noting, "When you offer a professional development course, we are not trying to create an expert, like you are just trying to give someone a couple of new skills … [T]he university course is to give people – help develop some competencies that they can apply of course in their work ... to apply more broad competencies." Overall, academic evaluators linked expertise to gaining knowledge of evaluator competencies through long-term university training for application in evaluation work.

Aside from developing expertise and keeping current with new knowledge in the field, evaluators also discussed collegiality with peers as a form of occupational solidarity or loyalty to the profession. As noted in the ancient adage that "iron sharpens iron," evaluators cited dependency on one another for referrals and peer support. In the process, they noted that they changed their professional behaviors by "*talking to other evaluators*" or “*asking other people questions*” about their practice. Participants explained that they “*refer clients to someone who is an expert*” if they are not familiar with certain evaluation approaches.

Table 4.1
Program Evaluators’ Description of Expertise in Evaluation

Code	Description	Excerpts from Interview Data
Identifying strengths	Evaluators identified professional strengths through self-assessment	“I document and keep track of my professional development hours on a regular basis”
Experiential learning	Evaluators developed evaluation skills through practicum or internships, on-the-job training, learning by doing, mentoring, learning from thought leaders, and teaching evaluation	“So, my expertise was developed on-the-job as well as training events and conferences and reading…,” “… my approach was to learn by doing,” “I am basically self-taught,” “I track...thought leaders …,” “…developing the courses and conducting the seminars…,” “I had a mentor...”
University courses	Evaluators developed expertise through university courses or training	“I had formal university training,” “…primarily from coursework …,” “…my first formal evaluation course was in graduate school”
Improving skills	Evaluators discussed how they improve their skills and master evaluation techniques through own reading, professional development workshops, doing research on evaluation, and teaching evaluation courses	“… doing research on evaluation…,” “…reading on my own,” “I provide professional development courses…,” “I supplemented it with little short courses in evaluation offered by the professional association…”

Table 4.1 (Continued)

Code	Description	Excerpts from Interview Data
Keeping up-to-date	Evaluators kept up-to-date by meeting informally, meeting formally through attending conferences and professional development workshops, integrating empirical research, reading evaluation-related journals, and submitting manuscripts to journals	"…I attend the AEA conference, the Canadian conference, I attend conferences around the world," "… meet occasionally … with other evaluators…but it's mostly social…," "…evaluation journal … see … what the hot topics are," "I continue to read. I submit papers," "I … review and integration of empirical research on the topic of interest"
Changing behaviors	Evaluators sought and incorporated feedback from colleagues to change their professional behaviors	"I talk about evaluation with my peers," "… having people I can ask if I have questions about my practice"
Collegiality	Evaluators practiced collegiality with peers through learning from each other, referring clients to someone if they are not an expert in an evaluation method, and volunteering in their professional associations	"I refer clients to someone who is expert in that particular area if I don't know…", "I find evaluation colleagues who are experts …," "I actively … participate in the activities of the profession…"

Sub-Scale 2: Innovation and Research. The intent of the qualitative questions in the Research and Innovation theme was to explore language used by evaluators to describe how they were institutionalizing intellectual knowledge to advance the discipline. Innovation and research refer to modifying or developing new knowledge, skills, ideas, methods, and new practical solutions to advance the discipline. Table 4.2 summarizes the codes, descriptors, and sample data within the theme of innovation and research. Evaluators described a range of activities and processes that they engaged in to advance the field. Participants reported purposefully conducting *research on evaluation* to generate and disseminate new knowledge in evaluation.

Although fewer participants reported conducting research on evaluation for the purpose of generating new theoretical knowledge, many associated "research on evaluation" with *modifying methods* created by other people or *creating hybrid approaches* for use in their evaluation practice. A participant noted, "There is nothing new under the sun. What you are seeing now as innovation in existing approaches is that we are creating hybrids of approaches." Moreover, several evaluators said they *adapted knowledge from other disciplines* to their evaluation work. Participants noted that evaluation is a trans-discipline. As such, practitioners coming from various backgrounds adapted methods from their disciplines to their evaluation practice. Several evaluators reported using technologies to innovate new ways of reporting evaluation findings.

Whether methodological or theoretical innovations, evaluators reported disseminating new knowledge by *sharing their insights* and *sharing reflections on their own practices* mainly in professional association conferences and workshops to advance the field. A few participants said they shared new knowledge through journal publications, although others contended that journal articles are not suitable for transmitted knowledge. A participant noted, "I don't want to write in a journal because I hate that style of writing."

Several evaluators noted that conducting research on evaluation to advance the field is a part of the job requirement for academic evaluators and therefore argued that practitioners are less suited for research. Academic evaluators said continuing to do research and publishing helps them to learn new things and advance the field.

Table 4.2
Program Evaluators' Descriptions of Innovation and Research.

Code	Description	Excerpts from Interview Data
Research on evaluation	Evaluators discussed how they intentionally generate new evaluation knowledge and participate in formal research on evaluation to benefit the field	"If you are an evaluator within a university context…research is part of your job", "…knowledge creation and dissemination…", "…continuing to do research and writing"
Cross-discipline adaptation	Evaluators applied knowledge from other disciplines to their evaluation work	"…we are a transdiscipline, … we cut across all the other disciplines…"
Sharing reflections and insights	Evaluators shared reflections on their own practices and insights to advance the profession	"I advance the profession through reflecting on my own practice", "I share my insights in conferences and workshops."
Modify methods	Evaluators modified methods created by other people for use in their own practice	"I have modified methods other people have used…", "we are creating hybrids of approaches"
Apply established methods	Evaluators applied established methods to novel situations	"I haven't developed anything new, but I adopted certain approaches"
Innovative reporting	Evaluators discussed innovative ways of preparing evaluation reports	"I do a lot more innovative reporting like newsletters, a screencast, and data parties"

Sub-Scale 3: Professional Autonomy. Scholarship from the sociology of professions refers to professional autonomy as the freedom professionals have from clients and the employing organization in order to perform their duties. In theory, evaluators must be free in framing the evaluation without *undue pressure or interferences* from the commissioners or funders of the evaluation. Participants emphasized the importance of professional autonomy in giving technical oversight about appropriate evaluation questions, methods, reporting, and ensuring that program stakeholders without a voice are involved in the evaluation process.

Besides giving technical oversight, professional autonomy related to an evaluator's personal agency to be able to *turn down jobs* or *to end their participation* when negotiation with the client or employer was unfruitful. Several evaluators noted that Autonomy came with confidence and level of mastery of evaluation approaches and techniques. Thus, personal agency tended to be exercised by the more experienced evaluators compared to novices and was also affected by the evaluator's place of work. For example, some internal evaluators with years of experience and mastery of evaluation techniques reported being constrained by the politics of their workplace and power differential. For example, one internal evaluator stated that,

> I had some professional autonomy, but I also had a boss, and I was within a structure of an organization, and so, therefore… that sometimes dictated what I did or did not do. (Interview respondent, July 27, 2017)

Participants noted the power dynamics in program evaluation and the notion that the person who commissions or pays for the evaluation owns the contract and ultimately decides on what they will accept as the product. This is reflected in the key evaluation questions selected, the methods used, and even in the ways the processes and results are used for decision making. For example, an external evaluation participant observed this power dynamic in requests for proposals (RFP), noting, "Oftentimes the requests for proposals will already specify the requirements for the evaluation, and we are supposed to fill in the details as prescribed."

However, several experienced academic evaluators reported that they do not accept the client's requests for proposals at face value. As a practice, they often re-arrange evaluation questions in the RFP and *can decide on the methods* to employ and then add questions, especially on unintended consequences or what has not worked well if not included in the RFPs. Experienced evaluators also said they *can turn down* evaluation jobs when they disagree with the client's or commissioner's intended approach and motive or if there is too much inherent bias within the program, and they *can choose what evaluation projects* to take on.

A key point made by many participants is that professional autonomy is a matter of balance and compromise. Participants noted that they collaborate and negotiate with other people on what makes sense. For example, an internal evaluator said they very diplomatically assert autonomy because of the layers of oversight on their work. Internal evaluators working in government noted that unlike external evaluators who depend on their clients for their next contracts or references from their clients, their employer cannot fire them although they can disagree with their professional judgment.

External evaluators, on the other hand, reported that sometimes they compromise their autonomy because of economic ramifications, but they *can refuse to do something* if the client's demand is unethical. An external evaluator noted, "I'm fine going their direction as long as there's nothing unethical." I will discuss ethics in the next section, but evaluators reported a client's demand for participants' confidential information and exclusion of negative findings from reports as examples of unethical behavior. Table 4.3 is a summary of codes and sample data for the professional autonomy theme.

Table 4.3
Program Evaluators' Descriptions of Professional Autonomy

Code	Description	Excerpts from Interview Data
Refuse to do something	Evaluators can refuse to do something they don't believe in if a client pressures them to do it	"I was not pressured to do things that I didn't think I should," "…didn't seem to make sense from an evaluative perspective"
Disagree with the client's approach	Evaluators described turning down evaluation work when they did not agree with a client's intended approach or motives	"…if there is too much inherent bias…," "…turned down a component of a project," "I have turned down contracts…," "I turned down projects…," "… I didn't agree with their approach"
Choose projects	Evaluators can decide what evaluation projects to take on	"I choose what projects I take on," "I … decide on what to take on, when to do an evaluation and what sort of evaluation to do"
Choose methods	Evaluators can decide what sort of evaluation methods to employ in an evaluation	"I have…the choice of methods," "…the best approach…," "…choose evaluation techniques…," "…I tell my clients well this is what is best…"
Turn down jobs	Evaluators can turn down jobs where they think that the client is unwilling to hear whatever might be found	"…if they are not willing to hear then I move on…," "…the client only wanted an evaluation to prove their point"
End participation	Evaluators describe that they can end their participation in a project if the evaluation is not going well	"…I can end it if I don't think it is going well," "I didn't think was worth my effort…"

Sub-Scale 4: Ethical Dispositions. Professional associations like the AEA and CES have developed general ethical principles to guide the activities of their members who engage in program evaluation work. The professionalism instrument will assess the following operational attributes of virtue, ethics, practical wisdom, behavior, and evaluator identity. All the participants reported that their practice of evaluation is guided by the principles of *integrity, competence, respect for people*, *common good and equity*, and *accountability*. Nearly all participants emphasized the importance of being *competent and accountable* for their performance. Participants reported taking responsibility in the *provision of information to clients to facilitate their decision making* and bringing their *best professional judgment* within the standard of evaluation practice.

Aside from being a reliable and trustworthy evaluator who fulfills their responsibilities to clients and employers to the best of their abilities, evaluators noted that practicing evaluation requires *integrity*. Participants explained they are honest, truthful, and transparent about their ethical values and that they *disclose conflicts of interests* to clients and commissioners when they may influence an evaluation. Several participants also noted that sometimes evaluation clients or commissioners demand confidential information about participants or sources of evidence. Evaluators perceived such requests as unethical and a breach of ethics or betrayal of the participants' trust in them. Nevertheless, participants expressed concern about the lack of clout in the field to protect evaluators from such demands. For example, one evaluator noted,

> …[W]e don't have enough professional clout to be able to back up any kinds of demands by individual evaluation practitioners that the professions support them in some sort of struggle with a particular client. They [guiding principles] are really, they sound good on paper, but what impact do they really have on evaluation practitioners? If you don't adhere to them, I mean what sanctions are going to be taken against you as a practitioner? If you do adhere to them, who is

going to defend you in a situation where you have a struggle with a client? (Interview participant, August 11, 2017)

Nearly all participants said they were concerned about *not doing harm* to those who participate in an evaluation, and several said they are concerned about *doing the right thing* regardless of the consequences. As one participant noted, “I know that I run the risk of being fired, and that’s okay.” Yet another participant said, “There is too much work out there if a client won’t hire me again because of negative findings.” Several participants also explained that they used *the ethics and standards of practice* as safeguards from undue interferences from clients and commissioners of evaluation. As a participant noted, “…the client was trying to push me in a certain direction, and I was able to say, well, the standards don’t allow me to, and you probably don’t want me to, either.” Nevertheless, some evaluators were concerned about losing credibility and income, as expressed by this participant: “I am at risk of either losing my job or losing credibility within the organization or losing influence if I refuse to give the names of the informants if demanded.”

Evaluators’ concern for the protection of participants was connected to the principle of *respect for people*. Participants reported minimizing respondents’ discomfort in participating in evaluation and respecting their autonomy to freely provide evaluation data. They explained that entails providing respondents with information on how participating in evaluation could be beneficial to them. Evaluators saw their work as being of service to others by doing good. Finally, the ultimate ethics of evaluation seemed to align with the principle of *common good and equity*. As noted by one participant, “Involving stakeholders who don’t have an established voice or people with less power and clout in the evaluation process is my most important ethical dilemma.”

Several participants said they ensured the voices of those at stake were included in an evaluation by *seeking the broader context in framing an evaluation*. They observed that sometimes this means having the courage to stand up to a client or employer or being incorruptible no matter the temptation to be dishonest. For example, a participant explained, “Sometimes you just have to put the client on the spot, and that cost me some fees for sure, but my name remained intact.” Table 4.4 summarizes the key ethical principles commonly mentioned by participants.

Table 4.4
Program Evaluators’ Descriptions of Ethical Dispositions

Code	**Description**	**Excerpts from Interview Data**
Competence	Evaluators emphasized providing stakeholders with the best possible information to the best of their abilities.	“…bring the best professional judgment to the work that we do…,” “I conduct evaluations to the best of my ability within the standards of the science”
Integrity – truthfulness and transparency	Participants described integrity as being transparent and truthful about their ethical values when conducting evaluations	“…the people I am partnering with want to know that we have the same set of values and are working towards the same kind of change in the world”
Honesty	Evaluators said honesty requires telling the whole truth like being upfront about their own values when people hire them	“Not everyone agrees with me. So, what I can do ethically is to say, this is who I am. This is how I think,” “to present myself for what my worth, not my self-view…”
Disclosure	Evaluators described disclosure as information that should be given to clients or commissioners, like disclosing conflicts of interests	“…if I am sort of liberal-leaning and social justice-oriented, people should know that’s what they are getting in me”

Table 4.4 (Continued)

Code	Description	Excerpts from Interview Data
Respect for people	Evaluators said they respected the autonomy of program participants to freely provide evaluation data without force or manipulation	"…the actual stakeholders who are experiencing the program … were put in a position of, it was coercive, if you want continued funding, you will participate in this study," "…respecting participants rights and confidentiality."
Doing good	Evaluators said they do good by providing respondents with information on how participating in evaluation would be beneficial to them	"…always respect the interest of your stakeholders"
Respect for people	Evaluators said that they care about how evaluation decisions affect program participants	"…how are we … not supporting a continuation of inequities," "involving stakeholders who don't have an established voice or people with less power and clout in the evaluation process"
Confidentiality	Evaluators said they act in the best interest of those who participate in an evaluation or doing to others as they would want to be done to themselves	"My biggest challenge is protecting personal information and maintaining the confidentiality of individual responses," "do no harm to others either by participating in the data collection or by the misuse of what is discovered"
Caring for others	Evaluators said they minimize respondents' discomfort in participating in an evaluation	"Reduce unnecessary risk or harm"
Safeguard or self-protection	Evaluators said they safeguard themselves by telling their clients or employers that they use program evaluation standards and that they follow a professional code of ethics to guide their work	"I am explicit about the evaluation standards and I tell my clients that I follow a code of ethics to guide my work," "I fall back on the program evaluation standards..."
Common good and equity	Evaluators said they seek the broader context in framing an evaluation to include concerns of all participants, especially those without voice and clout	The process has to be known, people have to be comfortable with the process and they have to never be surprised"

Sub-Scale 5: Credentialing. Altschuld and Engle (2015) define credentialing as the specification that "a person has requisite knowledge, skills, and practical experiences to be deemed worthy of a credential or designation of such status" (p. 6), such as by the review of an applicant's portfolio by other professionals in the same field. Participants described their beliefs in credentialing in five broad themes, including competence, professional development, conformity to standards, aligning credentialing to evaluator competencies, and protection of clients.

First, participants reported that credentialing had encouraged them to *reflect on* or to *self-assess* their skillsets with the goal of becoming a competent evaluator. They believed that credentialing would tell people that they have *reached the minimum standards* to be a *competent evaluator*. Although participants with doctoral degrees had mixed feelings about credentialing, others said credentialing would show their clients and employers that they had gone through their professional association's criteria to practice evaluation. Those who differed felt their competencies were greater than the two years of experience required by the Canadian PDP. Some interview participants stated that

> Like I know, I am like, I have a PhD, I have this you know all of these things, but at the same time I am like I think that we still really need the credentialing, because I might feel that way, but I wanna make sure that my client knows that too (Interview respondent, July 27, 2017).
>
> Because I have Ph.D. level research methods background and many *qual* and *quan* courses behind me, my work speaks for me. Credentialing would not help anything; it does not guarantee a strong methods background (CES Survey participant, August 6, 2019).

Furthermore, evaluators believed that being awarded a "Credentialed evaluator" (CE) designation (from the CES) meant that their peers had *recognized their expertise* or affirmed them as competent evaluators. Indeed, evaluators who experienced the credentialing process believed that credentialing made them showcase their work through their portfolios. Several credentialed evaluators reported seeing themselves now as *professional evaluators* because of credentialing. Gauthier observed that,

> Many evaluators, especially among CEs, are now introducing themselves as evaluators rather than economists, sociologists, psychologists, or political scientists as they did a decade ago. To me, that is associated with the magnet that represents the designation and the ability to call oneself something like "credentialed evaluator." (B. Gauthier, personal communication, March 18, 2019)

Second, the process of self-reflection and investing in themselves had spurred *professional development*. Credentialed evaluators reported that they were required by their professional association to demonstrate 40 hours of professional development each year to maintain their CE designations. This process makes them keep up-to-date on their practice. One evaluator noted that credentialing has "…helped me with my employer to justify certain professional development time."

Third, evaluation managers reported that credentialing would assure them that those who engage in evaluation *conform to certain professional standards* to authenticate their practice. Indeed, credentialed evaluators reported that credentialing makes their employers or supervisors see them as credible practitioners. In fact, this led to the fourth point evaluators made, that by gaining credibility and assurance of reaching minimum standards, credentialing protected evaluation clients and employers from hiring a noncompetent evaluator or someone who does not have enough of an evaluation background to do a good job. Several evaluation managers believed that credentialing would guarantee a minimal skill level because a lot of people improvise themselves as evaluators. Other managers lamented that they realized too late that they hired incompetent evaluators after spending several thousands of dollars. Nevertheless, participants were concerned that the federal government's lack of support downplays the importance of credentialing. Without such support, voluntary credentialing lacks clout, as noted by one respondent,

> With voluntary credentialing, if you can get organizations like the World Bank and the federal government and state governments to give extra points for credentials, voluntary credentials, that's how it would gain some momentum. (Interview respondent, July 26, 2017)

Furthermore, not all evaluators believed that credentialing would guarantee acceptable performance and questioned whether being a credentialed evaluator makes any difference. For example, several American evaluators said they believed the market approach was a better guarantee of performance on condition that employers do due diligence before hiring an evaluator. As one participant noted, "People who hire an evaluator have to do due diligence on their own; they have to be sure that the people they hire are qualified rather than relying upon some piece of paper." Other critics of credentialing are concerned about the potential exclusion of some evaluators from the practice of evaluation. For example, a participant noted,

> The cons are potentially designing a system that excludes really wonderful talented potential evaluators who will not and should not be asked to go through a

> traditional pathway… and produce white female evaluator after white female evaluator. And if we make a system that just makes more of that, it's just wrong. It is immoral to accept that. So, however this is designed, I want it to trouble the current pipeline… [and] make the pathway possibilities broader so that the field can be more inclusive. (Interview respondent, July 26, 2017)

Despite the difference of opinion, the fifth area where almost all evaluators agreed was the idea of *aligning credentialing to evaluator competencies*. Nearly all evaluators believed that evaluator competencies should provide a framework for teaching evaluation and that competencies should be used as a basis for credentialing. Several participants said using competencies for teaching would normalize the field and help in defining program evaluation. For example, one academic evaluator who reported using competencies as part of the curriculum to teach said students are responding positively and are beginning to recognize that evaluation is a profession. As one student reportedly put it, "Oh! I didn't realize that evaluation was a profession."

Another academic evaluator also said competencies provided a threshold of what should be taught and that they created opportunities for designing more evaluation courses. A participant noted, "So, as an academic providing some of that training, I can now look to the competencies and say okay, this is what we all agreed are the basic competencies. I can teach more than that, which I do, but I know I can't go less than that, right?"

Credentialed evaluators reported that credentialing heightened their *awareness of evaluator competencies*. Several people said the competencies made them realize what they needed to be a good or a competent evaluator and that people are beginning to understand that there is more to evaluation than what they practiced. A participant observed, "I think that it has made people who want to call themselves evaluators recognize that there is more to it than just calling yourself an evaluator and maybe reading a couple of books." Table 4.5 summarizes evaluators' beliefs in credentialing.

Table 4.5
Beliefs of Program Evaluators in Credentialing

Code	**Description**	**Excerpts from Interview Data**
Self-reflection	Evaluators described that being credentialed would encourage them to reflect on their skills	"…credentialing forces you to look at yourself…," "…be up-to-date on everything…," "I invest in myself; I consider myself a professional evaluator…"
Competence	Evaluators said credentialing would tell people that they have reached the minimum standards to be a competent evaluator	"I see a lot of value in the credentialed evaluator…," "…we are interested only in having people with the CE designation apply as potential contractors"
Affirmation	Evaluators said being a credentialed evaluator would mean that their peers have recognized their expertise or being affirmed by their peers	"…that process is one where your peers essentially look at your files and decide…," "I love the idea of the portfolio to sort of show your work."
Conformity to standards	Evaluators who are managers said credentialing would assure them that those who engage in evaluation conform to certain professional standards	"…to give the profession a different profile…," "conform to certain professional standards which are or have been assessed," "…them knowing that I am a credentialed evaluator matters …it lends credibility"
Protect clients	Evaluators believed credentialing protect stakeholders from hiring someone who does not have enough of an evaluation background to do a good job	"…the credential would at least give me that guarantee of minimal skill level because a lot of people improvise themselves as evaluators …," "…after I have spent $300,000, I realized that they are not giving me what I need"

Table 4.5 (Continued)

Code	Description	Excerpts from Interview Data
Professional development	Evaluators said credentialing would direct their professional development to maintain their credentialing status or to address weak skills	"…ongoing professional development…," "…by the end of it, it was so clear to me where my skills were weak," "…we are doing forty hours of maintenance"
Alignment to competencies	Evaluators noted that evaluator competencies should be aligned to credentialing as the framework for the assessment of competence	"I think the profession would have to be willing to accept evidence of completion of those skills…," "…there is more to it than just calling yourself an evaluator"
Awareness of competencies	Evaluators expressed that credentialing has made them aware of evaluator competencies	"…people become aware of the competencies," "I think the competencies are a good place to start," "…helpful in terms of understanding what you need to be good at an evaluator"
Teaching evaluation	Evaluators suggested that competencies should provide a framework for teaching evaluation	"…we know what the basic requirements are, we know where the threshold is, we know what kind of training ought to be provided," "It is part of the curriculum," "…influence on the design of more evaluation studies…," "…evaluation is going to be much more normalized"

Integration: Instrument Development and Pilot-Testing

Creswell and Clark (2018) refer to the point of integration in mixing as the second step where "the researcher identifies the results on which the quantitative feature will be built" (p. 87). The integration phase in this study involved the use of the exploratory qualitative results to build the quantitative professionalism assessment tool for program evaluators (PAToPE). As described in Chapter 3, the instrument development process involved three steps: 1) deriving items from qualitative data to build the quantitative instrument based on the themes derived from the sociology of professions theory, 2) using participant language in the wording of items, and 3) testing the instrument through cognitive interviews. Table 4.6 shows the four qualitative themes from the conceptual framework and the items derived from participant language.

Table 4.6
Items Derived from Qualitative Findings

Qualitative Themes from Conceptual Framework and Codes Interviews	**Quantitative Instrument Scales and Items Derived from Language Used by Evaluators in Describing Professionalism**
Expertise (research and innovation)	
Research on evaluation	I purposefully generate new evaluation knowledge to benefit the field
Share reflections	I advance the profession by sharing reflections on my own practice
Conferences	I share my insights at professional conferences
Research on evaluation	I participate in formal research studies on evaluation
Submit manuscripts	I submit manuscripts to evaluation journals
Read journals	I read evaluation-related journals to keep current
Conferences	I attend evaluation conferences to keep up-to-date
Research on evaluation	I integrate empirical research on evaluation to keep up-to-date
Established methods	I apply established methods to novel situations
Informal meeting	I meet with other evaluators informally to keep up-to-date
Own reading	I improve my skills in evaluation through reading on my own
Professional Autonomy	
Turn down work	I turn down evaluation work when I don't agree with my client's intended approach
Choose projects	I can decide what evaluation projects to take on.
End participation	I can end my participation in a project if the evaluation is not going well
Choose methods	I can decide what sort of evaluation method to employ.

Table 4.6 (Continued)

Qualitative Themes from Conceptual Framework and Codes Interviews	Quantitative Instrument Scales and Items Derived from Language Used by Evaluators in Describing Professionalism
Not coerced	I can refuse to do something I don't believe in if a client tries to get me to do it
Collegiality	I refer clients to someone who is an expert in an evaluation approach if I am not an expert in it
Disclose values	I am upfront about my own values when people hire me
Code of ethics	I tell my clients that I follow a professional code of ethics to guide my work
Turn down	I turn down jobs where I think that the client is unwilling to hear whatever might be found
Ethical Dispositions	
Ethical values	I am transparent regarding my ethical values when conducting evaluations
Provide information	I provide respondents with information on how participating in evaluation could be beneficial to them
Minimize discomfort	
Respect for people	I respect the autonomy of program participants to freely provide evaluation data
Use standards	I use the program evaluation standards as a safeguard in program evaluation
Best information	I provide stakeholders with the best possible information to improve their programs
Care about stakeholders	I am concerned about how evaluation decisions affect program participants
Context	I seek the broader context in framing an evaluation.
Disclose interests	I disclose my personal interests when they may influence an evaluation
Do-no-harm	I am concerned about not doing harm to those who participate in an evaluation
Credentialing	
Minimum standards	Credentialing would tell people I have reached the minimum standards to be a competent evaluator.
Reflect on skills	Being credentialed would encourage me to reflect on my skills
Conform to standards	Credentialing assures me that those who engage in evaluation conform to certain professional standards

Table 4.6 (Continued)

Qualitative Themes from Conceptual Framework and Codes Interviews	**Quantitative Instrument Scales and Items Derived from Language Used by Evaluators in Describing Professionalism**
Professional development	Credentialing would direct my professional development
Peer recognition	Being a credentialed evaluator would mean that my peers have recognized my expertise
Protects clients	Credentialing protects stakeholders from hiring someone who does not have enough of an evaluation background to do a good job
Align competencies to credentialing	Evaluator competencies should be aligned to credentialing
Awareness of competencies	Credentialing has made me aware of evaluator competencies
Use competencies to teach	Evaluator competencies should provide a framework for teaching evaluation

Chapter Summary

Chapter 4 provided the qualitative findings and operational definitions of professionalism themes derived from the sociology of professions theory. These definitions were used to develop items to measure professionalism of evaluators in North America. In the next chapter, I shall discuss the psychometric and statistical analyses of the quantitative data derived from PAToPE to answer my research questions two and three. I shall also discuss the results of a path analysis to explore the relationships between and among the five constructs.

CHAPTER 5

Quantitative Analyses and Findings

Introduction

This chapter presents the results of the quantitative analyses based on the survey to members of AEA and CES. Here, I present the results of factor analyses and the results from the first validation of the 59-item instrument to show how program evaluators demonstrate professionalism and how this reflects their attitudes towards credentialing and professionalization of evaluation.

Research Question 2

How is professionalism in evaluators assessed? The goal of this second research question was to develop a grounded and reliable professionalism measurement instrument to assess professionalism in evaluators. First, I assessed the psychometric qualities of the 59 items in PAToPE instrument through exploratory and confirmatory factor analyses. Second, I used the final 42 items identified through confirmatory factor analyses to assess the extent and ways in which Canadian and US program evaluators perceived and demonstrated professionalism in conducting evaluations. I assessed both their behavioral and attitudinal attributes of professionalism. The following sections present the results of exploratory and confirmatory factor analyses and structural equation modeling to assess measurement invariance and path analysis.

Quantitative Psychometric Item Analysis. Psychometric item analysis is one of the steps in an exploratory sequential mixed methods design (Creswell & Plano, 2018). I used psychometric item reduction and multivariate analysis techniques to assess the five domains from the literature that described professionalism and the subset of items that reliably measured each of these domains or factors. Adams and Wieman (2011) argue that the aim of psychometric validation is to increase confidence in the findings and use of the survey instrument in further studies. Because the professionalism assessment tool for program evaluators (PAToPE) is early in its development, I conducted an exploratory factor analysis (EFA) to uncover the underlying structures and confirmatory factor analysis (CFA) to assess how well my hypothesized model fit the data.

Exploratory Factor Analysis. I used the Principal Axis Factoring (PAF) extraction method in the Statistical Package for Social Sciences (SPSS) version 24 (IBM, 2016) and Mplus version 8.3 to analyze the factorability of the 59 items in the sample. The data came from responses to survey questions on a four-point Likert scale to assess the professionalism of evaluators in North America. Of the 456 participants who responded to the survey, I randomly selected approximately 50% of respondents (n=232) to conduct EFA and retained the other half (n=224) for cross-validation with confirmatory factor analysis (Fokkema & Greiff, 2017).

Of the 232 participants in the EFA sample, 221 completed the survey. The descriptive statistics of the variables showed that missing data on most of the 59 items of interest ranged between 4.3% - 15.9%. Missing data ranged from a low of 4% for Expertise to a high of 16% for Credentialing. Little's MCAR test indicated that the data were missing at random, x^2 (1463)= 1425.39, p >.755. As such, I used the listwise deletion approach to handle missing data.

I used several criteria to assess the factor structure of my model. First, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .76, indicating that enough items were predicted by each factor (Leech, Barrett, & Morgan, 2011). Second, Bartlett's test of sphericity was significant (x^2[1711] = 2026.56, $P < .001$), meaning variables were correlated highly enough to provide a reasonable basis for factor analysis (Leech et al., (2011). Third, the diagonals of the anti-image correlation matrix were all over .5, supporting the inclusion of each item in the factor analysis. Finally, the communalities were all above .3; .312 was the lowest, further confirming the existence of an underlying construct relating to all 59 variables before rotation. The determinant is 0.000000287, indicating that multicollinearity was not a problem for these data (Field, 2015).

The assumption of independent sampling was met, but the normality assumption was not met as the scale for the instrument was ordinal; a four-point scale of 1 = strongly disagree and 4 = strongly agree. In fact, Newsom (2019) noted that "[w]hen variables are measured on an ordinal scale and there are relatively few categories, 2-4 categories, estimation methods specifically designed for categorical variables are recommended" (p. 1). Furthermore, Osborne and Costello (2005) recommend the use of oblique rotation such as oblimin and scree plots when data do not meet assumptions.

I used direct oblimin rotation (an oblique rotation) because the underlying factors were theoretically related (Osborne & Costello, 2005). The initial eigenvalues showed that the first factor explained 13.56% of the variance, the second factor 9.17% of the variance, the third 6.22%, the fourth 4.44%, the fifth factor 3.77%, and the sixth factor 3.48% of the variance. The seventh through the 18th factors had eigenvalues of just over one, i.e., each factor explaining under 2% of the variance. I examined one-, two-, three-, four-, five-, and six-factor solutions in Mplus using the default weighted least square mean and variance (WLSMV) adjusted to compare the chi-square difference of test and model fit indices. I compared models with minus one factor to see the most parsimonious solution based on the goodness of fit indices summarized in Table 5.1.

Mplus provides the Comparative Fit Index (CFI), Root Means Squared Error of Approximation (RMSEA), the Tucker-Lewis index (TLI), and Standardized Root Mean Squared Residual (SRMR) when using MLSMV. The CFI evaluates incremental fit assessing whether the tested model fits better than the null model that treats all items as completely unrelated to each other. Absolute fit—the degree to which the relationships between variables implied by the model are similar to the relationships actually found in the data—were measured by RMSEA and SRMR, with RMSEA including a penalty for greater model complexity (Kline, 2017). Simulation studies suggest acceptable model fit should have a CFI ≥ 0.95, RMSEA ≤ 0 .06, and SRMR ≤ 0.08 (Hu & Bentler, 1999) cutoff points, although Browne and Cudeck (1993) suggest that standards of CFI ≥ .90, RMSEA ≤ .10, and SRMR ≤ .10 are minimally adequate.

Table 5.1 shows that a five-factor model is the most parsimonious solution to demonstrate the goodness of fit when compared to the other models. The 6-factor model has better fit statistics, but it is not supported by my theoretical framework. My preference is the five-factor solution, which supports my theoretical model based on the sociology of professions theory.

Table 5.1
Exploratory Factor Analyses and Model Fit Statistics

Factors in solution	Number of Parameters	Model X^2 fit test	X^2 difference test compared to a model with one less factor	RMSEA	90% CI f RMSEA	CFI	TLI	SRMR
1	59	4229.88***	743.33***	.084	.081, .087	.582	.567	.173
2	117	2892.76***	387.61***	.061	.057, .064	.789	.774	.120
3	174	2394.42***	278.86***	.050	.046, .054	.861	.845	.096
4	230	2127.10***	233.05***	.045	.040, .049	.895	.879	.087
5	285	1923.96***	192.42***	.040	.035, .044	.919	.903	.078
6	339	1745.55***	-	.035	.030, .040	.939	.924	.069

Note: $*p < .05$, $**p < .01$, $***p < .001$

I used the scree plot, which is a graphical representation of the eigenvalues on the y-axis against all 59 variables used in the factor analysis on the x-axis. The scree plot displays the proportion of the total variation in the dataset explained by each factor, and it is a useful way of showing the number of factors to retain. The scree plot in Figure 5.1 shows a ‘leveling off’ of the eigenvalues after the sixth factor, indicating that the first five factors accounting for most of the total variability in the data should be retained. Looking at the natural bend or break point occurring before the curve flattens out, each successive factor after the sixth factor is accounting for less and less amounts of total variance.

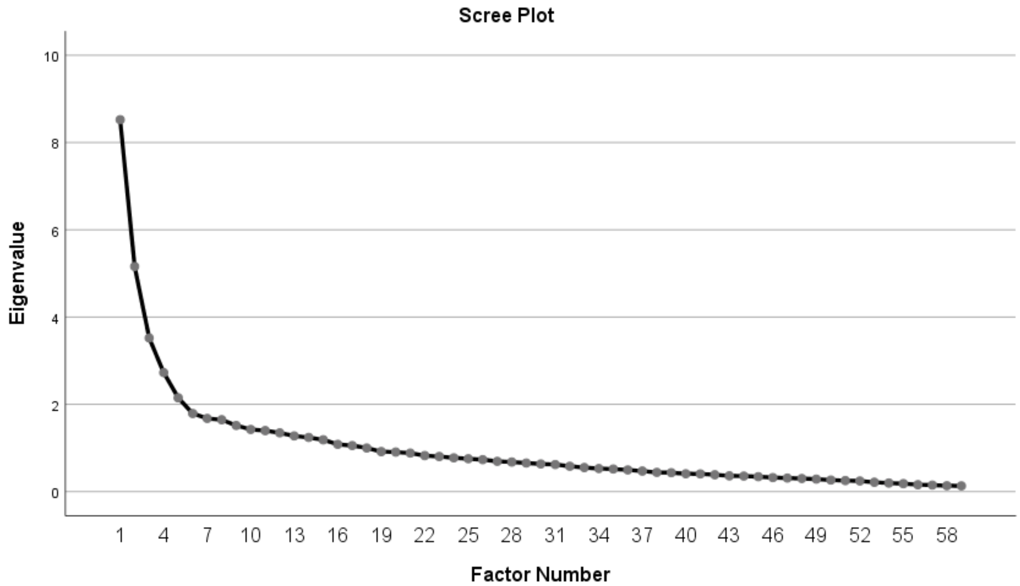

Figure 5.1. Scree Plot of PAToPE Instrument Results

Costello and Osborne (2005) argue that the best fit to the data is indicated by a table with the “cleanest” factor structure with item loadings above .30, few or no item cross-loadings, and no factors with fewer than three items. In fact, item developers like Benson and Clark (1982) suggest the retention of items whose discrimination value is .30 or greater. Table 5.2 provides a summary of the EFA in Mplus. The innovation and research factor had crossover items from expertise. Further observations indicate that these items index innovation and research according to the sociology of professions theory.

After rotation, the first factor accounted for 4.93% of the variance, the second factor accounted for 5.67%, the third factor accounted for 5.02%, the fourth factor accounted for 3.65%, and the fifth factor accounted for 2.47%. The five-factor model seemed to index five professionalism sub-scales from the sociology of professions theory: Expertise, Credentialing, Ethics, Innovation, and Professional Autonomy.

Table 5.2
Factor Loadings

Item#	Item	Factor Loadings: Factor 1: Expertise	Factor 2: Innovation & Research	Factor 3: Ethical Dispositions	Factor 4: Credentialing	Factor 5: Autonomy
Exp1	I self-assess to identify my professional strengths as an evaluator	0.502				
Exp2	I learn from evaluation thought leaders in areas of interest		0.405			
Exp4	I improve my skills in evaluation through reading on my own		0.564			
Exp7	I meet with other evaluators informally to keep up-to-date		0.438			
Exp8	I attend evaluation conferences to keep up-to-date		0.517			
Exp9	I attend professional development workshops to keep up-to-date	0.519				
Exp10	I integrate empirical research on evaluation to keep up-to-date		0.692			
Exp11	I read evaluation-related journals to keep current		0.626			
Exp12	I seek feedback on my work from evaluation colleagues	0.756				
Exp13	I incorporate feedback from my colleagues to make changes in my professional behavior	0.791				
Exp14	I have people I can ask if I have questions about my practice	0.643				
Exp15	I meet periodically to learn from other evaluators	0.455				
Exp16	I refer clients to someone who is an expert in an evaluation approach if I am not an expert in it	0.356				
Exp17	I submit manuscripts to evaluation journals		0.590			

Table 5.2 (Continued)

Item#	Item	Factor Loadings				
		Factor 1: Expertise	**Factor 2: Innovation & Research**	**Factor 3: Ethical Dispositions**	**Factor 4: Credentialing**	**Factor 5: Autonomy**
IR18	I purposefully generate new evaluation knowledge to benefit the field		0.874			
IR19	I apply knowledge from other disciplines to my evaluation work		0.406			
IR20	I advance the profession by sharing reflections on my own practice		0.663			
IR21	I modify methods created by other people for use in my own practice		0.421			
IR22	I share my insights at professional conferences		0.712			
IR23	I apply established methods to novel situations		0.517			
IR24	I participate in formal research studies on evaluation		0.605			
Et26	I provide stakeholders with the best possible information to improve their programs	0.547				
Et27	I disclose my personal interests when they may influence an evaluation			0.450		
Et28	I am upfront about my own values when people hire me			0.521		
Et29	I tell my clients that I follow a professional code of ethics to guide my work	0.511				
Et30	I respect the autonomy of program participants to freely provide evaluation data			0.636		
Et31	I am concerned about how evaluation decisions affect program staff			0.817		

Table 5.2 (Continued)

Item#	Item	Factor 1: Expertise	Factor 2: Innovation & Research	Factor 3: Ethical Dispositions	Factor 4: Credentialing	Factor 5: Autonomy
		Factor Loadings				
Et32	I am concerned about how evaluation decisions affect program participants			0.906		
Et33	I provide respondents with information on how participating in evaluation could be beneficial to them			0.515		
Et35	I am transparent regarding my ethical values when conducting evaluations			0.582		
Et36	I minimize respondents' discomfort in participating in evaluation			0.597		
Et37	I am concerned about not doing harm to those who participate in an evaluation			0.596		
Et38	I use the program evaluation standards as a safeguard in program evaluation	0.628				
Et39	I turn down jobs where I think that the client is unwilling to hear whatever might be found					0.588
Au40	I can refuse to do something I don't believe in if a client tries to get me to do it					0.554
Au41	I turn down evaluation work when I don't agree with my client's intended approach					0.836
Au43	I can decide what evaluation projects to take on					0.645
Au44	I can decide what sort of evaluation method to employ					0.508
Au47	I can end my participation in a project if the evaluation is not going well					0.535

Table 5.2 (Continued)

Item#	Item	Factor Loadings				
		Factor 1: Expertise	**Factor 2: Innovation & Research**	**Factor 3: Ethical Dispositions**	**Factor 4: Credentialing**	**Factor 5: Autonomy**
Au49	I seek the broader context in framing an evaluation	0.465				
Cre53	Credentialing would tell people I have reached the minimum standards to be a competent evaluator				0.878	
Cre51	Being credentialed would encourage me to reflect on my skills				0.868	
Cre58	Credentialing assures me that those who engage in evaluation conform to certain professional standards				0.881	
Cre52	Credentialing would direct my professional development				0.839	
Cre57	Being a credentialed evaluator would mean that my peers have recognized my expertise				0.808	
Cre56	Credentialing protects stakeholders from hiring someone who does not have enough of an evaluation background to do a good job				0.836	
Cre54	Evaluator competencies should be aligned to credentialing				0.809	
Cre55	Credentialing has made me aware of evaluator competencies				0.778	
Cre59	Evaluator competencies should provide a framework for teaching evaluation				0.587	

Note: Factor loadings below .4 were excluded except one theoretically important item.

The rotated factor matrix displays items clustered into five groups defined by the highest loading on each item. Overall, 11 items clustered on factor one, 14 items on factor two, 9 items on factor three, 9 items on factor four, five items on factor five, and 10 items were omitted since they did not meet the minimum criterion of factor loading of .4. Nevertheless, I retained one item with factor loading below .4 for theoretical reasons.

The first factor, which seems to index the Expertise sub-scale, had strong loadings above .70 for two items, .60 for two items, four items above .50, and the last three items loaded above .40 (.356 was the lowest loading). “I seek feedback on my work from evaluation colleagues” (.756) and “I incorporate feedback from my colleagues to make changes in my professional behavior” (.791) had the highest loadings from the first factor. One item that would fit this factor based on the sociology of professions theory loaded below .40: “I refer clients to someone who is an expert in an evaluation approach if I am not an expert in it.” I kept this item because collegiality with peers is a hallmark of a profession. Professionalism demands that professionals become more dependent on one another for referrals and access to jobs instead of competing with one another.

The second factor, which seemed to index the Innovation and Research sub-scale, had 11 items with factor loadings ranging between .4 and .8. Two items with the highest factor loadings are “I purposefully generate new evaluation knowledge to benefit the field” (.874) and “I share my insights at professional conferences” (.712). The third factor, which seemed to index ethical dispositions, loaded over .50 on eight items and .40 on one item. “I am concerned about how evaluation decisions affect program participants (.906) and “I am concerned about how evaluation decisions affect program staff” (.817)” had the highest loadings.

The fourth factor, which seemed to index the Credentialing sub-scale, had high loadings above .50 for all nine items in Table 5.2. “Credentialing would tell people I have reached the minimum standards to be a competent evaluator” (.878) and “Credentialing assures me that those who engage in evaluation conform to certain professional standards (.881) were the two items with the highest factor loadings.

The fifth factor, which seemed to index the Professional Autonomy sub-scale, loaded six items, with four items loading .50 and two items loading above .6. “I turn down evaluation work when I don't agree with my client's intended approach” (.836) and “I can decide what evaluation projects to take on” (.645) were the two items with the highest loadings.

I examined the distribution for each of the five sub-scales from the EFA model and the internal reliability using Cronbach’s alpha as shown in the descriptive statistics in Table 5.3. The overall EFA model had good internal reliability (Cronbach’s α = .88). The alphas were high for Credentialing (α = .96), Innovation and Research (α = .93) and Ethics (α = .93), Expertise (α = .89), and Professional Autonomy (α = .86). Authors on instrument development (Adams & Wieman, 2011; Benson & Clark, 1982) recommend a coefficient alpha of at least .80 or greater as desirable. Overall, these analyses indicated that five distinct factors are underlying evaluators’ responses to items in the professionalism instrument and that these factors are highly internally consistent.

Table 5.3
Descriptive Statistics for Five Factors of Professionalism (N = 383)

	No. of items	*M* (*SD*)	Skewness	Kurtosis	Alpha
Expertise	11	2.87 (0.86)	-1.896	3.920	.89
Innovation and Research	14	2.68 (0.83)	-1.725	3.585	.93
Ethics	9	3.04 (1.07)	-1.903	2.943	.93
Autonomy	6	2.27 (1.08)	-.762	-.273	.86
Credentialing	9	2.34 (1.18)	-.811	-.201	.96

Note: The rating scale used in the professionalism assessment tool was: strongly disagree (1), disagree (2), agree (3), and strongly disagree (4). The items are based on the EFA model.

I created composite sub-scale scores for the five factors based on the mean of the items that had their primary loadings on each factor. As shown in Table 5.3, evaluators reported higher ethical behaviors with a negatively skewed distribution, followed by expertise and innovation and research, both with a negatively skewed distribution, whilst belief in credentialing and professional autonomy behavior were moderate and had less skewed distributions. The skewness and kurtosis indicate non-normal distribution for Ethics, Expertise, and Innovation and Research.

Confirmatory Factor Analysis. I conducted a confirmatory factor analysis (CFA) using the lavaan package (v. 0.6-1) in R (Rosseel, 2012) and Mplus version 8.3 (Muthén & Muthén, 2017) to assess whether the measurement model identified in the exploratory factor analysis fitted the data. As earlier mentioned, I randomly selected approximately 50% of the respondents (n=224) for cross-validation with the confirmatory factor analysis. The CFA tested the measurement model for four scales of the instrument: *Professional Autonomy, Ethical Dispositions, Expertise, and Credentialing.*

I treated the 4-category Likert-type items on the survey as ordered categorical variables rather than continuous variables in the CFA estimation (Flora & Flake, 2017;

Flora, LaBrish, & Chalmers, 2012). As categorical variables, the association of the items and the underlying factor(s) was nonlinear. Consequently, I estimated all the CFA models with a diagonally weighted least squares estimator, which makes no assumptions about the distribution of the item responses and uses the polychoric, rather than product-moment correlation matrix (Li, 2016; Rhemtulla, Brosseau-Laird, & Savalei, 2012). However, I used the full weight matrix to compute robust standard errors and a mean- and variance-adjusted chi-square test statistic.

Assessing Scale Quality. To assess the model fit for the CFA, I used CFI, RMSEA, and SRMR. Lastly, besides evaluating the fit indices, I also used the size of the factor loadings to assess the quality of the model and, more specifically, to identify items that were not contributing to the factor structure. I used loadings of .4 as cut off point for both EFA and CFA.

While the model fit indices and factor loadings provide information about the dimensionality of items, I also assessed the reliability of responses in order to establish the appropriate interpretation and use and results from the scale. I used more rigorous assessments of reliability in the CFA. As such, I used Coefficient Alpha (Cronbach, 1951) and the three Coefficient Omegas, including hierarchical omega reliability (McDonald, 1999), all forms of internal consistency reliability to evaluate the amount of observed item variance that is true item variance as opposed to error variance. The four coefficients differ by how they quantify total variance and whether the number of items is taken into consideration (Alpha) or not (Omega). Although Alpha is the most commonly used coefficient, there are several well-documented problems with Alpha, which can often be overcome by using an alternative coefficient, such as Omega (Dunn, Baguley, & Brunsden, 2014; Trizano-Hermosilla & Alvarado, 2016).

Measurement of Model Fit. I tested both the model fit indices, and the results are summarized in Table 5.4. Overall, the model fit indices varied per Hu and Bentler's (1999) guidelines. A model fit was determined as "good" when all three fit indices meet the minimum threshold for fit, "marginal/ adequate" fit if any two of the three fit indices meet the minimum threshold for fit, and "poor," if all fit three indices failed to meet the minimum threshold of fit.

As shown in Table 5.4, the five-factor model—referred to as Model 1—showed a "poor" fit for data. Although the chi-square is significant and RMSEA is good, the CFI, TLI, and SRMR fell below Bentler's rigorous guidelines fit indices. The combination of small sample size and the quality of the four-point scale of *Agree/Disagree* may have affected the fit statistics. I conducted a post hoc modification to explore the model fit to the data.

Post Hoc Model Modification

The effects of model modifications and model fit statistics are summarized in Table 5.4. Kline (2017) recommends prespecifying a model if the fit is poor and if justifiable and supported by theory. I conducted a post hoc analysis in Mplus (Muthen & Muthen, 2019) version 8.3 since the hypothesized model poorly fitted the data. In fact, Byrne (2016) argues that a model can be modified if the fit of the originally hypothesized model is poor and if there are overlaps of items. I assessed the model modification indices (MI) to identify items for model modification, and I modified the model initialing starting with one item at a time to avoid distorting the data.

First, the Mplus output showed that item Exp15 *I meet periodically to learn from other evaluators* had a large model modification index (MI) value and the item overlapped with the Ethics and Innovation factors. As such, I deleted Exp15 from the model. As a result, Model 2 fit indices improved slightly as shown in Table 5.4. Second, Item Exp12 *I seek feedback on my work from evaluation colleagues* had a large MI value, indicating potential loading in the Ethics and Expertise factors. Therefore, I freed Exp12 from Expertise to Ethics factor, but the fit indices for Model 3 worsened as shown in the Table. I returned the Exp12 indicator to the Expertise factor.

Next, I deleted two items that overlapped with other factors and poor loading: Et31 *I am concerned about how evaluation decisions affect program staff* and Exp2 *I learn from evaluation thought leaders in areas of interes*t. Item Exp2 had poor loading on the Innovation and Research factor. I also freed Et26 *I provide stakeholders with the best possible information to improve their programs* from the Expertise factor to the Ethics factor due to poor loading. The fit indices for the model improved as can be seen in Model 4.

Fourth, I deleted Exp12 *I seek feedback on my work from evaluation colleagues* since it overlapped with Autonomy and Ethics factors. I also freed Et38 *I use the program evaluation standards as a safeguard in program evaluation* from Expertise to Autonomy. Furthermore, I freed Exp9 *I attend professional development workshops to keep up-to-date* from Innovation to Expertise. I freed Exp4 *I improve my skills in evaluation through reading on my own* from Innovation and Research to Expertise factor because of poor loading and freed Et38 *I use the program evaluation standards as a safeguard in program evaluation* from the Autonomy factor to the Expertise factor due to the high MI. The fit indices improved slightly as can be seen in Model 5. The fit indices for Model 6 increased significantly.

Fifth, I freed three indicators from the Innovation and Research factor to the Expertise factor since they seemed theoretically related to Expertise: Exp7 *I meet with other evaluators informally to keep up-to-date*; Exp8 *I attend evaluation conferences to keep up-to-date*, and Exp11 *I read evaluation-related journals to keep current*. However, the fit indices improved slightly as shown in Model 6.

Finally, I returned one indicator to the Innovation factor (Exp11). I also deleted Exp4 *I improve my skills in evaluation through reading on my own* and Au49 *I seek the broader context in framing an evaluation* due to poor loading. The model fit indices improved significantly as can be seen in Model 7. Hence, I selected Model 7 since it has moderate fit indices. As such, I used Model 7 with 42 indicators and five factors to conduct the statistical analyses to respond to research question number three.

Table 5.4
Exploratory Goodness-of-Fit Statistics and Model Modification

Five-Factor Model	Model X^2 fit test	RMSEA <.06 <.10	90% CI RMSEA	CFI .90≥ .95	TLI	SRMR ≤.08≤ .10	Model Fit
1	1775.956***	.053	.048, .057	.869	.862	.108	Poor
2	1633.184***	.050	.045, .055	.887	.880	.103	Poor
3	1699.083 ***	.053	.048, .057	.873	.866	.105	Poor
4	1456.177 ***	.048	.043, .053	.900	.894	.100	Poor
5	1373.500 ***	.047	.041, .052	.913	.908	.097	Marginal
6	1265.655***	.044	.039, .050	.919	.914	.095	Marginal
7	1155.689***	.046	.040, .052	.920	.915	.095	Moderate

Note: *p < .05, **p < .01, ***p < .001. CFI = comparative fit index, RMSEA = root mean-square of approximation, SRMR = standardized root mean square, n = the number of respondents used in assessing the model.

Overall, the global model fit was RMSEA = .046; CFI = 0.920; TLI = 0.915; SRMR = .095; X^2 = 5436.29, p < .001, and degree of freedom = 861. The items in the EFA model decreased by seven indicators from 49 to 42 in the CFA model. The final model showed that the indicators for the Expertise factor decreased from 11 to eight. The indicators for the Innovation and Research factor decreased from 14 to 10. However, the indicators for three factors remained the same as those in the EFA; the Ethics indicators remained at nine, the Autonomy indicators remained at six, and the Credentialing indicators remained the same at nine. Overall, I deleted six items and moved one indicator from the Expertise factor to Ethics factor (Et26) and one indicator from the Innovation and Research to Expertise factor. I retained two indicators whose loadings were slightly above the cutoff point of .4 since they were theoretically important to the model: IR24 *I participate in formal research studies on evaluation* and Au47 *I can end my participation in a project if the evaluation is not going well*.

Figure 5.2 illustrates the five-factor measurement model of professionalism produced using Mplus 8.3 (Muthén & Muthén, 2017). The figure shows the correlation between latent variables and the factor loadings. All the factor loadings in the model are significant and are shown in Table 5.5. The table shows standardized factor loadings for items within the final five scales.

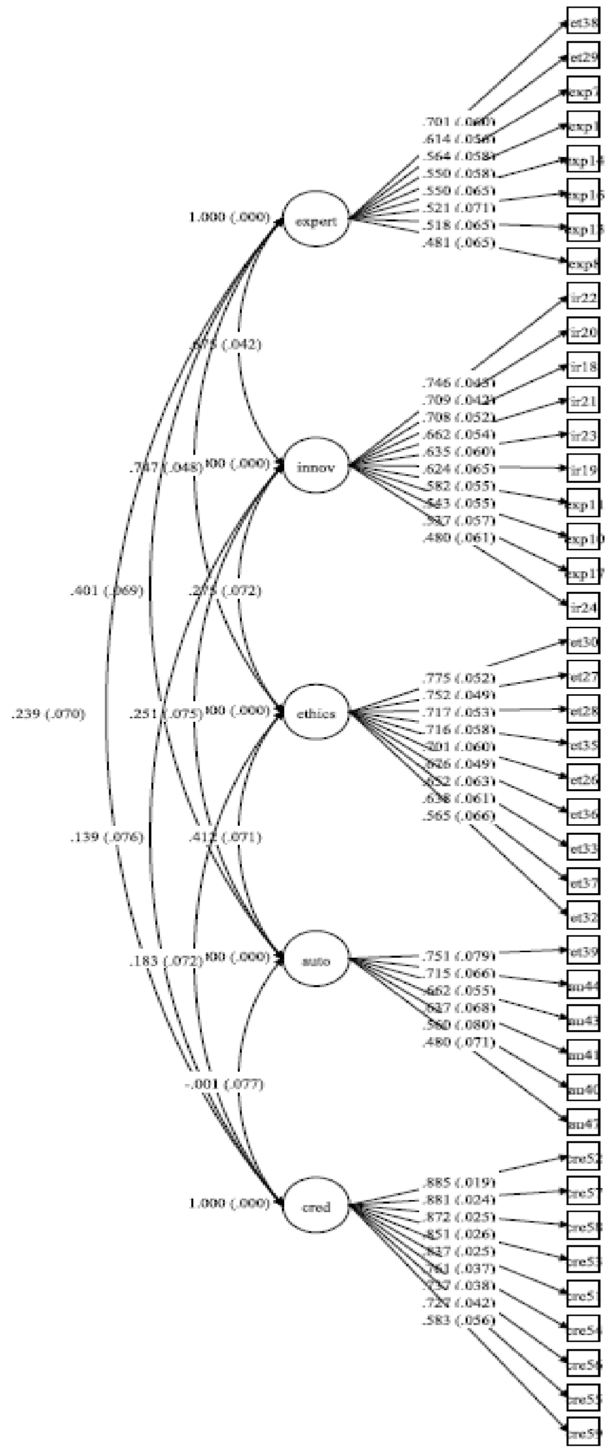

Figure 5.2. The Measurement Model of the Five Scales Professionalism Instrument
Note: RMSEA = .045; CFI = 0.924; TLI = 0.919; SRMR = .094; X^2 = 5070.208, and degree of freedom = 741. Standardized factor loadings are presented in Table 5.5.

Table 5.5
Standardized Factor Loadings

Items	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Expertise				
ET38	0.701	0.060	11.701	0.000
ET29	0.614	0.056	11.007	0.000
EXP7	0.564	0.058	9.672	0.000
EXP1	0.550	0.058	9.432	0.000
EXP14	0.550	0.065	8.463	0.000
EXP16	0.521	0.071	7.327	0.000
EXP13	0.518	0.065	7.975	0.000
EXP8	0.481	0.065	7.398	0.000
Innovation				
IR22	0.746	0.043	17.173	0.000
IR20	0.709	0.042	16.965	0.000
IR18	0.708	0.052	13.59	0.000
IR21	0.662	0.054	12.243	0.000
IR23	0.635	0.060	10.638	0.000
IR19	0.624	0.065	9.667	0.000
EXP11	0.582	0.055	10.497	0.000
EXP10	0.543	0.055	9.892	0.000
EXP17	0.537	0.057	9.435	0.000
IR24	0.480	0.061	7.857	0.000
Ethics				
ET30	0.775	0.052	15.042	0.000
ET27	0.752	0.049	15.33	0.000
ET28	0.717	0.053	13.585	0.000
ET35	0.716	0.058	12.258	0.000
ET26	0.701	0.06	11.773	0.000
ET36	0.676	0.049	13.798	0.000
ET33	0.652	0.063	10.362	0.000
ET37	0.638	0.061	10.464	0.000
ET32	0.565	0.066	8.514	0.000

Table 5.5 (Continued)

Items	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Autonomy				
ET39	0.751	0.079	9.537	0.000
AU44	0.715	0.066	10.837	0.000
AU43	0.662	0.055	12.017	0.000
AU41	0.637	0.068	9.389	0.000
AU40	0.560	0.08	6.995	0.000
AU47	0.480	0.071	6.721	0.000
Credentialing				
CRE52	0.885	0.019	46.775	0.000
CRE57	0.881	0.024	37.272	0.000
CRE58	0.872	0.025	35.029	0.000
CRE53	0.851	0.026	33.368	0.000
CRE51	0.837	0.025	33.853	0.000
CRE54	0.761	0.037	20.752	0.000
CRE56	0.737	0.038	19.378	0.000
CRE55	0.727	0.042	17.515	0.000
CRE59	0.583	0.056	10.328	0.000

Note: S.E. = Standardized errors of the parameter estimates, and Est./S.E. = Parameter estimate divided by standardized error of the parameter estimate.

Reliability. I calculated reliabilities for the final model using the semTools package[9] (Pornprasertmanit & Rosseel, 2016) in R. Table 5.6 shows that the reliability was acceptable for all the scales indicating in a classical test theory sense that a large proportion of variation in the observed raw summed scores on the survey was true variation in the summed scores as opposed to measurement error. Nunnally and Bernstein (1994) argued that composite reliability values of 0.6 to 0.70 are acceptable in exploratory research. As such, all the five sub-scales had adequate values. For all factors, the reliability suggests that a raw summed score from instrument responses could be

[9] Further description of the coefficients and the formulas for calculating them are found in the semTools documentation (https://cran.r-project.org/web/packages/semTools/semTools.pdf)

measured with precision in the current sample, thus the summed score would be a strong measure of the underlying construct. As such, I used the scores of the sub-scales to assess professionalism of evaluators.

Table 5.6
Reliability Coefficients of the Five-Factor Model

	Expertise	Innovation	Ethics	Autonomy	Credentialing
Alpha	0.670	0.794	0.831	0.722	0.902
Omega	0.686	0.794	0.833	0.725	0.906
Omega2	0.686	0.794	0.833	0.725	0.906
Omega3	0.704	0.778	0.834	0.721	0.905

Note: Omega is also referred to as conditional omega (Bollen, 1980; see also Raykov, 2001), Omega2 as unconditional omega (Bentler, 1972, 2009), and Omega3 referred to as hierarchical omega (McDonald, 1999).

Tests for Measurement Invariance – Multigroup Analysis. Multigroup CFA is a method of establishing measurement invariance, which means that the survey is measuring two or more groups in the same manner (Hirschfield & von Brachel, 2014). When measurement invariance holds, we know that differences in responses to the survey are true differences between the groups rather than the result of measurement error. Three of the four types of invariance classifications by Hirschfield and von Brachel (2014) are relevant for this purpose: 1) configural - is the factor structure the same for all groups? 2) metric (weak) - are the factor loadings equal across groups? and 3) scalar (strong) - are the item thresholds equal across groups? I am assuming measurement invariance although I was unable to formally assess it due to very small sample sizes. The number of parameters in this model is 264, but a rule of thumb of 10 - 20 participants per parameter (Kline, 2017) implies having a sample size of at least 2,640 to 5,280 – which was way beyond the current sample.

Path Analysis

I performed structural equation modeling in R using lavaan (v. 0.6-3) packages to test the relationships between the professionalism sub-scales hypothesized in my model (see the conceptual model in Figure 2.7). The model fit indices show $\chi^2(861, N = 212) = 5436.29, p <$.001, CFI = .920; RMSEA = .046, 90% RMSEA Confidence Interval (CI) [.040, .052], and SRMR = .095. Although the high X^2 and low p-value suggest a model misfit, the other fit indices suggest good model fit. As shown in Figure 5.3, this model hypothesized that Credentialing has a relationship with Expertise. This suggests that people with proof of tested competence are more likely to report increased Expertise. The cause of Credentialing is unexplained in the model, but this might relate to the agency of the professional associations to professionalize the field in response to contextual factors. The model shows an inverse and nonsignificant relationship with Innovation and Research, suggesting that Innovation and Research does not depend on the unexplained causes of Credentialing.

The hypothesized model also shows that there is a strong relationship between Expertise and Innovation and Research. This suggests that evaluators reporting high levels of Expertise are more likely to report engaging in Innovation and Research to promote the field. Expertise and Innovation have unrelated causes outside the model. For example, prolonged education or specialized training or many years of experience might be related to Innovation and Research. Access to funds for research could also contribute to innovation. In the qualitative analyses, academic evaluators with access to funds for research reported engaging in Innovation and Research. They also reported higher levels of Expertise on average compared to non-academic evaluators.

Furthermore, the model suggests that Expertise is related to increased Ethics and increased Autonomy, and these relationships are statistically significant. Thus, the model hypothesized an indirect relationship between Credentialing and Autonomy through Expertise and an indirect relationship between Credentialing and Ethics through Expertise. These indirect relationships suggest that credentialed evaluators (CEs) are more likely than non-credentialed evaluators (CEs) to report high levels of Expertise and are more likely to report behaving more ethically. These relationships will be explored to see how CEs and non-CEs compare on their reported levels of Expertise and Ethics in the next section.

The model also shows moderate covariation between Autonomy and Ethics after controlling for the other relationships in the model. Ethics and Autonomy have shared unexplained causes outside the model. For example, the qualitative study shows that the agency of professional associations affects how evaluators apply the ethical guidelines to their practice and exercise autonomy in making evaluative decisions. Some evaluators observed that unless professional associations have clout to enforce ethical guidelines and evaluation standards, including backing up evaluators when they are in some sort of struggle with a client on the integrity of evaluation, evaluators are on their own in dealing with clients. As such, evaluators cannot be sanctioned if they do not adhere to the guiding principles nor will the association protect them if a client breeches the guiding principles and evaluation standards. The lack of clout by professional associations seems to affect evaluator’s reported ethical behavior and Professional Autonomy when conducting an evaluation. It appears the Professional Autonomy and ethical behavior of evaluators is related to the professional clout of associations and these are external to the model.

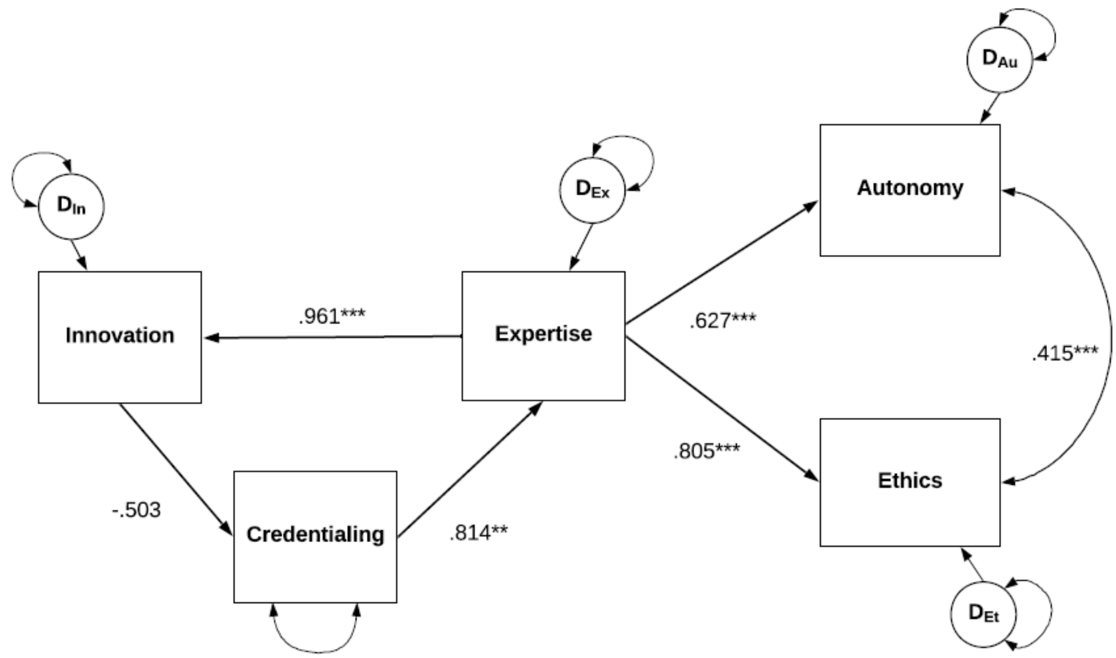

Figure 5.3. Path Model of Professionalism with Standardized Coefficients

Note: *p < .05, **p < .01, ***p < .001. The disturbances or unexplained causes of the endogenous variables are labeled D_{In} for Innovation, D_{Ex} for Expertise, D_{Au} for Autonomy, and D_{Et} for Ethics. Credentialing is the only exogenous variable in the model.

The hypothesized model showed a small negative relationship between residual covariance or unexplained causes of Professional Autonomy and Innovation and Research, but a small negative relationship between the residual covariance of Ethics and Innovation and Research.

Descriptive Statistical Results of the Items and Scales

Research Question 3

To what extent and in what ways do program evaluators in Canada and the US demonstrate professionalism in their practice? The goal of this third research question was to assess whether evaluators in North America reported demonstrating professionalism in their practice of evaluation. While in the preceding sub-section I conducted psychometric analyses of the data to uncover the latent variables of professionalism and quality of the new instrument, in this section I analyzed the survey data to understand whether there were differences in professionalism by country, evaluator workplace, gender, credentialed evaluator (CE) status, skill level, education, age, employment status, and level of experience. I used one-way Multivariate Analysis of Variance (MANOVA) to assess differences in professionalism behaviors and attitudes based on factor scores.

First, I started analyses by comparing the numbers and percentages of evaluators who rated the items “agree” and “strongly agree” (collapsed to “agree”), which is a 3 and 4 on the ordinal scale of the survey, and the factor loading of these items for each of the four domains. Next, I compared the means and standard deviations for Expertise, Innovation and Research, Ethics, Autonomy, and Credentialing by country to understand perceived evaluator professionalism. Finally, I used factor scores to compare evaluator professionalism by workplace, gender, CE, skill level, education, age, employment status, and level of experience.

Expertise. Table 5.7 shows the proportion of evaluators who reported various Expertise behaviors on this scale. The descriptive statistics in the Table indicate that a higher percentage of evaluators reported mastery of evaluation techniques through self-assessment, informal learning, and collegiality. In fact, the highest percentage of evaluators reported to have improved their skills in evaluation through self-assessment (94%), incorporating feedback from their colleagues (93%), and having people to ask if they have questions about their practice (92%).

A moderate percentage (84%) reported using ethical guidelines and program evaluation standards to guide their work. As I pointed out earlier in discussing the qualitative descriptors of Expertise, there is contention in the field on how evaluators acquire Expertise. Some interview participants contended that Expertise is acquired through completion of long-term university competencies-based training where evaluators gain knowledge of evaluator competencies. As such, they argued that people get skills for immediate use from attending professional development workshops and informal learning, but not Expertise.

Furthermore, the highest percentage of evaluators also reported collegial behavior, a very important theoretical assertion in the sociological studies of professions. For example, evaluators in this study reported getting feedback from their colleagues and referring clients to someone else if they are not experts in that situation. This may be an indication that evaluators are beginning to specialize in evaluation methods. Interestingly, the items with higher endorsement had lower factor loadings compared to the items with high factor loadings that measure Expertise. This insight may relate to the lack of enforcement of codes of ethics and minimum standards of practice by the professional associations both in Canada and the US.

Table 5.7
Expertise Behaviors and Corresponding Factor Loadings

Expertise Items	**n**	**% Agree**	**Factor Loading**
I self-assess to identify my professional strengths as an evaluator	428	94	0.550
I incorporate feedback from my colleagues to make changes in my professional behavior	418	93	0.518
I have people I can ask if I have questions about my practice	425	92	0.550
I refer clients to someone who is an expert in an evaluation approach if I am not an expert in it	350	87	0.521
I use the program evaluation standards as a safeguard in program evaluation	377	84	0.701
I tell my clients that I follow a professional code of ethics to guide my work	362	83	0.614
I meet with other evaluators informally to keep up-to-date	423	82	0.564
I attend evaluation conferences to keep up-to-date	416	83	0.481

Note: n = the number of respondents who completed each survey item. The factor loadings are based on CFA.

Innovation and Research. Table 5.8 shows how participants endorsed the Innovation and Research items in the measurement model. The participants in the study showed variability in their levels of engagement in Innovation and Research. The findings indicate that most participants in this study reported being consumers of research, not creators of new knowledge to advance the field. This is shown by the relatively small percentage of evaluators (27%) who reported purposefully generating new knowledge to advance the field compared to most participants (60% - 90%) who were primarily consumers of Innovation and Research. In fact, the highest percentage of evaluators (99%) in the study reported applying knowledge from other disciplines and modifying methods created by other people (95%) to their work. This finding provides insights into the interdisciplinary nature of evaluation and how evaluation methods are influenced by practices in other disciplines.

The qualitative analyses point to the lack of funding like Michael Scriven's Faster Forward Fund for empirical research on evaluation, especially at the level of professional associations. Furthermore, very few participants in this study reported that they submitted manuscripts to evaluation journals. The qualitative analyses indicated that the academic members of the associations are more likely than non-academic evaluators to engage in research and to publish their study findings in journals. It appears that sharing insights at conferences is the predominant way in which evaluators engage in generating new knowledge in the field. A moderate percentage (68%) of participants in this study reported sharing their insights at professional conferences while others kept abreast with changes and innovations in the field by attending conferences and reading evaluation-related journals.

Table 5.8
Innovation and Research Behaviors and Corresponding Factor Loadings

Innovation and Research Items	n	% Agree	Factor Loading
I apply knowledge from other disciplines to my evaluation work	423	99	0.624
I modify methods created by other people for use in my own practice	408	95	0.662
I apply established methods to novel situations.	405	89	0.635
I read evaluation-related journals to keep current	425	70	0.582
I integrate empirical research on evaluation to keep up-to-date	408	68	0.543
I share my insights at professional conferences	387	68	0.746
I advance the profession by sharing reflections on my own practice	406	67	0.709
I participate in formal research studies on evaluation	393	52	0.480
I purposefully generate new evaluation knowledge to benefit the field	428	44	0.708
I submit manuscripts to evaluation journals	377	27	0.537

Note: n = the number of respondents who completed each survey item.

Ethical Dispositions. Table 5.9 shows the proportion of evaluators who reported various aspects of ethical behavior. An overwhelming majority of evaluators in this study endorsed items relating to evaluator integrity, particularly the principles of honesty and transparency, followed by the principles of respect for people. Nearly all (99%) respondents said being transparent and truthful about their ethical values and providing the best possible information to stakeholders (97%) were the two most endorsed ethical behaviors. This was closely followed by the respondents' endorsement of items on respect for people, especially respecting the participants' Professional Autonomy to freely provide information and concern about how evaluation decisions will affect program participants.

Regarding evaluator integrity, respondents were concerned about disclosing any conflicts of interests and being upfront about their own values when people hire them. Participants equally endorsed the principle of doing no harm to evaluation participants, providing them with information on how participating in the evaluation would benefit them, and minimizing the participants' discomfort in participating in the evaluation. Interestingly, the items with the highest level of endorsement were those with high factor loadings of .7 while items with factor loadings of .6 were endorsed by 80% - 90% of respondents. The higher endorsement may relate to self-affirmation bias or may indicate that program evaluators are knowledgeable and use guiding principles in their practice of evaluation. However, use of the guiding principles is voluntary since the professional associations are not yet enforcing these ethical principles.

Table 5.9
Ethical Behaviors and Corresponding Factor Loadings

Ethical Disposition Items	n	% Agree	Factor Loading
I am transparent regarding my ethical values when conducting evaluations	411	99	0.716
I provide stakeholders with the best possible information to improve their programs	407	97	0.701
I respect the autonomy of program participants to freely provide evaluation data	393	94	0.775
I am concerned about how evaluation decisions affect program participants	391	94	0.565
I minimize respondents' discomfort in participating in evaluation	384	93	0.676
I provide respondents with information on how participating in evaluation could be beneficial to them	370	89	0.652
I disclose my personal interests when they may influence an evaluation	364	87	0.752
I am concerned about doing no harm to those who participate in an evaluation	362	87	0.638
I am upfront about my own values when people hire me	320	76	0.717

Note: n = the number of respondents who completed each survey item.

Professional Autonomy. Table 5.10 shows the proportion of evaluators who reported various Professional Autonomy behaviors. Overall, a higher percentage of program evaluators in the sample endorsed items relating to Autonomy from clients compared to items relating to Autonomy from the employing organization. Interestingly, the descriptive statistics in the table show that a higher percentage of respondents reported having the Autonomy to select methods (83%) compared to a much smaller percentage (35%) of those who reported Autonomy in ending their participation or terminating employment. In addition, while a high percentage (82%) of respondents said they can refuse to do something they do not believe in if a client tries to make them do it, a relatively moderate number indicated that they can turn down evaluation jobs in situations where the client’s actions might compromise the integrity of the evaluation. For example, 62% of evaluators said they can turn down jobs if they do not agree with the client's intended approach, and 47% turn down jobs where they perceive that the client is unwilling to hear whatever might be found. Table 5.10 displays the different levels of Professional Autonomy behaviors demonstrated by the respondents in this study.

Table 5.10
Professional Autonomy Behaviors and Corresponding Factor Loadings

Professional Autonomy Items	**n**	**% Agree**	**Factor Loading**
I can decide what sort of evaluation method to employ	408	83	0.715
I can refuse to do something I don't believe in if a client tries to get me to do it	407	82	0.560
I turn down evaluation work when I don't agree with my client's intended approach	405	62	0.637
I can decide what evaluation projects to take on	408	61	0.662
I turn down jobs where I think that the client is unwilling to hear whatever might be found	414	47	0.751
I can end my participation in a project if the evaluation is not going well	404	35	0.480

Note: n = the number of respondents who completed each survey item.

It is also interesting to note that the item with the lowest factor loading (*I can end my participation in a project if the evaluation is not going well*) was endorsed by a lower percentage of evaluators. This result is consistent with the hypothesis discussed in Chapter Two regarding Autonomy from employing organizations, which will be discussed in the concluding chapter regarding the status of evaluation as a profession.

Credentialing. Table 5.11 shows the percentage of evaluators who reported their beliefs in Credentialing. Overall, the majority (89%) of respondents endorsed the item relating to using competencies as a framework for pre-service training of evaluators. This was followed by the participants' endorsement of items that related Credentialing as proof of education and experience necessary to demonstrate competence in conducting evaluations. For example, 74% of the participants indicated that Credentialing would tell people they have reached the minimum standards to be a competent evaluator. In addition, 69% believed that being a credentialed evaluator would mean that their peers have recognized their Expertise. The moderately endorsed (52% - 63%) items related to evaluators' reported belief that Credentialing assures employers and clients that evaluators conformed to professional standards and that Credentialing would direct their professional development. The comparison of factor loadings and percentage of evaluators who endorsed an item shows variability in belief in Credentialing. The least (38%) endorsed item was that *Credentialing has made me aware of evaluator competencies*.

Table 5.11
Belief in Credentialing and Corresponding Factor Loadings

Credentialing Items	n	% Agree	Factor Loading
Evaluator competencies should provide a framework for teaching evaluation	391	89	0.761
Credentialing would tell people I have reached the minimum standards to be a competent evaluator	395	74	0.851
Evaluator competencies should be aligned to credentialing	390	74	0.583
Being a credentialed evaluator would mean that my peers have recognized my expertise	395	69	0.881
Being credentialed would encourage me to reflect on my skills	396	63	0.837
Credentialing assures me that those who engage in evaluation conform to certain professional standards	393	63	0.872
Credentialing would direct my professional development	395	63	0.885
Credentialing protects stakeholders from hiring someone who does not have enough of an evaluation background to do a good job	392	52	0.736
Credentialing has made me aware of evaluator competencies	393	38	0.727

Note: n = the number of respondents who completed each survey item.

Differences in Item Means and Effect Size by Country

Table 5.12 shows the item means and standard deviations on professionalism items by country. The table shows that both American and Canadian evaluators highly endorsed ethical dispositions and that the main difference between Canada and the US related to Credentialing. Although both American and Canadian evaluators endorsed Credentialing, Canadians reported feeling more strongly about it shown by the means, the medium, and the effect sizes. The large effect sizes indicated that there was a practical significance in the mean differences between US and Canadian evaluators.

For example, a statistically significant difference was found between American and Canadian evaluators whereby Canadian evaluators believed more so than Americans that (a) Credentialing had made them aware of evaluator competencies (Canada M=2.81, US M=1.68, *d* = .906), (b) Evaluator competencies should be aligned to credentialing (Canada M=3.26, US M=2.63, *d* = .572), (c) Being a credentialed evaluator would mean that their peers have recognized their Expertise (Canada M=3.05, US M=2.58, *d* = .453), (d) Being credentialed would encourage them to reflect on their skills (Canada M=3.01, US M=2.57, *d* = .424), and (e) Credentialing would tell people they have reached the minimum standards to be a competent evaluator (Canada M=3.16, US M=2.72, *d* = .421). Overall, 83% of Canadian evaluators indicated that they “intend to apply” for CE status compared to 17% of American evaluators. As noted by one respondent,

> If suddenly I was going to lose contracts because I couldn’t prove that our staff was credentialed, then we would have to do it, but otherwise, I wouldn’t do it. (Interview respondent, July 27, 2017)

Furthermore, Canadian evaluators reported behaving differently than American evaluators on Expertise and one Ethics items respectively: (a) I tell my clients that I follow a professional code of ethics to guide my work (Canada M=3.19, US M=2.69, *d* = .384), and (b) I am transparent regarding my ethical values when conducting evaluations (Canada M=3.70, US M=3.50, *d* = .313). With regard to Professional Autonomy, a statistically significant difference was found between American and Canadian evaluators whereby American evaluators said (a) they would turn down jobs where they think that the client is unwilling to hear whatever might be found (Canada M=1.63, US M=1.96, *d* = .271), and (b) they can end their participation in a project if the evaluation is not going well (Canada M=1.84, US M=2.09, *d* = .218). There was no statistical difference between Canadian and American evaluators on the submission of manuscripts to evaluation journals (Canada M=1.82, US M=1.82, *d* = .0). The details of the mean differences and effect sizes for each item are shown in Table 5.12.

Table 5.12
Professionalism Means, Standard Deviations, P-values, and Effect Sizes by Country

Quantitative instrument scales	**Canada**		**United States**			
The language used by professional evaluators in describing professionalism	***Mean (scale 0-4)***	***SD***	***Mean (scale 0-4)***	***SD***	***p-values***	***Effect size***
Exp1 I self-assess to identify my professional strengths as an evaluator	3.43	.514	3.29	.684	.053*	.221
Exp7 I meet with other evaluators informally to keep up-to-date	3.20	.757	3.05	.816	.091	.188
Exp13 I incorporate feedback from my colleagues to make changes in my professional behavior	3.35	.646	3.17	.887	.047**	.220
Exp14 I have people I can ask if I have questions about my practice	3.43	.734	3.33	.755	.226	.134
Exp16 I refer clients to someone who is an expert in an evaluation approach if I am not an expert in it	2.75	1.45	2.68	1.44	.653	.049
Et38 I use the program evaluation standards as a safeguard in program evaluation	3.13	1.10	2.83	1.17	.022**	.261
Et29 I tell my clients that I follow a professional code of ethics to guide my work	***3.19***	***1.22***	***2.69***	***1.33***	***.001****	***.386***
Exp8 I attend evaluation conferences to keep up-to-date	3.19	.905	3.06	.980	.225	.136
Exp11 I read evaluation-related journals to keep current	2.94	.830	2.88	.829	.535	.072
Exp10 I integrate empirical research on evaluation to keep up-to-date	2.75	.924	2.70	1.01	.607	.051
Exp17 I submit manuscripts to evaluation journals	1.82	1.10	1.82	1.06	.996	.000
IR18 I purposefully generate new evaluation knowledge to benefit the field	2.14	1.18	2.30	1.09	.206	.143
IR19 I apply knowledge from other disciplines to my evaluation work	3.55	.608	3.52	.697	.682	.045
IR20 I advance the profession by sharing reflections on my own practice	2.81	1.02	2.65	1.01	.148	.158
IR21 I modify methods created by other people for use in my own practice	3.40	.668	3.19	.880	.021**	.257
IR22 I participate in formal research studies on evaluation	2.54	1.19	2.66	1.17	.384	.102
IR23 I apply established methods to novel situations	3.10	.778	2.94	.951	.116	.178
IR24 I share my insights at professional conferences	2.38	1.15	2.33	1.09	.682	.045
Et26 I provide stakeholders with the best possible information to improve their programs	3.75	.587	3.60	.730	.053*	.218
Et27 I disclose my personal interests when they may influence an evaluation.	3.23	1.20	3.24	1.01	.829	.009
Et28 I am upfront about my own values when people hire me	2.73	1.61	2.81	1.46	.629	.053
Et30 I respect the autonomy of program participants to freely provide evaluation data	3.64	.746	3.44	.951	.045**	.225
Et32 I am concerned about how evaluation decisions affect program participants	3.53	.82	3.51	.78	.772	.013

Table 5.12 (Continued)

Quantitative instrument Scales ***The language used by professional evaluators in describing professionalism***	**Canada** *Mean (scale 0-4)*	*SD*	**United States** *Mean (scale 0-4)*	*SD*	*p-values*	*Effect size*
Et33 I provide respondents with information on how participating in evaluation could be beneficial to them	3.431	.89	3.28	.998	.166	.157
Et35 I am transparent regarding my ethical values when conducting evaluations	***3.70***	***.459***	***3.50***	***.709***	***.005****	***.313***
Et36 I minimize respondents' discomfort in participating in evaluation	3.43	.832	3.27	.948	.112	.175
Et37 I am concerned about not doing harm to those who participate in an evaluation	3.32	1.09	3.34	.960	.882	.020
Et39 I turn down jobs where I think that the client is unwilling to hear whatever might be found	***1.63***	***1.63***	***1.96***	***.960***	***.067***	***.271***
Au40 I can refuse to do something I don't believe in if a client tries to get me to do it	3.03	1.22	3.00	1.23	.830	.024
Au41 I turn down evaluation work when I don't agree with my client's intended approach	2.25	1.37	2.48	1.32	.126	.172
Au43 I can decide what evaluation projects to take on	2.42	1.32	2.65	1.24	.113	.182
Au47 I can end my participation in a project if the evaluation is not going well	1.84	1.14	2.09	1.15	.052*	.218
Au44 I can decide what sort of evaluation method to employ	2.96	.86	3.02	.89	.552	.080
Cre53 Credentialing would tell people I have reached the minimum standards to be a competent evaluator	***3.16***	***1.01***	***2.72***	***1.06***	***.001****	***.421***
Cre58 Credentialing assures me that those who engage in evaluation conform to certain professional standards	2.81	1.01	2.56	.98	.020**	.253
Cre51 Being credentialed would encourage me to reflect on my skills	***3.01***	***1.11***	***2.57***	***1.00***	***.001****	***.424***
Cre52 Credentialing would direct my professional development	***2.94***	***1.08***	***2.58***	***1.05***	***.002****	***.340***
Cre57 Being a credentialed evaluator would mean that my peers have recognized my expertise	***3.05***	***.92***	***2.58***	***1.09***	***.001****	***.453***
Cre56 Credentialing protects stakeholders from hiring someone who does not have enough of an evaluation background to do a good job	2.61	.93	2.38	1.02	.041**	.232
Cre55 Credentialing has made me aware of evaluator competencies	***2.81***	***1.34***	***1.68***	***1.20***	***.001****	***.906***
Cre54 Evaluator competencies should be aligned to credentialing	***3.26***	***.99***	***2.63***	***1.15***	***.001****	***.572***
Cre59 Evaluator competencies should provide a framework for teaching evaluation	3.28	.90	3.04	1.07	.003***	.236

Note: *p < .05, **p < .01, ***p <.001. The effect size (d), was computed for groups with different sample size (Canada=118 and US=248) (Cohen's d, Hedges' g). Effect size of >0 to.1 = no effect, .2 to .4 = small; .5 to .7 = medium, and .8 to ≥ 1.0 = large (Cohen, 1988). SD is standard deviation. Items with effect size of .3 and above are boldfaced and italicized.

Multivariate Analysis of Variance (MANOVA)

Results of evaluators' professionalism scores were analyzed per sub-scale at the level of country, evaluator workplace, gender, credentialed evaluator (CE) status, skill level, education, age, employment status, and level of experience. The analyses are presented in boxplots, standard deviation error bars, and tables to answer research question three. The tables show the statistical significance of the mean differences between these groups for the Multivariate Analysis of Variance (MANOVA). The boxplots represent an assortment of sub-scale scores divided into four equal-sized groups or quartiles with each group depicting 25% of all scores. The box in the middle of each plot represents the middle 50% of sub-scale scores while the line in the box represents the midpoint or the median score, and the lower and upper whiskers represent scores outside the middle 50%.

The boxplots represent the range and pattern of responses for each sub-scale on a four-point Likert scale of 1 to 4, with 1 indicating strongly disagree and 4 strongly agree. A shorter box indicates a high level of agreement among evaluators on that sub-scale while taller boxes indicate variability of opinions on the items making up the sub-scale. Furthermore, overlaps or boxes that are much higher or lower than the others indicate a difference in the mean scores for the different evaluator categories for those sub-scales. I have used standard deviation error bars to further illustrate the variability of responses between groups and conducted MANOVA to determine whether there were any differences between groups on professionalism sub-scale scores. The error bars represent one standard deviation of the mean score for each sub-scale. The mean of the sub-scale is represented by the dot in the middle of the bars.

I have used the median and mean since the semi-interquartile range (based on median) and the standard deviation (based on mean) are the most commonly used statistics to describe variability in a set of scores (Thorndike & Thorndike-Christ, 2010). The wider the error bars, the more variable the responses are, and the shorter the error bars, the higher the level of agreement among the participants in the study.

Evaluator Professionalism by Country. Figures 5.4, 5.5, and 5.6 depict evaluator reported behaviors on Expertise, Innovation and Research, Ethics, Autonomy, and beliefs in Credentialing sub-scales, and overall Professionalism for Canadian and American evaluators in the sample. Here, I sought to understand whether there were differences in Professionalism behaviors by country by comparing the sum scores of the sub-scales and overall Professionalism score. Overall, the boxplots indicate that both Canada and the US evaluators in the sample consistently agreed to behaving ethically when conducting program evaluations as indicated by the relatively elevated Ethics boxplot placed well above the other boxplots in Figure 5.4. The North American evaluators in the sample reported variable opinions in Credentialing shown by the longer boxplot, although their reported level of Expertise appeared higher compared to the other sub-scales.

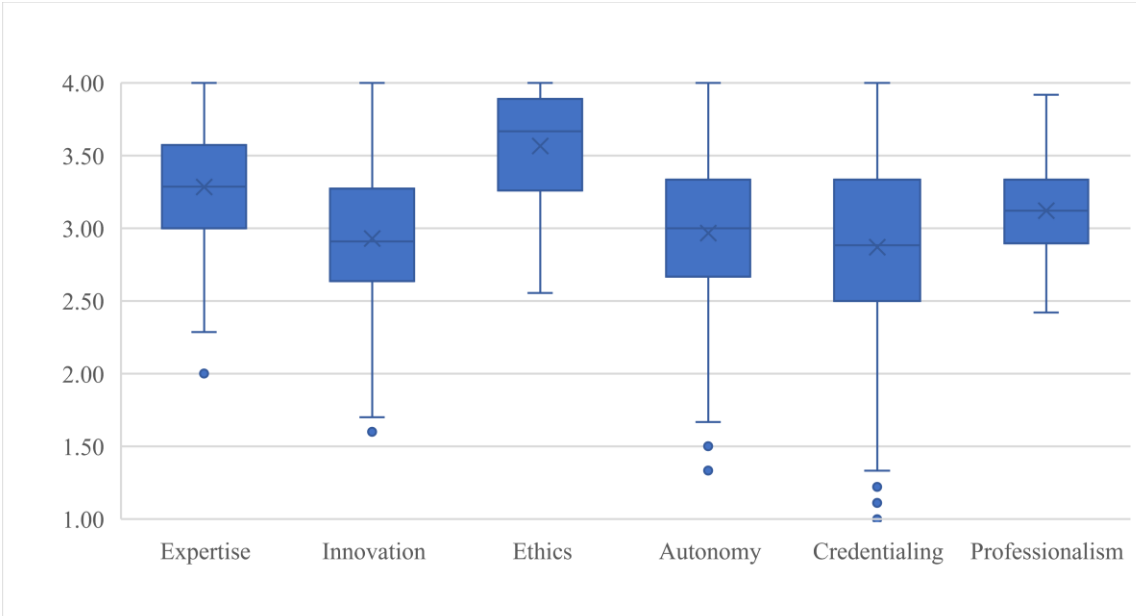

Figure 5.4. Total Professionalism measures for USA and Canada evaluators.

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4).

To further understand the reported overall Professionalism behavior by country, I compared American and Canadian evaluators in the sample on the overall Professionalism score. Figure 5.5 shows that Canadian evaluators in the sample reported higher levels of overall Professionalism on average compared to American evaluators as indicated by the elevated position of their boxplot. It would be interesting to speculate the reasons why the Canadian evaluators in the sample perceived being more professional evaluators than their American counterparts in the sample.

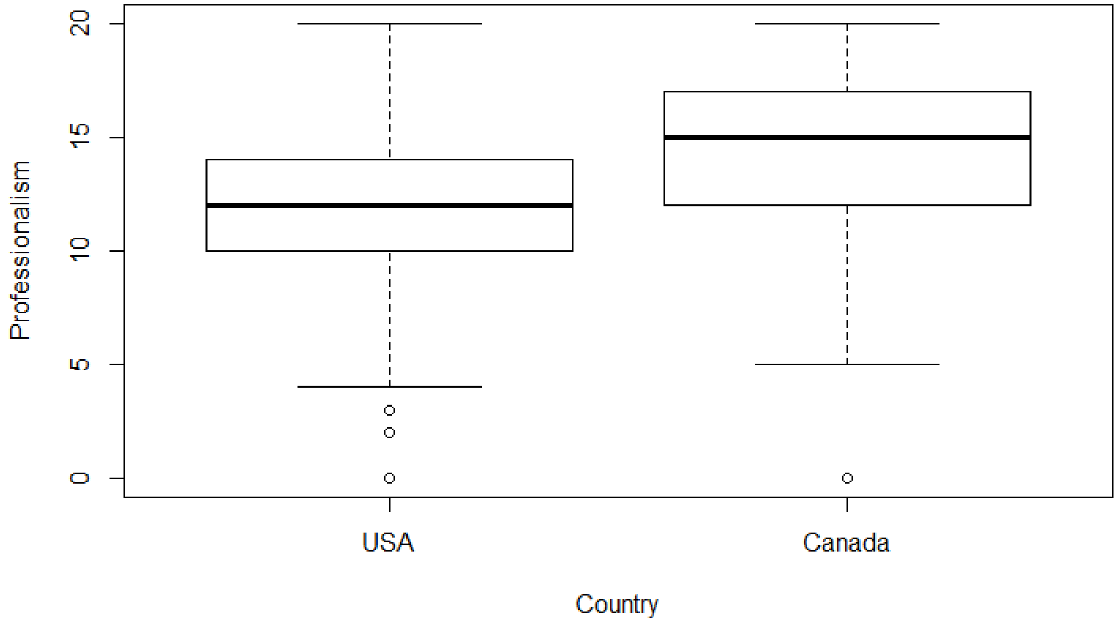

Figure 5.5. Comparison of Overall Professionalism Means by Country.

I sought to understand the differences in professionalism behaviors by country by comparing the US and Canadian evaluators in the sample on all five Professionalism sub-scales. Figure 5.6 indicates that there were similar levels of agreement among Canadian and US evaluators in the sample on Ethics, Expertise, and Innovation and Research sub-scales, although their agreement on Innovation and Research was on the lower level of the four-point scale compared to Expertise and Ethics. Furthermore, the boxplots indicate that the median scores for Expertise and Ethics sub-scales, and overall Professionalism appear above the “Agree” point on the scale, suggesting that evaluators in the sample highly endorsed these Professionalism sub-scales on a higher level of the scale.

Although Canadian evaluators endorsed the Credentialing sub-scale at a much higher level of the four-point scale, they held variable opinions on Credentialing compared to US evaluators, indicated by the longer boxplot. The US evaluators in the sample had more outliers on the Credentialing sub-scale on both the "Strongly Agree" and "Strongly Disagree" points compared to Canadian evaluators with outliers only in the "Strongly Disagree" point. Both US and Canadian evaluators had outliers on "Strongly Disagree" points of the Autonomy sub-scale.

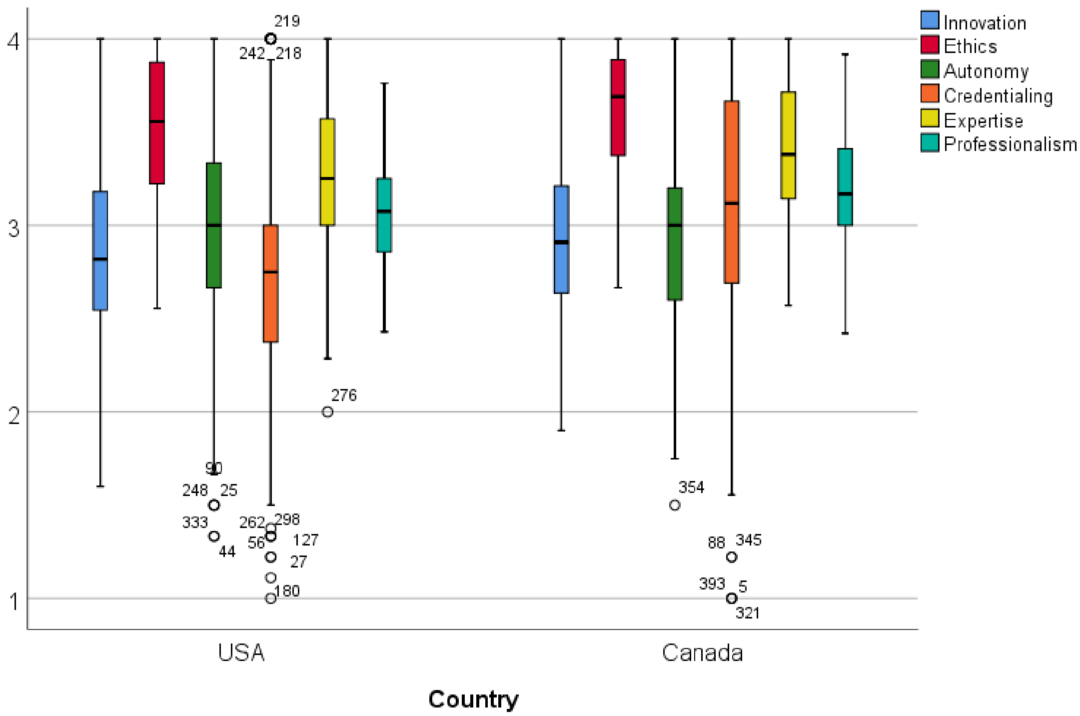

Figure 5.6. Professionalism Sub-Scales Comparison by Country.

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4).

To further understand the professionalism behavioral patterns and beliefs in Credentialing of evaluators, I compared American and Canadian evaluators in the sample on the sub-scales looking at one standard deviation for the mean. The error bars in Figure 5.7 represent the sub-scales and the overall Professionalism scale by country. The first error bar on the top level represents the overall Professionalism, the second bar from the top represents the Expertise sub-scale, the third bar represents the Expertise sub-scale, the fourth bar from the top represents the Professional Autonomy sub-scale, the fifth bar represents the Ethics sub-scale, and the sixth bar represents the Innovation and Research sub-scale. This order is the same for both Canada and the US.

Figure 5.7 further shows that North American evaluators in the sample reportedly had mixed feelings and variability of opinions on the Credentialing and Autonomy sub-scales, illustrated by the wider standard deviation error bars. The Canadian evaluators had higher means for the Credentialing and Expertise sub-scales and the overall Professionalism scale compared to American evaluators, indicating that Canadian evaluators in the sample were more likely than American evaluators to agree and strongly agree on these sub-scales. Furthermore, although the error bars reinforce similarity of the Canadian and American evaluators' agreement on the Ethics and the Innovation and Research sub-scales, American evaluators in the sample had a slightly higher mean score on the Professional Autonomy sub-scale.

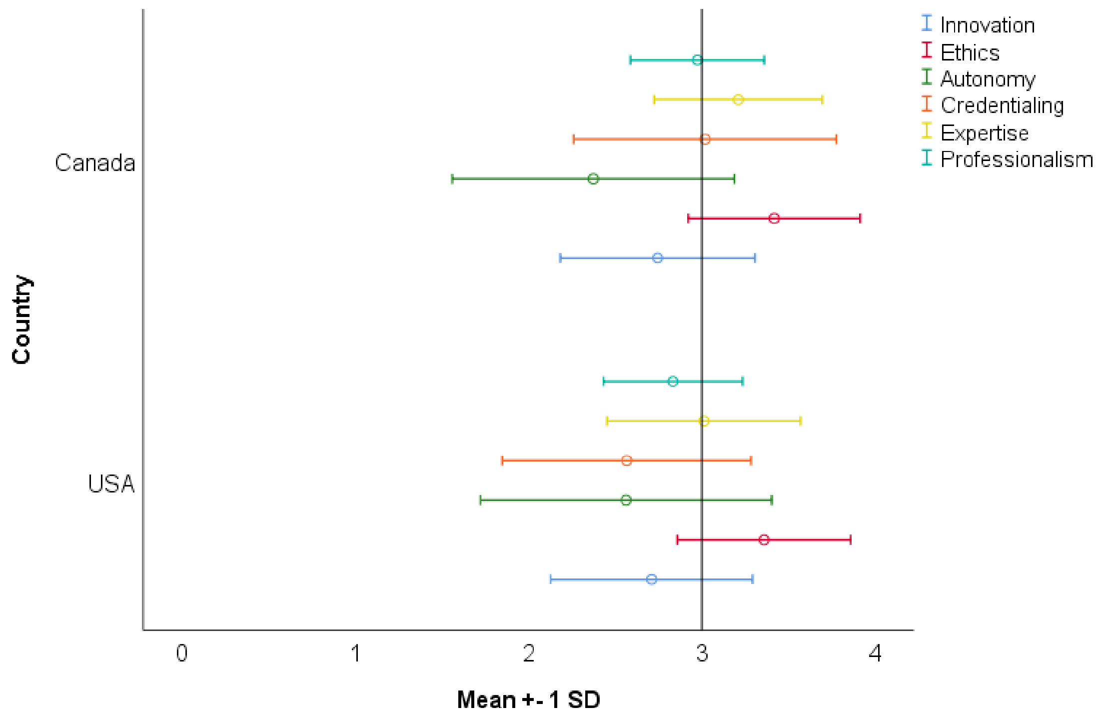

Figure 5.7. Variability in Professionalism Sub-Scale Means by Country

Note: The rating scale used for assessing the five sub-scales of professionalism was based on a four-point scale of strongly disagree (1), disagree (2), agree (3) - point of the vertical line, strongly agree (4). The round dots represent the means.

I conducted a one-way MANOVA to assess if the country differences on a linear combination of the Professionalism sub-scales were statistically significant. The assumptions of independence of observations and homogeneity of variance/covariance were checked and met. Box's M (13.27) test was non-significant, (p = .600) > α (.001), indicating that there were no significant differences between the covariance matrices. As such, I used Wilk's Lambda criterion to test the difference by country since the assumption of homogeneity of covariance was not violated. Furthermore, bivariate scatterplots were checked for multivariate normality.

There was a statistically significant difference in professionalism in evaluators based on their country, F (5, 348) = 9.820, p < .001, Wilk's ∧ = .876, and a large effect size (partial eta squared) of .124, according to Cohen (1988). Power to detect the effect was 1. Thus, the hypothesis that evaluator professionalism behaviors and their attitudes to Credentialing was influenced by their country of practice was confirmed. The multivariate η^2 = .124 indicated that approximately 12.4% of multivariate variance of professionalism behaviors was associated with the evaluator's country.

Given the significance of the overall test, the univariate main effects were examined, and the results are depicted in Table 5.13. To protect against Type I error, I used the traditional Bonferroni procedure and tested each of the five ANOVAs at the statistical significance at p < .01 (.05 divided by the number of dependent variables or number of ANOVAs conducted). Examination of the coefficients for the linear combinations distinguishing the US and Canada indicated that the three factors that contributed the most to distinguishing the two countries were Credentialing, Expertise, Autonomy sub-scales, and the overall Professionalism. The US and Canadian evaluators in the sample did not differ significantly on the Innovation and Research and the Ethics sub-scales.

Based on the results of the univariate main effects in Table 5.13, Canadian evaluators reportedly felt that Credentialing is more important on average compared to American evaluators, with a large observed power of 1 and medium effect size of .082 (Cohen, 1988). The table also shows that Canadian evaluators in the sample reported higher levels of Expertise and the overall Professionalism on average compared to American evaluators, with small effect sizes of .028 and .024 respectively (Cohen, 1988).

Table 5.13
Professionalism Mean Differences by Country (at p < .05 level)

		Canada		USA					
	df	M	SD	M	SD	F	*Partial Eta (d)*	*p*	*Observed Power*
Expertise	352	3.21	.48	3.03	.54	9.98	.028	.002***	.883
Innovation	352	2.75	.56	2.71	.59	.332	.001	.565	.089
Ethics	352	3.43	.48	3.37	.49	1.07	.003	.302	.178
Autonomy	352	2.37	.81	2.56	.84	3.93	.011	.048*	.507
Credentialing	352	3.01	.76	2.56	.70	31.49	.082	.001***	1.000
Professionalism	352	2.99	.38	2.86	.39	8.71	.024	.003***	.837

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed from MANOVA table. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect; and .039 to .2 = desired effects (Cohen, 1988).

I conducted a non-parametric independent samples Mann-Whitney Test (t-test) to compare the difference in means by country. The results confirmed that there were statistically significant differences between the US and Canadian evaluators on three sub-scales: Professional Autonomy (*p* <.026, effect size *d* = .2), Credentialing (*p* <.001, effect size *d* = .6), and Expertise (*p* = .001, effect size *d* = .4), including the overall Professionalism (*p* = .005, effect size *d* = .4). The post hoc differences and somewhat small to medium to large effect sizes indicated that professionalism behaviors in the Credentialing and the Expertise sub-scales and the overall Professionalism were practically important.

Although the significant overall *F* and significant post hoc differences indicated that there was a tendency for the evaluators' country to be related to their professionalism behaviors, the effect size was very small for the Professional Autonomy sub-scale given the statistical power of 1. There were no statistically significant differences in the Innovation and Research and the Ethics sub-scales for the US and Canadian evaluators.

Evaluator Professionalism by Evaluator Workplace. Figure 5.8 depicts program evaluators' reported professionalism behaviors on the Expertise, Innovation, Ethics, Autonomy sub-scales and their beliefs in the Credentialing sub-scale by workplace. Here, I sought to understand how evaluation managers, external, internal, and academic evaluators compared on their perceptions and reported demonstration of professionalism. First, the figure shows that Ethics boxplots are shorter and a little higher than the other boxplots at the higher values of "Agree" to "Disagree" points on the scales for all four categories of evaluators. It appears that North American evaluators in the sample had a high level of agreement on the Ethics sub-scale regardless of their place of work. However, evaluation managers showed variability of opinions on the Ethics sub-scale.

Furthermore, despite similar levels of agreement on the Expertise sub-scale by all evaluator workplaces, the external and academic evaluators expressed a higher level of agreement on this sub-scale with responses ranging between the "Agree" and "Strongly Agree" points on the scale. While academic evaluators reported higher level of agreement on the Innovation and Research sub-scale than those in the other evaluator workplaces, internal evaluators showed the lowest level of agreement on the Innovation and Research sub-scale based on their mean and median scores. Interestingly, academic evaluators showed much variability of opinions on the Credentialing sub-scale, but evaluation managers had the most agreement on Credentialing, followed by internal evaluators, meaning these two groups reported believing more strongly that credentialing was important in evaluation practice. Finally, external and academic evaluators showed a higher level of agreement on the Professional Autonomy sub-scale while internal evaluators and evaluation managers showed much variability of opinions on this sub-scale.

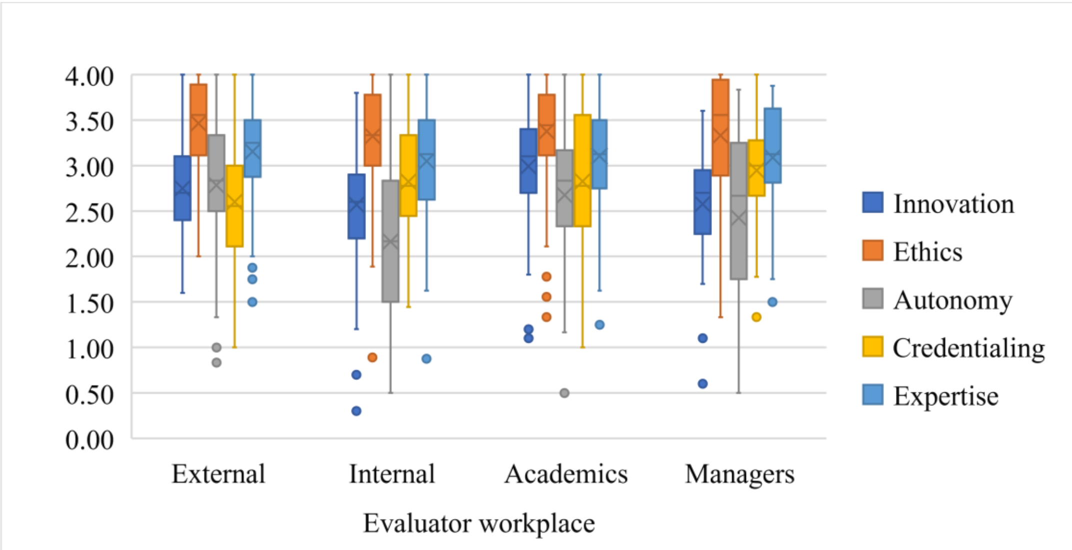

Figure 5.8. Variability in Professionalism Behavior by Workplace

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4). The "X" in the boxplots indicates the mean scores for each sub-scale.

To understand the variability of these responses further, I plotted error bars of one standard deviation of the means for each Professionalism sub-scale. The error bars in Figure 5.9 represent the sub-scales and the overall Professionalism scale by country. The first error bar on the top level represents evaluation managers, the second bar from the top represents academic evaluators, the third bar represents internal evaluators, and the fourth bar from the top represents external evaluators. This order is the same for all subsequent graphs representing each Professionalism sub-scale for both Canada and the US.

Figure 5.9 confirms less variability of opinions on Ethics by evaluators in the study except the US evaluation managers whose opinions varied a lot on this sub-scale. The Canadian evaluation managers and academic evaluators consistently agreed and endorsed ethical behaviors at higher mean scores compared to their counterparts in the US. Interestingly, the internal and external evaluators in both Canada and the US showed somewhat similar endorsement of the Ethics sub-scale based on their mean scores. The somewhat similar means is a clue that the difference in the Ethics means scores for Canada and the US evaluators may not be statistically insignificant.

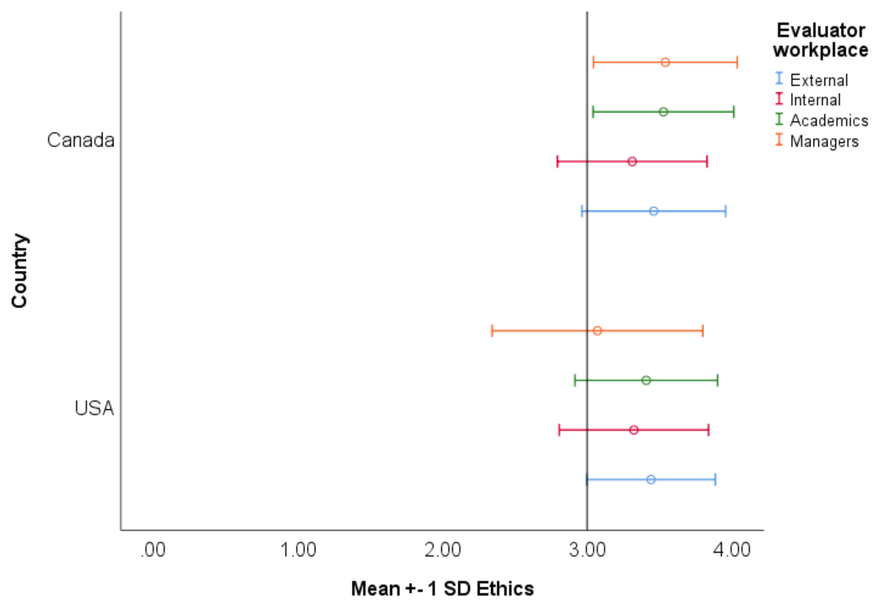

Figure 5.9. *Variability in Ethics Means by Workplace*
Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on “agree” and the round dots represent the means.

Figure 5.10 illustrates that there is variability of opinions on Expertise, especially by the Canadian evaluation managers, and all US evaluator groups except academic evaluators. The four error bars represent evaluators by workplace for both Canada and the US. The first error bar represents evaluation managers, the second bar represents academic evaluators, the third bar represents internal evaluators, and the fourth bar represents external evaluators in the same order for both countries.

Overall, the North American evaluators in the sample endorsed items on the Expertise sub-scale between the “Agree” (3) and “Strongly Agree” (4) points of the scale. Aside from Canadian internal evaluators whose mean score was the lowest on the Expertise sub-scale as a group, the Canadian evaluators had higher means on expertise compared to American evaluators. More specifically, Canadian academic and external evaluators reported higher means for the Expertise sub-scale than their US counterparts. Interestingly, within country comparisons in the US indicated that academic evaluators reported a slightly higher level of expertise than the evaluation managers and the internal and external evaluators.

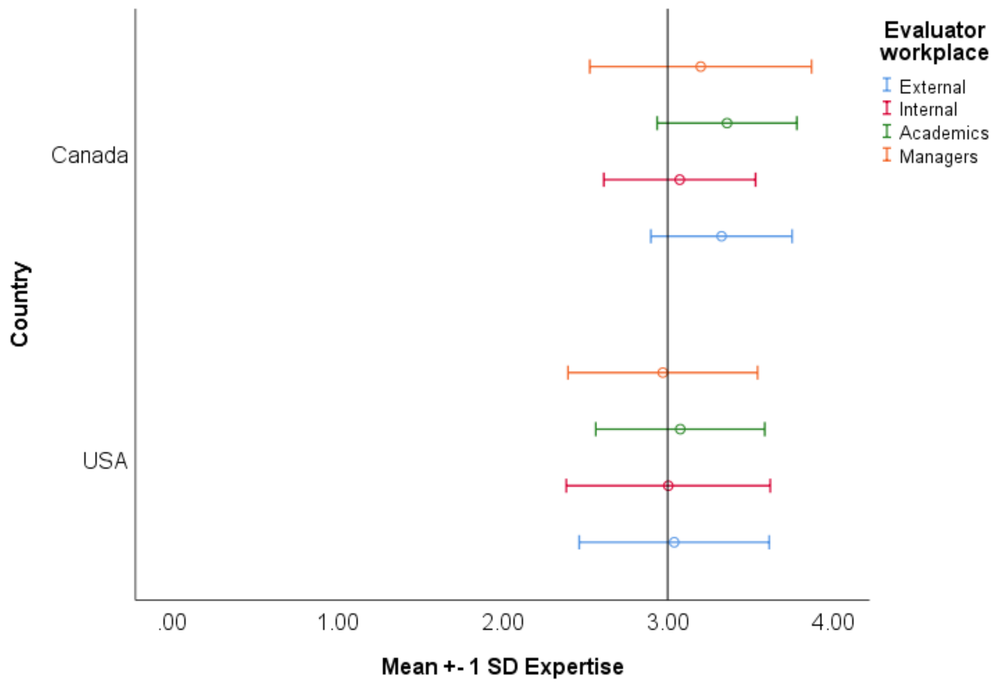

Figure 5.10. Variability in Expertise Means by Evaluator's Workplace

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

Figure 5.11 confirms that academic evaluators in Canada and the US reported higher levels of Innovation and Research shown by their almost identical means scores for this sub-scale. As in the previous graph, the order of the error bars in all ensuing graphs for evaluator workplace starts with evaluation managers at the top, followed by academic evaluators, internal evaluators, and the fourth bar represents external evaluators in the same order for both countries. The figure shows that evaluation managers and internal evaluators in Canada and the US reported the lowest mean scores for the Innovation and Research sub-scale compared to the other evaluation groups. However, Canadian evaluation managers demonstrated more variability of opinions on the

Innovation and Research sub-scale than their counterparts in the US. In fact, external evaluators in both Canada and the US showed similar means scores and variability of opinions on this sub-scale. As a group, North American evaluators in the sample endorsed the Innovation and Research sub-scale at the lower levels of Disagree (2) and Agree (3) points on the four-point scale.

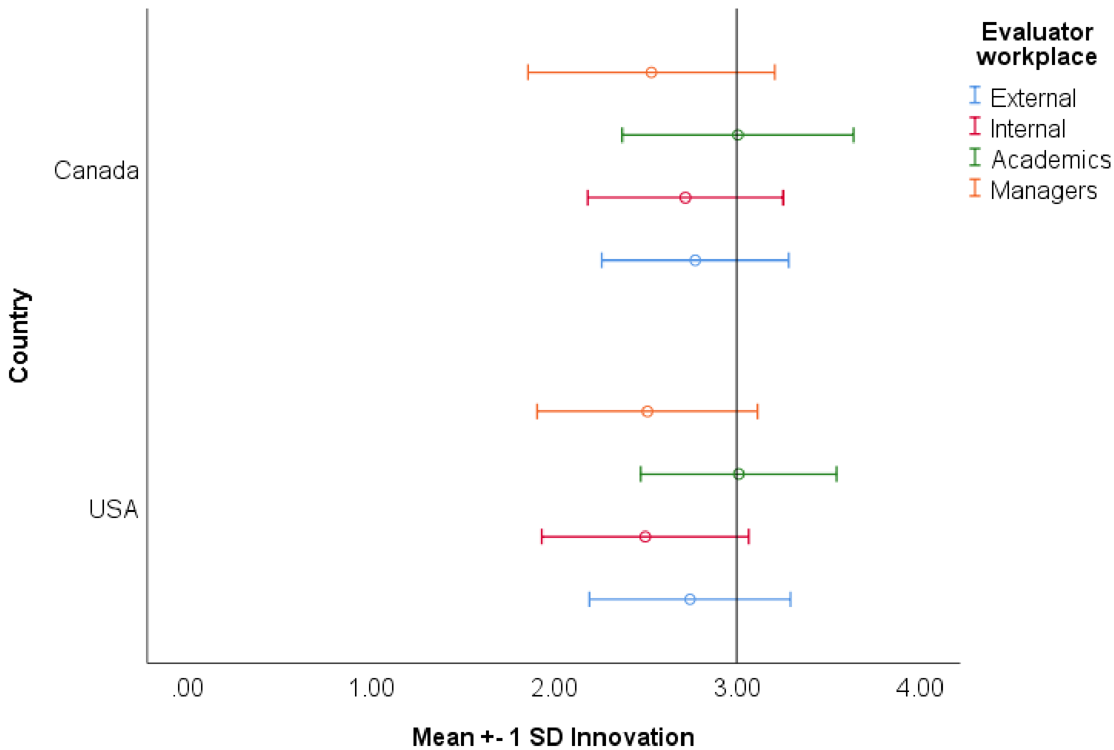

Figure 5.11. Variability in Innovations Means by Evaluator's Workplace

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

Figure 5.12 confirms that external evaluators in the sample in both countries and US academic evaluators reported higher levels of the Professional Autonomy sub-scale than other evaluator groups. In fact, internal evaluators in Canada and the US reported the lowest levels of professional autonomy compared to all evaluator groups based on their mean scores. Not only did US external evaluators in the sample report higher levels of professional autonomy compared to their Canadian counterparts, their mean score was higher than for the other evaluator groups in the US. Evaluation managers and internal evaluators in the US and Canada had much variability of opinions on the Professional Autonomy sub-scale as demonstrated by the wider error bars. Interestingly, evaluators in the sample endorsed the Professional Autonomy sub-scale items between the "Disagree" (2) and "Agree" (3) levels of the scale.

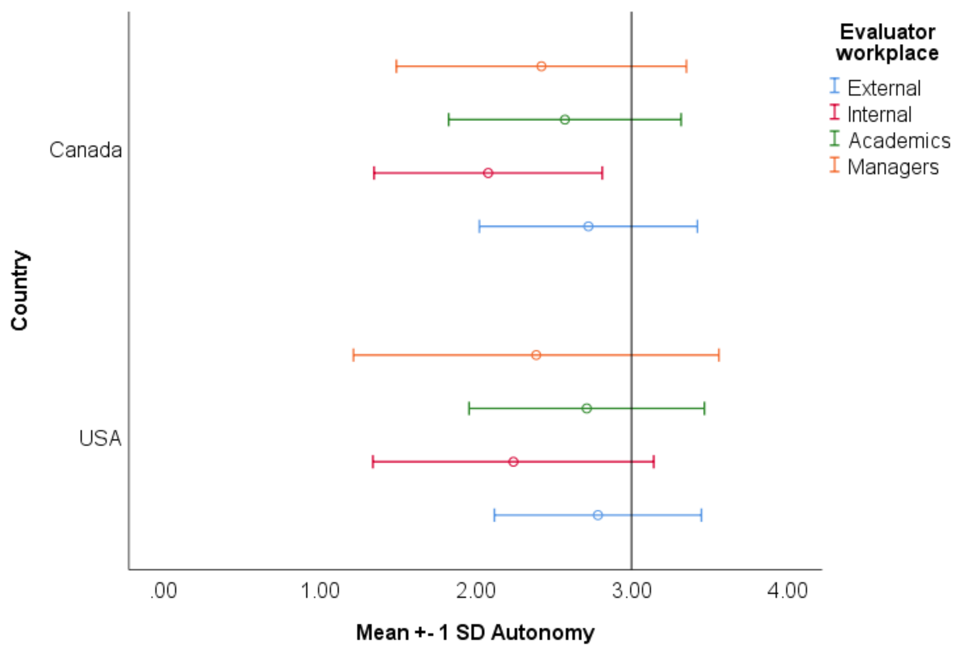

Figure 5.12. Variability in Professional Autonomy Means by Evaluator's Workplace and Country

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

Figure 5.13 shows that while the estimated mean scores for Canadian evaluators on the Credentialing sub-scale appeared at the higher range of “Agree” (3) and “Strongly Agree” (4) points on the sub-scale except for the external evaluators, the mean scores of the US evaluators’ responses fell within “Disagree” (2) and “Agree” (3) points on the scale. The figure confirms that Canadian evaluation managers and academic evaluators reported higher beliefs in credentialing of evaluators than internal and external evaluators as a group. Although the Canadian external evaluators had higher mean score than their US counterpart, they appeared to have more variability of opinions on the Credentialing sub-scale than their American counterparts, shown by the wider error bars. Furthermore, within country comparisons indicated that academic and external evaluators in the US reported higher levels of beliefs in credentialing than internal evaluators and evaluation managers. Interestingly, academic evaluators in the US demonstrated much variability in opinions on the Credentialing sub-scale, shown by the wider error bar.

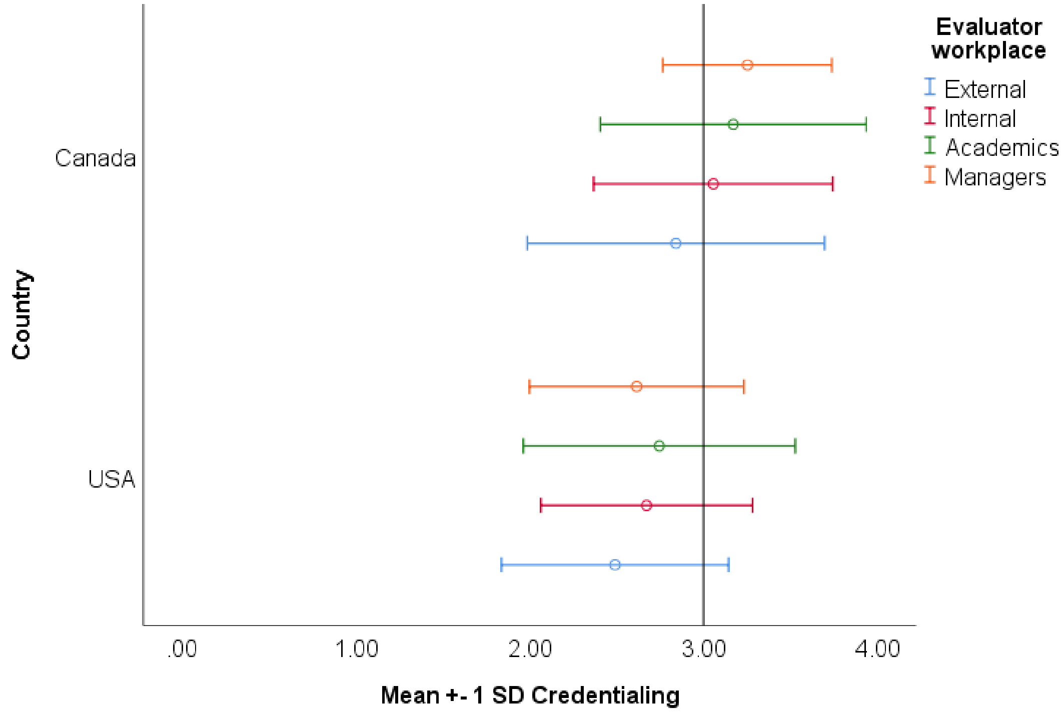

Figure 5.13. Variability in Credentialing Means by Evaluator's Workplace and Country

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

Figure 5.14 shows that academic evaluators in Canada and the US had higher the overall Professionalism mean scores than all other evaluator groups compared within country. The figure also shows that external evaluators in Canada had higher means than their US counterparts. Evaluation managers and external evaluators in Canada had a somewhat similar level of professionalism behaviors. Interestingly, internal evaluators in both Canada and US and evaluation managers in the US had the lowest the overall Professionalism means compared to the rest of the groups.

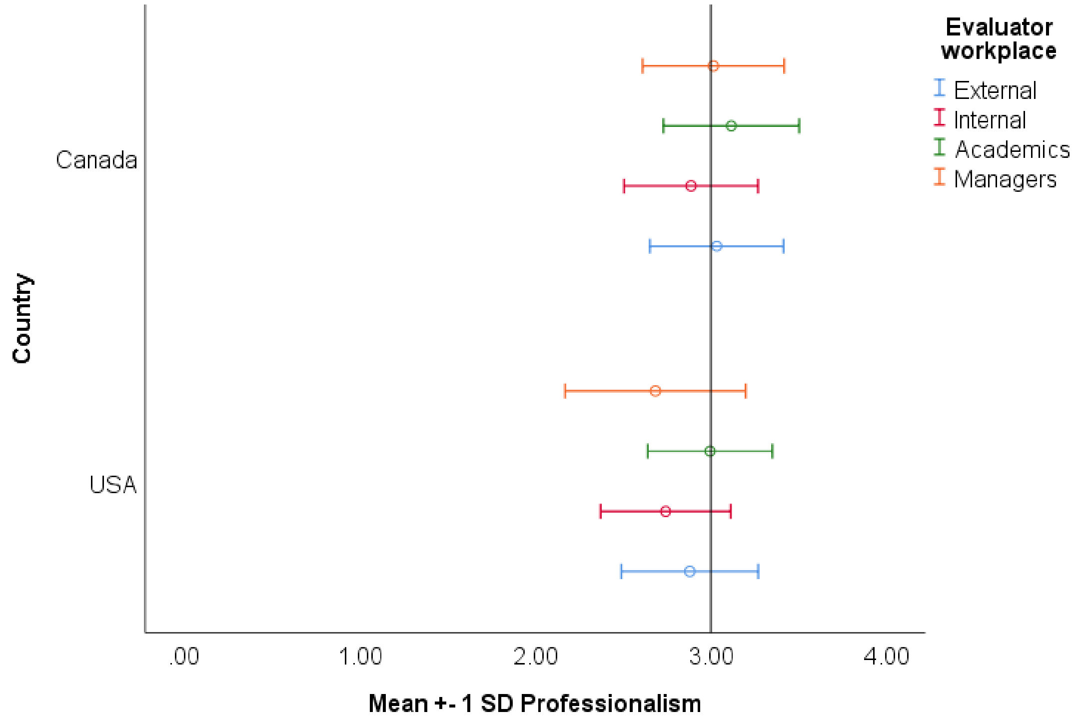

Figure 5.14. Variability in Professionalism Means by Workplace and Country

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

The key takeaway from these analyses represented in the boxplots and the standard deviation error bars is that evaluators in Canada and the US varied in opinions and reported different levels of perceived Professionalism based on their workplace. At the sub-regional level, the Canadian and US evaluators in the sample seemed to agree that ethics was the most important aspect of professionalism based on their endorsement of the items in the Ethics sub-scale. At the country level, the key difference between Canada and the US was credentialing. The error bars showed that Canadian evaluators reported higher levels of expertise on average compared to American evaluators. In fact, Canadian academics and external evaluators had the highest expertise means scores on average than all other evaluator groups. Finally, academic evaluators reported a higher level of innovation and research and the overall Professionalism scale on average than other evaluator groups; external evaluators and academic evaluators reported the highest levels of professional autonomy on average than other evaluators; and evaluation managers reported the greatest belief in credentialing on average that other evaluator groups.

I conducted a one-way MANOVA to test the hypothesis that evaluator workplace had a relationship with evaluator professionalism behaviors. The assumption of independence of observations was checked and met. Box's M (73.43) test was significant, p (.008) < α (.05), indicating that the assumption of homogeneity of variance/covariance was not met. This was confirmed by the significance of the Levene's test for the Ethics (p = .043) and the Professional Autonomy (p = .001) sub-scales. I used Pillai's trace criterion (a test statistic that is very robust and not highly linked to assumptions about the normality of the distribution of the data) to assess the statistical significances among the evaluator workplace on a linear combination of the five professionalism sub-scales since the assumption of homogeneity of variance-covariance was violated.

Overall, the one-way MANOVA revealed a significant difference in evaluator professionalism behavior by workplace, Pillai's $\vee$ = .234, F (15, 1008) = 5.68, p < .001, and a medium effect size (partial eta squared) of .080, according to Cohen (1988). Power to detect the effect was 1. Thus, the hypothesis that evaluator professionalism behaviors and their attitudes to credentialing had a relationship with their workplaces was confirmed. The multivariate η^2 = .08 indicates that approximately 8% of multivariate variance of professionalism behavior is associated with evaluator workplace.

Given the significance of the overall multivariate test, Table 5.14 shows the univariate main effects. Using the traditional Bonferroni procedure, I tested each of the five ANOVAs at the statistical significance of $\alpha < .01$ (.05 divided by the number of dependent variables or number of ANOVAs conducted). The results show that evaluator workplace had a statistical significant univariate effects on two Professionalism sub-scales at the .01 adjusted alpha level: Innovation and Research $F(3, 338) = 8.83$, $p < .001$, medium effect size (partial eta square = .073, power = .976), and Professional Autonomy $F(3, 338) = 13.35$, $p < .001$, large effect size (partial eta square = .106, power to detect the effect was .999). The overall Professionalism behavior $F(3, 338) = 4.19$, $p < .006$, small effect size (partial eta square = .036, power = .854) was also significantly different by evaluator workplace. The evaluator workplace had no significant differences on Expertise ($p = .535$), Ethics ($p = .225$), and Credentialing ($p = .016$).

Table 5.14
Univariate Mean Differences by Evaluator Workplace (at α < .01)

	Error df	External		Internal		Academics		Managers		F	*Partial Eta Squared*	*p*	*Observed Power*
		M	SD	M	SD	M	SD	M	SD				
Innovation	338	2.75	0.53	2.57	0.61	2.99	0.59	2.58	0.64	8.826	.073	.001***	.976
Ethics	338	3.46	0.45	3.32	0.54	3.38	0.60	3.33	0.66	1.462	.013	.225	.182
Autonomy	338	2.78	0.68	2.17	0.83	2.67	0.72	2.42	1.01	13.350	.106	.001***	.999
Credentialing	338	2.60	0.72	2.82	0.64	2.83	0.78	2.95	0.59	3.471	.030	.016	.556
Expertise	338	3.16	0.52	3.05	0.57	3.10	0.55	3.09	0.64	.730	.006	.535	.073
Professionalism	338	2.95	0.37	2.82	0.40	3.01	0.38	2.89	0.46	4.193	.036	.006***	.671

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d) was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect – and .039 to .2 = desired effects (Cohen, 1988).

I conducted a post hoc Games-Howell non-parametric test to understand which pairs of means were significantly different for the four levels of evaluator workplace because the assumption of homogeneity of variance-covariance was violated. I had previously controlled for Type I error across the five univariate ANOVAs by testing each at the .01 alpha level. To be consistent with this decision, I tested each comparison at the .01 adjusted alpha level for the ANOVAs divided by the five pairwise comparisons for the dependent variables (.01/5 = .002 adjusted alpha level).

Significant workplace pairwise differences were obtained in two sub-scales for Innovation and Research and Professional Autonomy and in the overall Professionalism scale. Significant pairwise differences were obtained in the Innovation and Research sub-scale between academic evaluators and both internal evaluators and evaluation managers. The mean scores for the Innovation and Research sub-scale were statistically significantly different between academic evaluators and internal evaluators ($p = .001$) and academic evaluators and evaluation managers ($p = .011$), but not between academic and external evaluators ($p = .018$) or between external evaluators and evaluation managers ($p = .479$). There was also no significant difference between internal evaluators and evaluation managers ($p = 1$) on the Innovation and Research sub-scales. Academic evaluators were more likely than the other evaluator groups to report higher Innovation and Research behavior based on the mean scores.

The mean scores for the Professional Autonomy sub-scale were statistically significantly different between external and internal evaluators ($p < .001$) and between academic and internal evaluators ($p < .001$), but not between evaluation managers and both academic evaluators (p = .577) and external evaluators (p = .230) or between evaluation managers and internal evaluators (p = .545). External and academic evaluators were more likely than internal evaluators or managers to report higher perception of professional autonomy behaviors based on the means scores.

The overall Professionalism score was statistically significantly different between academic and internal evaluators ($p < .005$), but not between academic evaluators and both evaluation managers ($p < .540$) and external evaluators ($p < .685$). Academic evaluators were more likely than internal evaluators to report higher perception on overall professionalism based on the mean scores. Finally, there was no statistically significant difference in pairwise comparisons for Ethics, Expertise, and Credentialing by workplace at the adjusted alpha level of .002.

Evaluator Professionalism by Credentialed Evaluator Status. I was interested in understanding if the professionalism behavioral patterns and beliefs in credentialing was affected by Credentialed Evaluator (CE) status. Here, I compared credentialed evaluators (CEs) and non-credentialed evaluators (Non-CEs) on the five sub-scales of my model to understand their perceptions and demonstration of professionalism behaviors. Figure 5.15 shows the level of endorsement of the items for each sub-scale by both the CEs and non-CEs.

The figure shows that CEs endorsed credentialing, ethics, expertise, and innovation and research more highly than non-CEs, and they had high levels of agreement on ethics compared to non-CEs. Interestingly, both CEs and non-CEs had mixed feelings and much variability of opinions on professional autonomy, demonstrated by the wider boxplots. Notably, both CEs and non-CEs endorsed professional autonomy at similar levels, demonstrated by the somewhat equal median for this sub-scale.

The CEs' median scores appeared higher than those of non-CEs, especially on the Credentialing, Expertise, and Innovation sub-scales. This difference was evident in the relatively higher overlaps in the CEs' boxplots for the Credentialing, Expertise, and Innovation sub-scales compared to non-CEs. The sizes and the elevated positions of the boxplots suggest that the difference in opinions and reported professionalism behaviors on these sub-scales between CEs and non-CEs was likely significant.

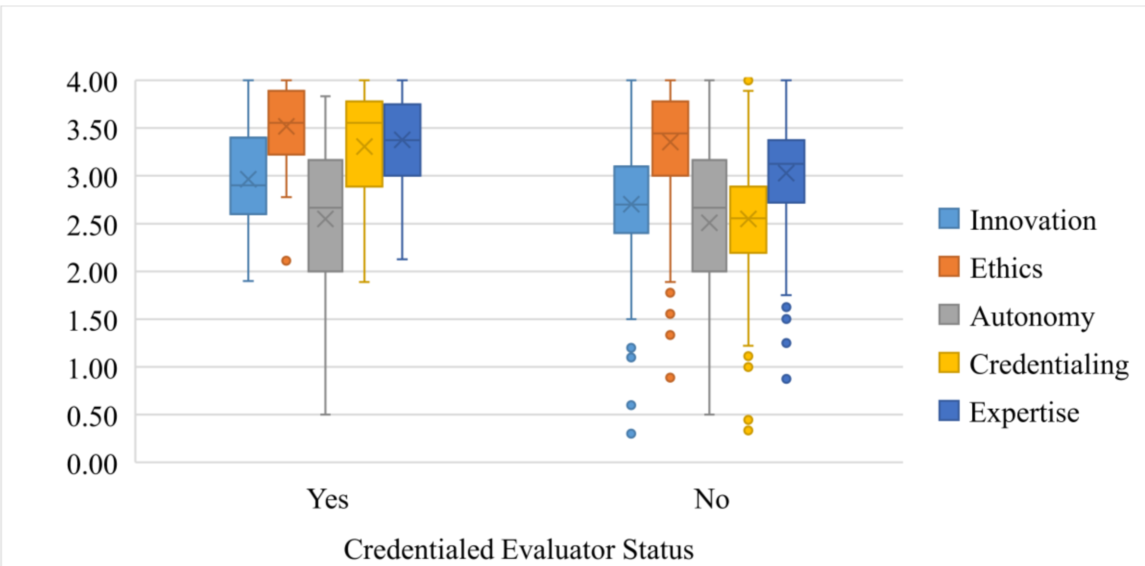

Figure 5.15. Variability in Professionalism Behaviors by Credentialed Evaluator Status

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4). The "X" in the boxplots indicates the mean scores for each sub-scale.

Figure 5.16 provides the comparison of one standard deviation of means for CEs and non-CEs to understand their differences in opinions and reported behaviors for the five sub-scales. The error bars in Figure 5.16 represent the sub-scales and the overall

Professionalism scale by CEs and non-CEs evaluators' status. The first error bar on the top level represents the overall Professionalism scale, the second bar represents the Expertise sub-scale, the third bar from the top represents the Credentialing sub-scale, the fourth bar represents the Professional Autonomy sub-scale, the fifth bar from the top represents the Ethics sub-scale, and the sixth bar represents the Innovation and Research sub-scale. This is the same order for both CEs and non-CEs.

Overall, both CEs and non-CEs had variability of opinions on the Professional Autonomy and Credentialing sub-scales although non-CEs reported more mixed feelings about credentialing, shown by the wider error bar. According to the positions of the round dots on the bars, the mean scores for Non-CEs were lower than those of CEs on the Credentialing, Innovation, and Expertise sub-scales and on overall Professionalism scale. The figure also shows that both the CEs and non-CEs had similar means score for the Professional Autonomy and Ethics sub-scales, although CEs had a slightly higher mean score for the Ethics sub-scale.

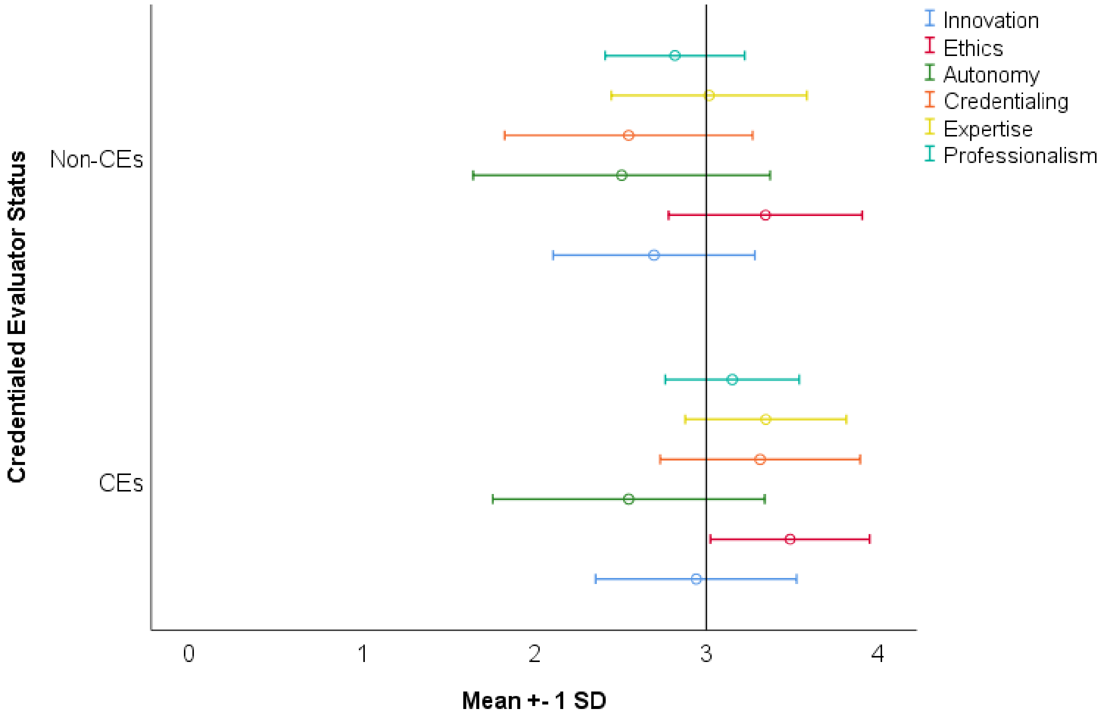

Figure 5.16. Variability in Professionalism Means by Credentialed Evaluators Status

Note. The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

I conducted a one-way MANOVA to test the hypothesis that Credentialed evaluator status was related to reported professionalism behaviors. The assumptions of independence of observations were checked and met. Box's M (34.38) test was significant (p = .005), indicating that the assumption of homogeneity of variance/covariance was not met. Although the Levene's test of equality of variances was not significant, indicating the assumption of homogeneity of variance was not violated, I used Pillai's Trace test criterion to assess the statistical significances among the credentialing status on a linear combination of the five Professionalism sub-scales.

The result of the initial one-way MANOVA was statistically significant for the five Professionalism sub-scales in relation to CE status, Pillai's $\vee$ = .156, *F* (5, 323) = 11.982, p = 001, and unusually large effect size (multivariate η^2) of .156. Power to detect the effect was 1. Thus, the hypothesis that evaluator Professionalism behaviors and their beliefs in credentialing was related to their CE status was confirmed. The multivariate η^2 = .156 indicates that approximately 16% of multivariate variance of professionalism behaviors is associated with evaluators' CE status. Examination of the coefficients for the combinations distinguishing CEs status indicated that the Innovation and Research sub-scale ($\beta = .261$, p = 006, multivariate $\eta^2 = .023$, power = .565), the Credentialing sub-scale ($\beta = .760$, p = 001, multivariate $\eta^2 = .120$, power = 1), and the Expertise sub-scale ($\beta = .348$, p = 001, multivariate $\eta^2 = .045$, power = .910) contributed the most to distinguishing evaluators' reported professionalism behaviors for CEs and non-CEs.

Given the significance of the overall multivariate test, the univariate main effects were examined, and Table 5.15 provides a summary of the results. To protect against Type I error, I used the traditional Bonferroni procedure and tested each ANOVA at the statistical significance at $\alpha < .01$ (.05 divided by the number of dependent variables or number of ANOVAs conducted). The results show that credentialed evaluator status had statistically significant univariate effects on three Professionalism sub-scales at the .01 adjusted alpha level: Innovation and Research ($p < .006$), Credentialing ($p < .001$), and Expertise ($p < .001$). The overall Professionalism scale was also significantly different by evaluator credentialed status $F(3, 338) = 4.19$, $p < .006$, partial eta square = .036, power =

.854. The CE status had no significant differences on the Professional Autonomy (p = .758) and Ethics (p = .058) sub-scales.

Table 5.15
Professionalism Mean Differences by Credentialed Evaluator Status (at $\alpha < .01$ level)

		CEs		Non-CEs					
	df Error	M	SD	M	SD	F	*Partial Eta (d)*	*p*	*Observed Power*
Innovation	327	2.97	0.57	2.70	0.58	7.587	.023	.006***	.565
Ethics	327	3.52	0.42	3.35	0.55	3.622	.011	.058	.248
Autonomy	327	2.55	0.79	2.51	0.86	.095	.000	.758	.014
Credentialing	327	3.31	0.59	2.55	0.71	44.579	.120	.001***	1.000
Expertise	327	3.38	0.42	3.03	0.56	15.505	.045	.001***	.910
Professionalism	327	3.18	.36	2.84	.39	28.003	.079	.001***	.996

Note. * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect, and .039 to .2 = desired effects (Cohen, 1988).

I conducted a non-parametric independent samples Mann-Whitney Test (t-test) at the adjusted alpha level of .01 (.01/5 = .002 adjusted alpha level) to compare the difference in means by CE status. There was a statistically significant difference between CEs and non-CEs evaluators on the Credentialing sub-scale (p <.001) with unusually large effect size (d = .123), the Expertise sub-scale (p <.001) with a medium effect size (d = .045), and the overall Professionalism scale (p <.001) with a medium effect size (d = .079), according to Cohen (1988). The mean differences and effect sizes were practically important, meaning CEs in this study were more likely than non-CEs to define themselves as evaluation professionals and have stronger beliefs in credentialing. There were no statistically significant differences in the Innovation and Research and the Ethics sub-scales for the US and Canadian evaluators at the adjusted alpha level of .002.

Evaluator Professionalism by Gender. I assessed if there was a difference in program evaluators' reported professionalism behavior by gender for each of the five sub-scales depicted in Figure 5.17. The sizes of the boxplots indicated that evaluators consistently agreed on ethics irrespective of their gender. Interestingly, both male and female evaluators endorsed ethics, expertise, and credentialing at similar levels, demonstrated by the somewhat equal median scores for these sub-scales, although male evaluators had more variable opinions on credentialing shown by the longer shape of their boxplot. The median scores for male evaluators on the Professional Autonomy and Innovation and Research sub-scales appeared higher than those of the female evaluators, and female evaluators had much variability of ratings on the Professional Autonomy sub-scale, demonstrated by the longer shape of their boxplot.

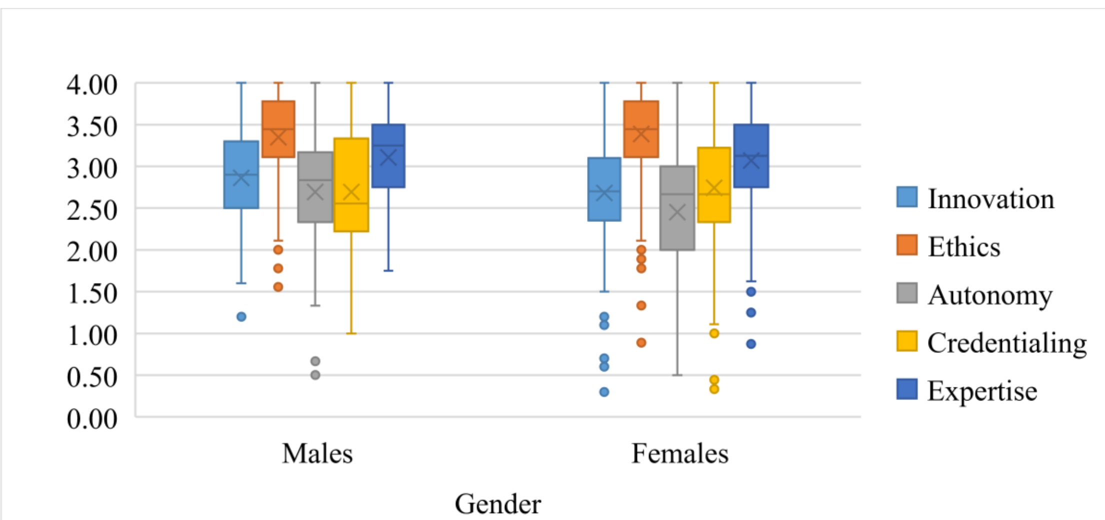

Figure 5.17. Variability in Professionalism Behavior by Gender

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4). The "X" in the boxplots indicates the mean scores for each sub-scale.

Figure 5.18 depicts a comparison of one standard deviation of means for female and male evaluators to further explore how their perceptions differed on the five sub-scales and the overall Professionalism scale. The error bars in Figure 5.18 represent the five sub-scales and the overall Professionalism scale by gender. The first error bar on the top level represents the overall Professionalism scale, the second bar represents the Expertise sub-scale, the third bar from the top represents the Credentialing sub-scale, the fourth bar represents the Professional Autonomy sub-scale, the fifth bar from the top represents the Ethics sub-scale, and the sixth bar represents the Innovation and Research sub-scale. This order is the same for both genders.

Overall, this comparison confirmed that female and male evaluators' agreement on ethics was between the "Agree" (3) and "Strongly Agree" (4) points of the scale, although they had mixed ratings on the Credentialing and Professional Autonomy sub-scales, shown by the wider error bars for each group. Male evaluators had a great deal of variability of opinions on credentialing compared to female evaluators. The figure also shows that both genders had similar means for the Expertise, Professional Autonomy, Ethics, and Credentialing sub-scales, and the overall Professionalism scale. However, the figure shows that male evaluators had somewhat higher means score on the Innovation and Research sub-scale than the female evaluators.

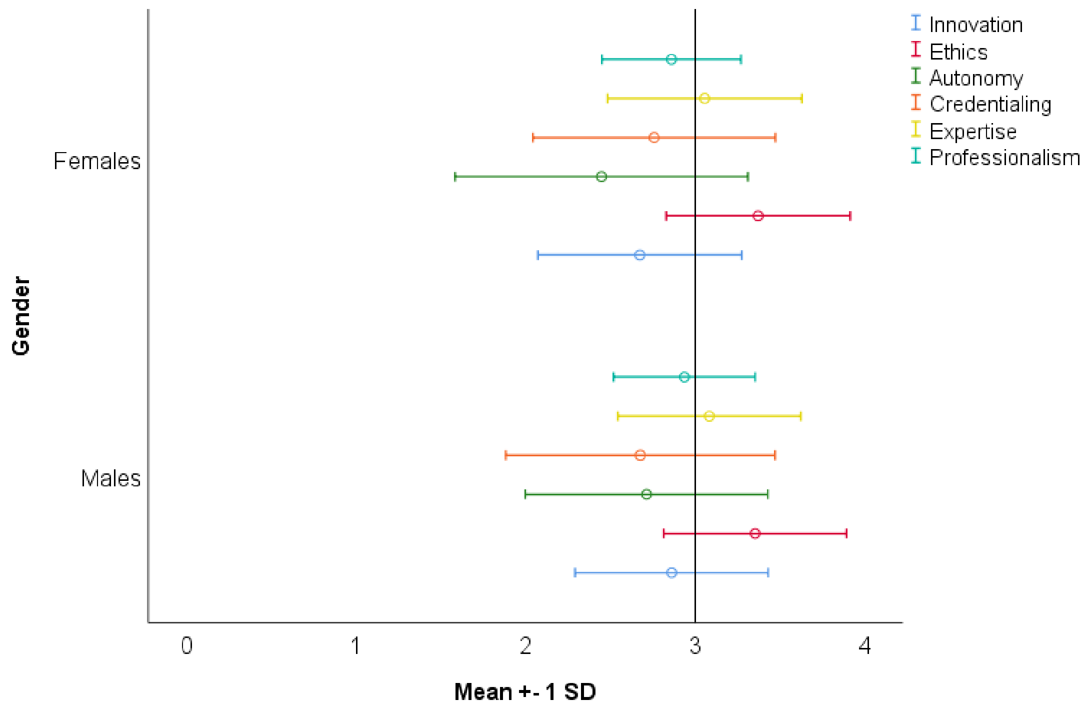

Figure 5.18. Variability in Professionalism Means by Gender

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

I conducted a one-way MANOVA to test the hypothesis that gender had a significant impact on evaluator professionalism behaviors. Bivariate scatterplots were checked for multivariate normality. The assumptions of independence of observations and homogeneity of variance/covariance were checked and met. In fact, Box's M (20.33) test was non-significant (p = .175), indicating that the assumption of homogeneity of variance/covariance was met. As such, I used Wilk's Lambda test criterion to assess the statistical significances of gender on a linear combination of the five professionalism sub-scales since the assumption of homogeneity of variance/covariance was not violated.

Overall, the one-way MANOVA revealed a significant difference in evaluator professionalism behavior based on the evaluators' gender, Wilk's Lambda ∧ = .962, *F* (5, 364) = 2.849, $p < .015$, a small effect size (partial eta squared) of .038. Power to detect the effect was .649. Thus, the hypothesis that evaluator professionalism behaviors and attitudes to credentialing was related by their gender was confirmed. The multivariate η^2 = .038 indicates that approximately 4% of multivariate variance of Professionalism behaviors is associated with evaluator's gender. Examination of the coefficients for the combinations distinguishing gender indicated that innovation and research ($\beta = .184$, *p* = 008, multivariate $\eta^2 = .019$, power = .760) and professional autonomy ($\beta = .244$, *p* = 011, multivariate $\eta^2 = .017$, power = .718) contributed the most to distinguishing evaluator's professionalism behaviors for female and male evaluators.

Given the significance of the overall multivariate test, the univariate main effects were examined, and the result is shown in Table 5.16. I used .01 adjusted alpha level (.05 divided by the number of dependent variables or number of ANOVAs conducted) to avoid Type 1 Error. The univariate ANOVAs indicated that innovation and research and professional autonomy, when examined alone, were statistically significantly different for female and male evaluators. The table shows that female and male evaluators did not show any statistically significant difference in their reported expertise ($p = .549$), belief in credentialing ($p = .561$), ethics ($p = .553$), and overall professionalism behaviors (p = .144). This was evident by the large p-values and very negligible and small effect sizes, indicating no effect in the difference in means.

Table 5.16
Professionalism Mean Differences by Gender (at p < .01 level)

		Males		Females					
	df Error	M	SD	M	SD	F	*Partial Eta (d)*	*p*	*Observed Power*
Innovation	368	2.863	.570	2.680	.593	7.153	.019	.008***	.760
Ethics	368	3.349	.539	3.385	.524	.353	.001	.553	.091
Autonomy	368	2.695	.709	2.451	.859	6.466	.017	.011**	.718
Credentialing	368	2.692	.579	2.741	.708	.339	.001	.561	.089
Expertise	368	3.106	.500	3.068	.569	.360	.001	.549	.092
Professionalism	368	2.953	.399	2.886	.392	2.148	.006	.144	.309

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect, and .039 to .2 = desired effects (Cohen, 1988).

I conducted a non-parametric independent samples Mann-Whitney Test (t-test) at the adjusted alpha level of .01 (.01/5 = .002 adjusted alpha level) to compare the difference in means by gender. There was a statistically significant difference between male and female evaluators on the Professional Autonomy sub-scale (p <.009) with a small effect size (η^2 = .019) and the Innovation and Research sub-scale (p <.005) with a small effect size (η^2 = .017), according to Cohen (1988). Male evaluators reported higher levels of professional autonomy and innovation and research behaviors compared to female evaluators, but the difference in means was not practically important given the relatively small effect sizes.

Evaluator Professionalism by Skill Level. Figure 5.19 depicts program evaluators' reported professionalism behaviors on the Expertise, Innovation and Research, Ethics, Professional Autonomy sub-scales and their belief in credentialing by their skill levels. Here, I sought to understand how evaluators' professionalism behaviors were affected by their level of skills in evaluation practice. First, as shown in the Figure, evaluator professionalism behaviors appear to increase with seniority in practice, particularly in the Professional Autonomy, Innovation and Research, Ethics, and Expertise sub-scales. Evaluators who self-identified as experts tended to endorse these professionalism behaviors at the higher levels of the scale shown by their elevated boxplots and medians scores for these sub-scales.

Aside from ethics, which was endorsed at a relatively high level of the four-point scale by evaluators in all three skill levels shown by the relatively elevated position of the Ethics boxplot compared to other boxplots, expert evaluators in the study showed much variability of ratings on credentialing shown by their longer boxplot. Interestingly, belief in Credentialing appeared to be inversely associated with seniority in evaluation skills, whereas professional autonomy behavior appeared somewhat positively related to evaluators' level of skill.

While novice evaluators were more likely to report believing in credentialing than the other evaluator groups, expert evaluators had more agreement and endorsement of the Professional Autonomy sub-scale items at the higher levels of the four-scale than the other evaluator skill levels. Novice evaluators had more variability of opinions on professional autonomy than the senior evaluators – indicating professional autonomy may come with advancement in skills levels. Furthermore, novice and proficient evaluators appeared to have somewhat similar medians for the Credentialing sub-scale compared to the more skilled expert evaluators.

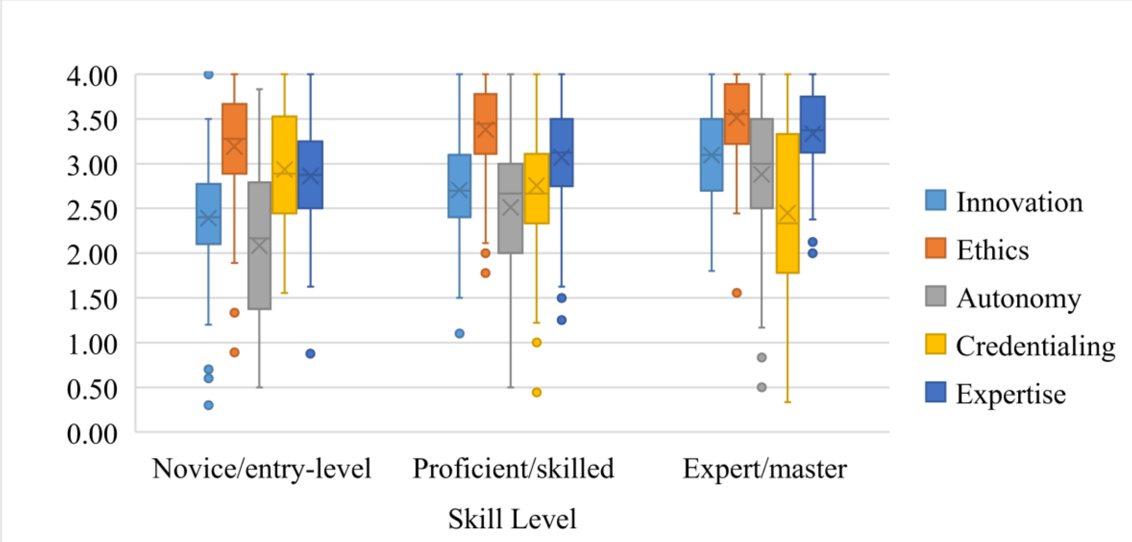

Figure 5.19. Variability in Professionalism Behavior by Skill Level

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4). The “X” in the boxplots indicates the mean scores for each sub-scale.

To understand the variability of the responses by skill level, I plotted error bars of one standard deviation of the means depicted in Figure 5.20. The error bars represent the five professionalism sub-scales by skill level. The first error bar on the top level represents the Expertise sub-scale, the second bar represents the Credentialing sub-scale, the third error bar from the top represents the Professional Autonomy sub-scale, the fourth bar represents the Ethics sub-scale, and the fifth bar from the top represents the Innovation and Research sub-scale. This order is the same for all three levels of skill levels.

Although evaluators of all skill levels demonstrated consistent agreement on ethics evidenced by the shorter error bars, ethics was more highly endorsed by expert evaluators compared to the other groups. This figure also confirms that while expert evaluators had much variability of ratings on credentialing, novice evaluators also had mixed feelings on professional autonomy, shown by the wider error bars. Evaluators who self-identified as novice or entry-level professionals had the highest means scores in the Credentialing sub-scale. The figure also confirms that evaluator professionalism behaviors in professional autonomy, expertise, innovation and research, and ethics appear to be gained reportedly with seniority in skill level. Novice evaluators had lower mean scores on these sub-scales compared to senior evaluators in the proficient and expert category.

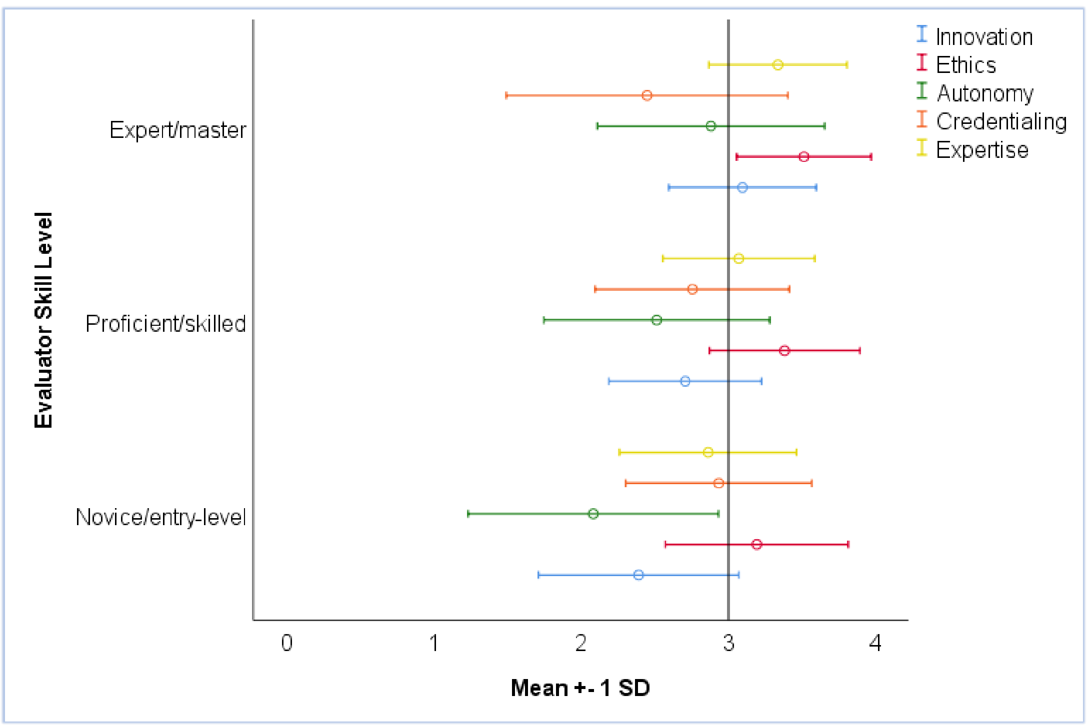

Figure 5.20. Variability in Professionalism Means by Skill Level

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on “agree” and the round dots represent the means.

I conducted a one-way MANOVA to test the hypothesis that an evaluator's skill level has a relationship with reported evaluator professionalism behaviors. The assumption of independence of observations was checked and met. Box's M (66.47) test was statistically significant (*p* = .001), thus the assumption of homogeneity of variance/covariance was not met. The assumption of equal variance was not met since the Levene's test was significant for the Credentialing sub-scale (p = .001). I used Pillai's trace criterion (a test statistic that is very robust and not highly linked to assumptions about the normality of the distribution of the data) to assess the statistical significances among the evaluators' skill level on a linear combination of the five professionalism sub-scales since the assumption of homogeneity of variance/covariance was violated.

Overall, the result of the one-way MANOVA showed a statistically significant multivariate effect for the five professionalism sub-scales as a group in relation to evaluators' skill level, Pillai's trace $\wedge$ = .231, F (10, 736) = 9.592, $p < .001$, and a large effect size (partial eta squared) of .115, according to Cohen (1988). Power to detect the effect was 1. Thus, the hypothesis that evaluator professionalism behaviors and their beliefs in credentialing was associated with their levels of skill in evaluation practice was confirmed. The multivariate η^2 = .115 indicated that approximately 12% of multivariate variance of professionalism behaviors was associated with evaluators' skill levels.

Given the significance of the overall multivariate test, the univariate main effects were examined, and Table 5.17 provides a summary of the results. To protect against Type I error, I used the Bonferroni procedure and tested each of the five ANOVAs at the .01 level (.05 divided by five dependent variables). The univariate ANOVA results showed that evaluator level of skills had statistically significant univariate effects on all five professionalism sub-scales and the overall Professionalism sub-scale at the adjusted alpha level of .01: Innovation and Research $F(2, 371) = 28.840$, $p < .001$, large effect size (partial eta square = .135, power = 1); Ethics $F(2, 371) = 6.343$, $p < .002$, small effect size (partial eta square = .033, power = .745); Professional Autonomy $F(2, 371) = 17.464$, $p < .001$, medium effect size (partial eta square = .086, power = .998); Credentialing $F(2, 371) = 8.202$, $p < .001$, small effect size (partial eta square = .042, power = .872); Expertise $F(2, 371) = 14.271$, $p < .001$, medium effect size (partial eta square = .071, power = .992), and overall Professionalism $F(2, 371) = 13.157$, $p < .001$, medium effect size (partial eta square = .066, power = .986).

As shown in Table 5.17, self-reported expert evaluators had higher means on the Innovation and Research ($m = 3.170$), Professional Autonomy ($m = 2.882$), Expertise ($m = 3.297$) and Ethics ($m = 3.487$) sub-scales on average than the other evaluator groups. Self-reported novice valuators had higher means on the Credentialing sub-scale ($m = 2.935$) on average than the other evaluator groups.

Table 5.17
Univariate Mean Differences by Evaluator Skill Level (at p < .01)

	Error df	Novices M	Novices SD	Proficient M	Proficient SD	Experts M	Experts SD	F	*Partial Eta Squared*	*p*	*Observed Power*
Innovation	371	2.39	0.68	2.71	0.52	3.10	0.50	28.840	.135	.000***	1.000
Ethics	371	3.19	0.62	3.38	0.51	3.51	0.46	6.343	.033	.002***	.745
Autonomy	371	2.08	0.85	2.51	0.77	2.88	0.77	17.464	.086	.000***	.998
Credentialing	371	2.94	0.63	2.76	0.66	2.45	0.96	8.202	.042	.000***	.872
Expertise	371	2.86	0.60	3.07	0.52	3.34	0.47	14.271	.071	.000***	.992
Professionalism	371	2.73	0.44	2.90	0.36	3.06	0.38	13.157	.066	.001***	.986

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect, and .039 to .2 = desired effects (Cohen, 1988).

I conducted a post hoc Games-Howell non-parametric test to understand which pairs of means were significantly different for the three levels of evaluator skill. I used the Games-Howell test because the assumption of homogeneity of variance-covariance was violated. I had previously controlled for Type I error across the five univariate ANOVAs by testing each at the .01 alpha level using the Bonferroni method. To be consistent with this decision, I tested each comparison at the .01 adjusted alpha level for the ANOVAs divided by the five pairwise comparisons for the dependent variables (.01/5 = .002 adjusted alpha level).

Significant skill level pairwise differences were obtained in all five sub-scales and in the overall Professionalism scale at the adjusted alpha level of $p < .002$. The pairwise comparison showed that self-reported expert evaluators' innovation and research behaviors were statistically different than the proficient evaluators ($p = .003$) and novice evaluators ($p = .001$). There were also significant differences between proficient evaluators and novice evaluators on the Innovation and Research sub-scale ($p = .003$). On the Ethics sub-scale, expert evaluators reported higher levels of ethical behaviors on average than novice evaluators ($p = .003$). Pairwise comparisons showed no significant differences between expert and proficient evaluators on the Ethics sub-scale ($p = .095$) and proficient and novice evaluators ($p = .079$) on this sub-scale. On the Expertise sub-scale, expert evaluators reported higher levels of expertise behaviors on average than novice evaluators ($p = .001$) and proficient evaluators ($p = .001$). Pairwise comparisons showed no significant differences between proficient evaluators and novice evaluators on the Expertise sub-scale ($p = .041$).

Significant evaluation skill level pairwise differences were also obtained in the Professional Autonomy sub-scale for all evaluator groups, indicating that autonomy was reportedly gained with the level of skill since expert evaluators reported higher means on the Professional Autonomy sub-scale, followed by proficient evaluators, and lastly novice evaluators. For the Credentialing sub-scale, pairwise comparisons revealed that novice evaluators had significantly greater belief in credentialing than expert evaluators (p = .002), but there was no statistically significant difference between novice evaluators and proficient evaluators (p = .130) and between expert evaluators and proficient evaluators (p = .029) on belief in credentialing. Finally, both expert and proficient evaluators reported a higher overall professionalism score than novice evaluators. There was a statistically significant difference between expert and novice evaluators (p = .001) and between expert evaluators and proficient evaluators (p = .006) on overall professionalism behavior. Pairwise comparisons showed no significant differences between proficient evaluators and novice evaluators on the overall Professionalism scale (p = .013)

Overall, expert evaluators were more likely than proficient and novice evaluators to report having professionalism behaviors in professional autonomy, expertise, innovation and research, and ethics while novice evaluators were more likely than expert and proficient evaluators to report belief in evaluators being credentialed.

Evaluator Professionalism by Levels of Education. Figure 5.21 depicts program evaluators' reported professionalism behaviors on expertise, innovation, ethics, autonomy and their beliefs in credentialing by levels of education. Here, I sought to understand how the evaluator's level of education had a relationship with their professionalism behaviors. First, the figure shows that all evaluator education groups in the study endorsed ethics at the higher levels of the scale on “Agree” (3) and “Strongly Agree” (4) points on the scale as shown by the sizes of the elevated ethics boxplots. Interestingly, evaluators with bachelor’s degrees had the most agreement on ethics, but also had much variability of opinions on professional autonomy compared to those with master’s and doctoral degrees.

Unlike evaluators with bachelor’s and doctoral degrees who endorsed credentialing at lower levels of the scale, evaluators with master’s degrees showed higher levels of agreement on credentialing based on their median and mean scores. Furthermore, evaluators with doctoral degrees were more likely than those with bachelor’s and master’s degrees to endorse innovation and research and expertise items more highly, evidenced by their higher median and mean scores. Finally, it appears that reported expertise, autonomy, and innovation behaviors increased with seniority in education. In fact, evaluators with doctoral degrees endorsed these professional behaviors at higher levels of the scale than those with bachelor’s and master’s degrees.

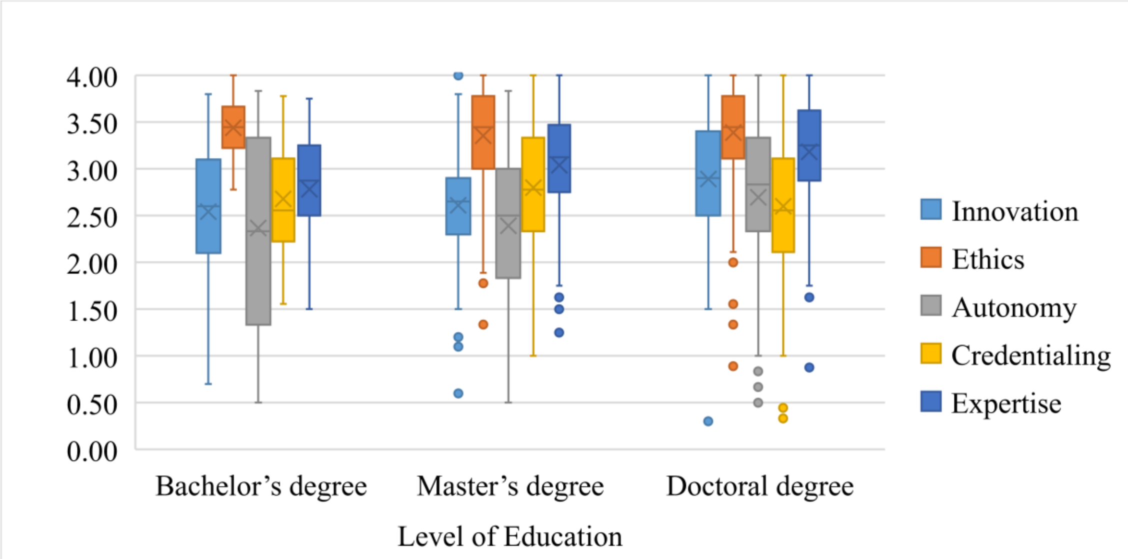

Figure 5.21. Variability in Professionalism Behavior by Level of Education

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4). The "X" in the boxplots indicates the mean scores for each sub-scale.

To understand the variability of these responses, I plotted error bars of one standard deviation of the means depicted in Figure 5.22. The five error bars represent the five professionalism sub-scales by education level. The first error bar on the top level represents the Expertise sub-scale, the second bar represents the Credentialing sub-scale, the third error bar from the top represents the Professional Autonomy sub-scale, the fourth bar represents the Ethics sub-scale, and the fifth bar from the top represents the Innovation and Research sub-scale. This order is the same for all three levels of education.

Figure 5.22 confirms that while evaluators with doctoral degrees expressed much variability of opinions on credentialing, those with bachelor's degrees had much variability of opinions on professional autonomy. Furthermore, it appears that reported professionalism behaviors in innovation and research, expertise, and professional autonomy increased with seniority in educational levels. Interestingly, all evaluator educational groups endorsed ethics on the “Agree” (3) to “Strongly agree” (4) level of the scale, indicating that the level of education did not seem to negatively affect evaluators’ reported ethical behaviors.

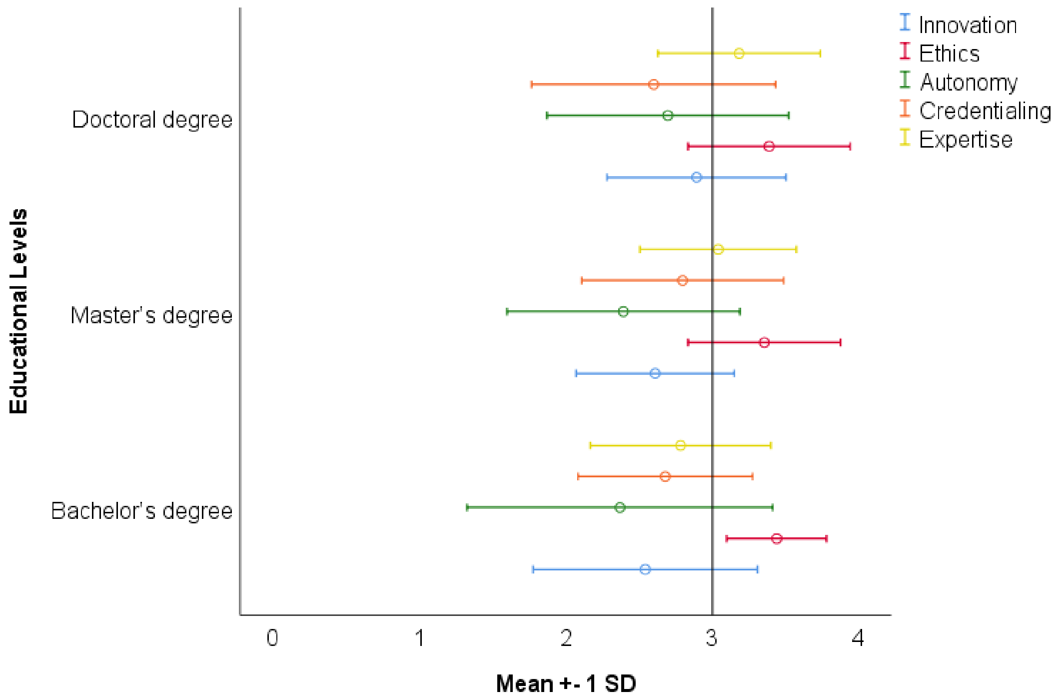

Figure 5.22. Variability in Professionalism by Educational Levels

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on “agree” and the round dots represent the means.

I conducted a one-way MANOVA to test the hypothesis that evaluators’ levels of education (bachelor’s, master’s, and doctoral degrees) had a relationship with their professionalism behaviors. The assumption of independence of observations was checked and met, but the assumption of homogeneity of variance/covariance was violated. Box's

M (59.618) test was statistically significant (p = .003), although Levene's test of equality of error variances was not significant. I used Pillai's trace criterion to assess the statistical significances among the evaluators' levels of education on a linear combination of the five professionalism sub-scales and Games-Howell non-parametric post hoc tests because of the violation of the assumption of homogeneity of variances/covariance.

Overall, the result of MANOVA was statistically significant, Pillai's $\vee$ = .111, F (10, 708) = 4.145, p < .001, medium effect size (partial eta squared) of .055, according to Cohen (1988). Power to detect the effect was .999. Thus, the hypothesis that evaluator professionalism behaviors and their attitudes on credentialing had a relationship with their levels of education was confirmed. The multivariate η^2 = .055 indicates that approximately 6% of multivariate variance of professionalism behaviors was associated with evaluators' level of education.

Given the significance of the overall test, the univariate main effects were examined, and the result of the one-way MANOVA is shown in Table 5.18. To protect against Type I error, I used the Bonferroni procedure and tested each of the five ANOVAs at the .01 level (.05 divided by the number of dependent variables). The follow-up univariate ANOVA results showed that the Innovation and Research (p < .001), Professional Autonomy (p < .003); and Expertise (p < .004) sub-scales were statistically significant at the adjusted alpha level of .01. The comparisons showed no significant results for the Ethics (p = .718) and the Credentialing (p = .056) sub-scales, and the overall Professionalism scale (p < .032).

The univariate testing indicated that evaluators with postgraduate training (master's and doctoral degrees) were more likely than those with bachelor's degrees to report higher levels of professional autonomy, expertise, and innovation behaviors. Evaluators with doctoral degrees reported higher levels of professional autonomy (m = 2.70), innovation and research (m = 2.89), expertise (m = 3.18), and overall professionalism (m = 2.96) on average than those with bachelor's and master's degrees. Evaluators with master's degrees reported more preference for being credentialed (m = 2.80) on average compared to those with bachelor's and doctoral degrees. Interestingly, all evaluator education groups reported similar means for the Ethics sub-scale.

I conducted a post hoc test to understand which pairs of means were significantly different for the Professional Autonomy, Innovation and Research, Expertise, Credentialing, and Ethics sub-scales, and the overall Professionalism scale for the three levels of evaluator education. To protect against Type I error, I maintained the .01 alpha level across comparisons for the five sub-scales to be consistent with the decision for testing the five ANOVAs at the .01 level. With the Bonferroni method, I tested each comparison at the alpha level for the ANOVA divided by the number of comparisons (.01/5 = .002).

Table 5.18
Univariate Mean Differences by Evaluator Education per Sub-scale (at p < .01)

		Bachelor's Degree		Master's Degree		Doctoral Degrees					
	df	M	SD	M	SD	M	SD	F	*d (Eta Squared)*	*p*	*Observed Power*
Innovation	357	2.54	0.77	2.61	0.54	2.89	0.61	10.439	.055	.000***	.988
Ethics	357	3.44	0.34	3.35	0.52	3.39	0.55	.332	.002	.718	.103
Autonomy	357	2.37	1.04	2.39	0.79	2.70	0.83	5.800	.031	.003***	.869
Credentialing	357	2.68	0.60	2.80	0.69	2.60	0.83	2.901	.016	.056	.565
Expertise	357	2.78	0.61	3.04	0.53	3.18	0.55	5.735	.031	.004***	.865
Professionalism	357	2.78	0.33	2.86	0.38	2.96	0.43	3.473	.019	.032	.648

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect, and .039 to .2 = desired effects (Cohen, 1988).

Significant education level pairwise differences were obtained in two Professionalism sub-scales at the adjusted alpha level of $p < .002$. The pairwise comparison showed that evaluators with doctoral degrees were statistically significantly different on the Innovation and Research sub-scale than those with masters ($p = .001$). There were also significant differences between evaluators with doctoral degrees and those with master's degrees on the Professional Autonomy sub-scale ($p = .003$). The post hoc comparisons indicated no statistically significant results on the Ethics, Credentialing, and Expertise sub-scales, and the overall Professionalism scale by level of education. Overall, evaluators with doctoral degrees were more likely than those with master's or bachelor's degrees to report having professionalism behaviors in professional autonomy and innovation and research.

Evaluator Professionalism by Employment Status. Figure 5.23 shows the study participants' reported professionalism behaviors on expertise, innovation, ethics, autonomy and beliefs in credentialing by employment status. Here, I sought to understand how practicing evaluation part-time or full-time might influence program evaluators' perceptions and demonstration of professionalism behaviors. First, as shown in Figure 5.23, the sizes of the boxplots indicated that evaluators working part-time and full-time had similarity of opinions on innovation and research, ethics, and credentialing, but they had somewhat variable opinions on professional autonomy and expertise. Program evaluators who worked full-time endorsed items on the Expertise sub-scale at the higher values of "Agree" (3) and "Strongly Agree (4) points on the scale. Both groups of evaluators showed variability of opinions on professional autonomy and credentialing.

However, both groups endorsed items on the Ethics sub-scale the most, followed by the Expertise sub-scale. There seemed to be a difference in the mean scores by employment status for the Expertise and the Autonomy sub-scales based on the somewhat different median scores.

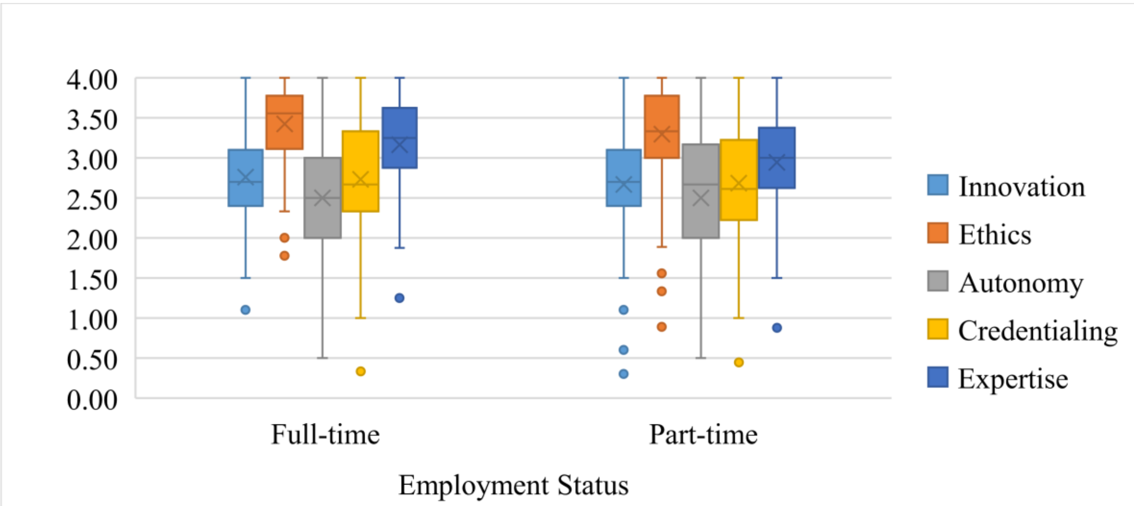

Figure 5.23. Variability in Professionalism Behavior by Employment Status

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4).

To understand the participants' responses by employment status, I plotted error bars of one standard deviation of the means depicted in Figure 5.24. The five error bars represent the five professionalism sub-scales by employment status. The first error bar on the top level represents the Expertise sub-scale, the second bar represents the Credentialing sub-scale, the third error bar from the top represents the Professional Autonomy sub-scale, the fourth bar represents the Ethics sub-scale, and the fifth bar from the top represents the Innovation and Research sub-scale. This order is the same for all three levels of education.

This figure confirms the participants' variability of opinions on the Credentialing and the Professional Autonomy sub-scales, indicated by the longer error bars for both groups. Furthermore, evaluators who worked part-time and full-time endorsed ethical behaviors at higher mean scores than the other professionalism sub-scales. Aside from the higher Expertise sub-scale mean score for evaluators who worked full-time, both groups had similar mean scores for the Credentialing, and the Professional Autonomy sub-scales. However, the mean scores for the Innovation and Research sub-scale was slightly higher for evaluators who worked full-time compared to those who worked part-time.

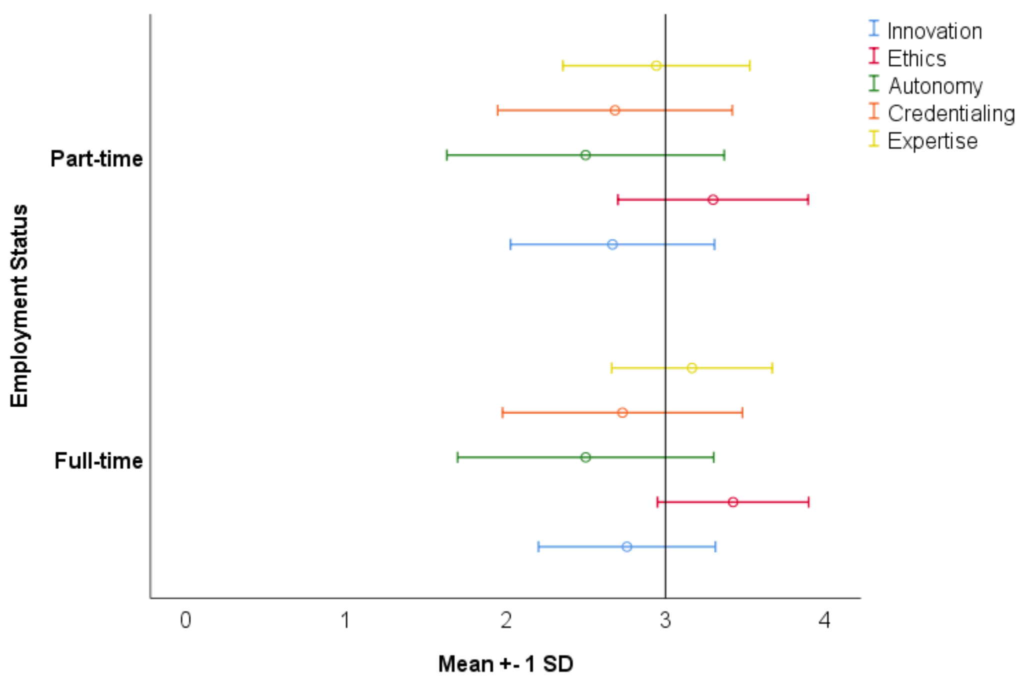

Figure 5.24. Variability in Professionalism by Employment Status

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on "agree" and the round dots represent the means.

I conducted a one-way MANOVA to assess if the mean differences for employment status on a linear combination of the professionalism sub-scales were statistically significant. The assumption of independence of observations was checked and met. Although Box's M (31.07) test was significant (p = .010), thus the assumption of homogeneity of variance/covariance was not met, the Levene's test of equality of variance was significant for the Ethics sub-scale (p = .042). As such, I used Pillai's Trace criterion (a test statistic that is very robust and not highly linked to assumptions about the normality of the distribution of the data) to assess the statistical significances among the employment status linear combination of the five professionalism sub-scales.

Overall, the result of the one-way MANOVA was statistically significant, Pillai's $\vee$ = .046, F (5, 374) = 3.637, p < .003, partial eta squared = .046. Power to detect the effect was .926. Thus, the hypothesis that evaluator professionalism behaviors and their attitudes to credentialing was associated with working full-time or part-time was confirmed. The multivariate η^2 = .046 indicates that approximately 5% of multivariate variance of professionalism behaviors was associated with evaluator employment status.

Given the significance of the overall multivariate test, the univariate main effects were examined, and the result of the one-way MANOVA is shown in Table 5.19. I used a .01 adjusted alpha level (.05 divided by the number of dependent variables or number of ANOVAs conducted) to avoid Type 1 Error. Significant univariate effects were found only on the Expertise (p < .001) sub-scale. Examination of the coefficients for the combinations distinguishing professionalism behavior by employment status indicated that the Expertise sub-scale ($\beta = .223$, *p* = 001) contributed the most to distinguishing evaluator's professionalism behaviors for evaluators working full-time and part-time. As shown in Table 5.19, evaluators working full-time reported a higher mean score for the Expertise sub-scale (*m* = 3.1654) and the overall Professionalism scale (*m* = 2.936).

The table shows that evaluators working full-time and part-time did not show any statistically significant difference in their reported innovation and research, ethics, professional autonomy, belief in credentialing, and overall professionalism behaviors. This was evident by the large p-values and very negligible and small effect sizes, indicating no effect in the difference in means.

Table 5.19
Significant Univariate Effects by Seniority in Years of Employment Status (at $\alpha < .01$ level)

		Full-Time		Part-Time					
	df Error	M	SD	M	SD	F	*Partial Eta (d)*	*p*	*Observed Power*
Innovation	378	2.759	.553	2.668	.638	2.137	.006	.145	.131
Ethics	378	3.422	.473	3.297	.595	5.159	.013	.024	.377
Autonomy	378	2.501	.801	2.500	.867	.000	.000	.993	.010
Credentialing	378	2.731	.570	2.684	.734	.359	.001	.549	.025
Expertise	378	3.165	.502	2.943	.585	15.620	.040	.001***	.913
Professionalism	378	2.936	.372	2.835	.422	5.971	.016	.015	.443

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect, and .039 to .2 = desired effects (Cohen, 1988).

I conducted a non-parametric independent samples Mann-Whitney Test (t-test) to compare the difference in means by employment status for the five dependent variables. To protect against Type I error, I maintained the .01 alpha level across comparisons for the five sub-scales to be consistent with the decision for testing the five ANOVAs at the

.01 level. With the Bonferroni method, I tested each comparison at the alpha level for the ANOVA divided by the number of comparisons (.01/5 = .002). The test found a statistically significant difference between evaluators working part-time and full-time on the Expertise sub-scale (p <.001) with a medium effect size (d = .040), according to Cohen (1988). Evaluators working full-time reported higher levels of expertise on average compared to evaluators working part-time, and the difference in means was practically important. There were no statistically significant differences in innovation and research, ethics, professional autonomy, credentialing, and overall professionalism behaviors for evaluators working full-time and part-time at the adjusted alpha level of .002.

Evaluator Professionalism by Years of Evaluation Practice. Figure 5.25 depicts program evaluators' reported professionalism behaviors on the Expertise, Innovation, Ethics, and Autonomy sub-scales and their beliefs in credentialing by years of evaluation experience. Here, I sought to understand how years of evaluation experience were related to evaluator professionalism behaviors. As shown in Figure 5.25, evaluator professionalism behaviors appeared to increase with years of evaluation experience, particularly for professional autonomy, innovation and research, and expertise. Evaluators with more years of experience tended to endorse the items of these professionalism behaviors at the higher levels of the scale than those with fewer years of experience. Evaluators' reported ethical behavior also appeared to positively increase with years of evaluation experience except for the decline for evaluators with 31 or more years of evaluation experience. However, participants in the study endorsed items of the

Ethics sub-scale at the high levels of the scale between “Agree” (3) and “Disagree” (4) irrespective of how long they had practiced program evaluation.

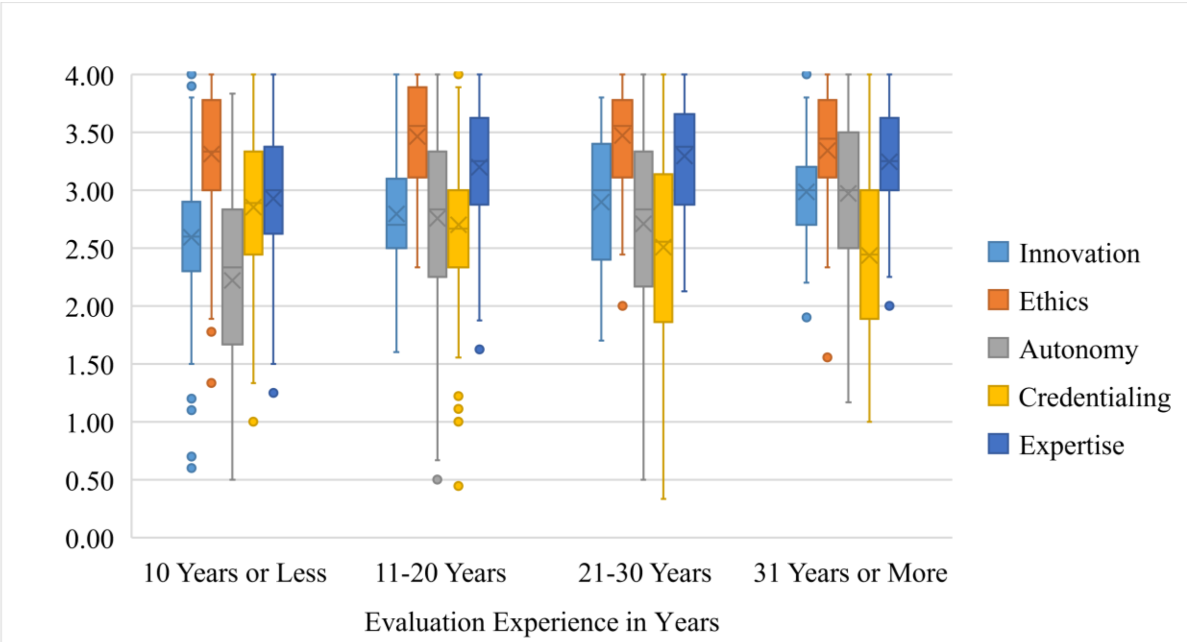

Figure 5.25. Variability in Professionalism Behavior by Years of Experience

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4). The “X” in the boxplots indicates the mean scores for each sub-scale.

Interestingly, there appears to be an inverse association between years of experience in evaluation and belief in credentialing as shown in Figures 5.25 and 5.26, especially for newer evaluators. Evaluators with 10 years or less experience in evaluation were more likely to believe that evaluators should be credentialed compared to those with over 21 years of evaluation experience. In fact, senior evaluators in the practice endorsed items on the Credentialing sub-scale between the “Disagree” (2) and “Agree” (3) level of the scale while novice evaluators endorsed the items at the “Agree” (3) or “Strongly Agree” levels of the scale. As shown in Figure 5.23, evaluators with 21 – 31 or more years of experience had much variability in opinions on credentialing than those with less than 20 years of experience.

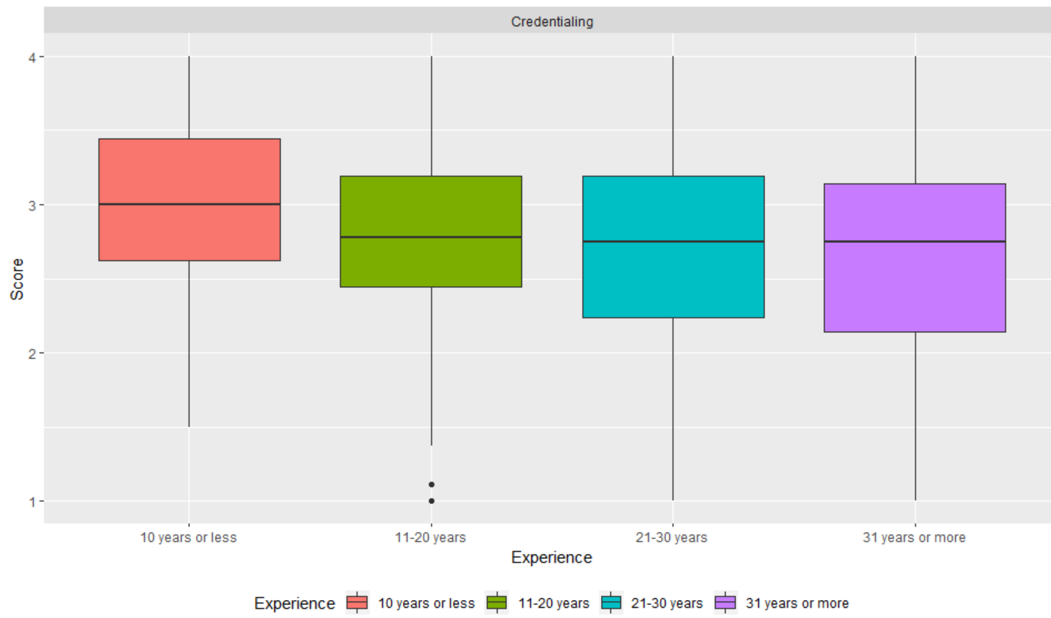

Figure 5.26. Variability in Belief in Credentialing by Seniority in Years of Experience.

Note: The rating scale used in the professionalism assessment survey was: strongly disagree (1), disagree (2), agree (3), strongly agree (4).

To understand the variability of these responses and the association between years of evaluation experience and professionalism, I plotted error bars of one standard deviation of the means depicted in Figure 5.27. The five error bars represent the five professionalism sub-scales by years of experience in evaluation. The first error bar on the top level represents the Expertise sub-scale, the second bar represents the Credentialing sub-scale, the third error bar from the top represents the Professional Autonomy sub-scale, the fourth bar represents the Ethics sub-scale, and the fifth bar from the top represents the Innovation and Research sub-scale. This order is the same for all four categories of years of experience in evaluation.

This figure confirms that evaluators' reported professionalism behaviors in the Professional Autonomy, Innovation and Research, and Expertise sub-scales seemed to generally increase with years of evaluation experience while their belief in credentialing decreased by years of evaluation experience. The wider error bars show much variability of ratings on credentialing, especially by senior evaluators with more than 21 years of evaluation experience compared to more agreement by novice evaluators on credentialing shown by their shorter error bar. The relationship is different for the Professional Autonomy sub-scale. Novice evaluators with 20 years or less experience had much variability on professional autonomy. Although all evaluators in the sample consistently agreed and endorsed ethical behaviors at the higher levels of the scale, the Ethics sub-scale means for evaluators with over 30 years of experience in evaluation was somewhat lower than those with 11 to 30 years of evaluation experience.

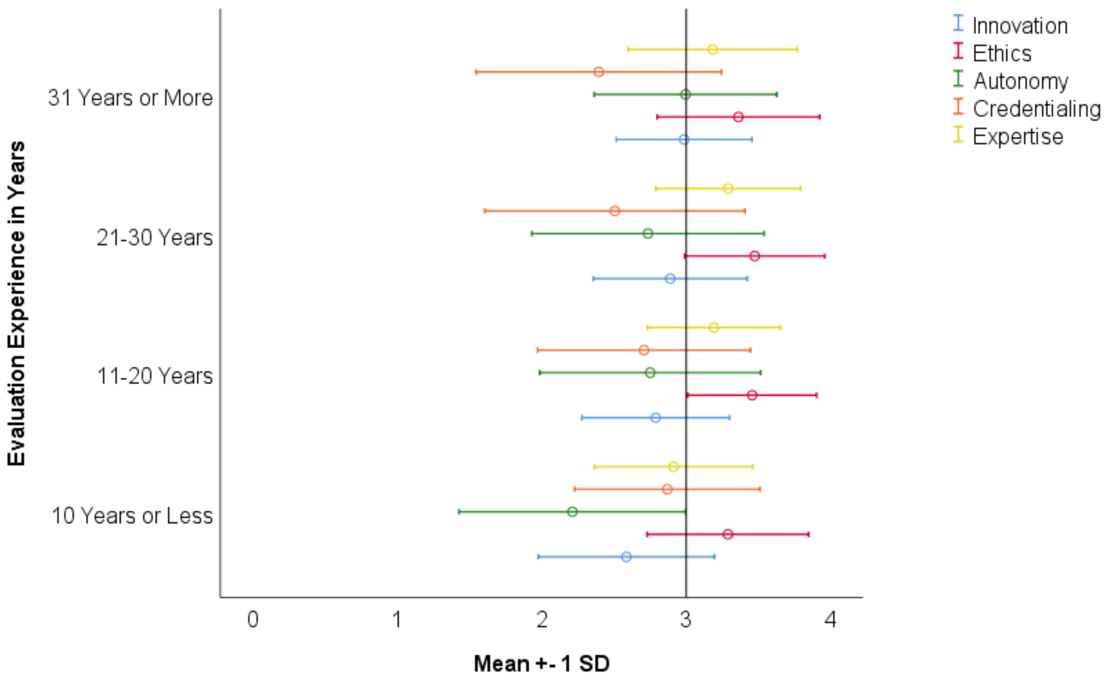

Figure 5.27. Variability in Professionalism Means by Experience in Years

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on “agree” and the round dots represent the means.

I conducted a one-way MANOVA to examine the association between evaluator professionalism behaviors and years of experience in evaluation practice. The assumption of independence of observations was checked and met. Box’s M (66.253) test of equality of covariance matrices was significant (*p* = .033), so the assumption of homogeneity of variance/covariance was not met. However, the Levene’s test of equality of variances was significant for the Credentialing sub-scale (p = .012). As such, I used Pillai’s Trace criterion to assess the statistical significance among the evaluator experience in years on a linear combination of the five professionalism sub-scales.

Overall, the result of the initial MANOVA showed a statistically significant multivariate effect for the five Professionalism sub-scales as a group in relation to evaluators years of experience in evaluation, Pillai's $V = .247$, F (15, 1125) = 6.738, $p < .001$, with a large effect size (partial eta squared) of .082, according to Cohen (1988). Power to detect the effect was 1. Thus, the hypothesis that evaluators' professionalism behaviors and their beliefs in credentialing was related to their experience in years of evaluation practice was confirmed. The multivariate $\eta^2 = .082$ indicates that approximately 8% of multivariate variance of professionalism behaviors was associated with evaluators' years of experience in evaluation.

Given the significance of the overall test, the univariate main effects were examined, and the results of the one-way MANOVA are shown in Table 5.20 at the .01 adjusted alpha level. The table shows that seniority in years of evaluation experience was statistically significant for the Innovation and Research ($p < .001$); the Professional Autonomy ($p < .001$); the Credentialing ($p < .001$), and the Expertise ($p < .001$) sub-scales. The table shows that evaluators with 31 years or more of evaluation experience reported higher means on the Innovation and Research ($m = 2.9857$) and the Professional Autonomy ($m = 2.9714$) sub-scales, but evaluators with 21-30 years of evaluation experience reported higher means on the Ethics ($m = 3.4731$) and the Expertise ($m = 3.2955$) sub-scales. Finally, evaluators with 10 years or less of evaluation experience reported higher means on the Credentialing ($m = 2.8541$) sub-scale on average than the more senior evaluators.

I conducted a post hoc test to understand which pairs of means were significantly different. To protect against Type I error, I maintained the .01 alpha level across comparisons for all five sub-scales to be consistent with the decision for testing the five ANOVAs at the .01 alpha level. With the Bonferroni method, I tested each comparison at the alpha level for the ANOVA divided by the number of comparisons (.01/5 = .002). Because the assumption of homogeneity of variance-covariance was violated, I used Games-Howell post hoc criterion.

Significant years of experience pairwise differences were obtained in three sub-scales and in the overall Professionalism scale at the adjusted alpha level of $p < .002$. There was statistically significant difference on the Innovation and Research sub-scale between evaluators with 21-30 years and novice evaluators with 10 years or less experience ($p = .001$), evaluators 31 years or more experience and novice evaluators with 10 years or less experience ($p = .001$), but no significant difference between evaluators 11-20 years and novice evaluators with 10 years or less experience ($p = .017$). These pairwise comparisons showed that evaluators with 21 years or more years of experience were more likely than novice evaluators with 10 years or less years of experience to report innovation and research behaviors, suggesting that innovation and research behaviors was gained with years of evaluation experience.

Table 5.20
Univariate Mean Differences by Years of Evaluation Experience (at p < .01 level)

		10 Years or Less		11-20 Years		21-30 Years		31 Years or More					
	df	M	SD	M	SD	M	SD	M	SD	F	*d (Eta Squared)*	*p*	*Observed Power*
Innovation	377	2.59	0.60	2.80	0.51	2.90	0.54	2.99	0.47	8.534	.064	.001***	.971
Ethics	377	3.31	0.54	3.46	0.45	3.47	0.48	3.34	0.56	2.809	.022	.039	.436
Autonomy	377	2.22	0.78	2.76	0.77	2.71	0.80	2.97	0.62	17.547	.123	.010***	1.000
Credentialing	377	2.85	0.63	2.70	0.74	2.51	0.90	2.43	0.83	5.682	.043	.001***	.842
Expertise	377	2.93	0.54	3.20	0.46	3.30	0.50	3.25	0.46	12.310	.089	.001***	.998
Professionalism	377	2.81	0.37	2.99	0.36	2.99	0.41	2.99	0.40	6.852	.052	.001***	.977

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect, and .039 to .2 = desired effects (Cohen, 1988).

Statistically significant pairwise differences were obtained in the Professional Autonomy sub-scale for all evaluator groups with 11 years or more experience than novice evaluators with 10 years or less experience, suggesting that professional autonomy is gained with years of evaluation experience. There was statistically significant differences on the Professional Autonomy sub-scale between evaluators with 11-20 years of experience and novice evaluators with 10 years or less experience (p = .001), between those with 21-30 years of experience and novice evaluators with 10 years or less experience (p = .001), and between evaluators with 31 years or more experience and novice evaluators with 10 years or less experience (p = .001). There were no statistically significant pairwise differences on the Professional Autonomy sub-scale between evaluators with 21-30 years of experience and those with 11-20 years of experience (p = .983), evaluators with 31 years or more and those with 11-20 years of experience (p = .355), and between evaluators with 31 years or more and those with 21-30 years of experience (p = .274).

Statistically significant pairwise differences were obtained in the Expertise sub-scale for all evaluator groups with 11 years or more experience than novice evaluators with 10 years or less experience, suggesting that evaluators reportedly gained expertise with years of evaluation experience. The pairwise comparisons showed statistically significant differences on the Expertise sub-scale between evaluators with 11-20 years of experience and novice evaluators with 10 years or less experience (p = .001), between those with 21-30 years of experience and novice evaluators with 10 years or less experience (p = .001), and between evaluators with 31 years or more experience and novice evaluators with 10 years or less experience (*p* = .003). There were no statistically significant pairwise differences on the Expertise sub-scale between evaluators with 21-30 years of experience and those with 11-20 years of experience (*p* = .584), evaluators with 31 years or more and those with 11-20 years of experience (*p* = .949), and between evaluators with 31 years or more and those with 21-30 years of experience (*p* = .960).

Statistically significant differences were found in pairwise comparisons by evaluator years of experience on the overall Professionalism scale only between evaluators with 11-20 years of experience and novice evaluators with 10 years or less experience (*p* = .001). There were no statistically significantly differences in overall reported professionalism behavior between all other groups of evaluators by years of experience. Pairwise comparisons showed no statistically significant differences in the Ethics and Credentialing sub-scales by years of experience in evaluation at the adjusted alpha level of .002.

Pairwise comparisons at the adjusted alpha level of $\alpha < .002$ showed that novice evaluators with 10 years or less evaluation experience reported significantly fewer Professionalism behaviors in the Innovation and Research, Expertise, and Professional Autonomy sub-scales compared to those with over 11 years of evaluation experience based on the difference in means.

Evaluator Professionalism by Level of Seniority in Age. Finally, I assessed how program evaluators' professionalism behaviors were associated with their age. Figure 5.28 depicts a generational difference in evaluator professionalism by age between the older and the younger evaluators. First, the figure suggests that evaluators reportedly acquire professional autonomy, innovation and research, and expertise behaviors with seniority in age, shown by the slightly upward linear increase in median and mean scores from the youngest to the oldest evaluators.

The boxplots indicate that older evaluators reported more professional autonomy, innovation and research, and expertise professional traits than younger evaluators, shown by their endorsement of the items in these sub-scales at higher levels on the scale. However, the evaluators' reported ethical behavior seemed to fluctuate with age, with those in the 55 – 64 age group more likely to strongly endorse the items on the Ethics sub-scale shown by their elevated boxplots. Nevertheless, the elevated boxplots suggest that evaluators of all ages endorsed the items on the Ethics sub-scale. In addition, there seemed to be much variability of opinions on professional autonomy among evaluators in the 34 – 44 and 55 – 64 years age groups.

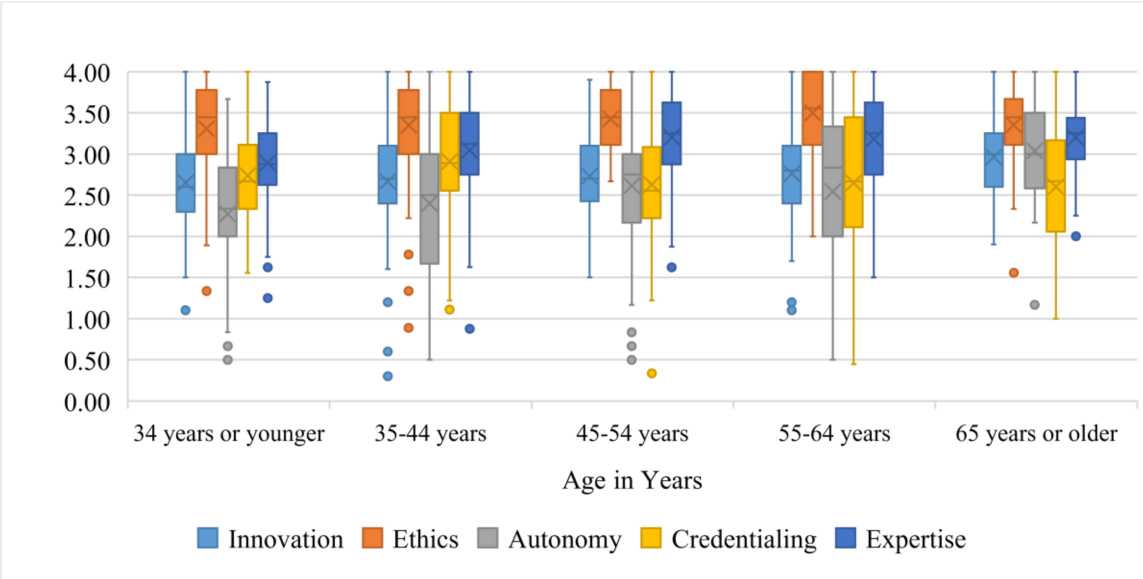

Figure 5.28. Variability in Professionalism Behavior by Seniority in Age

Note: The rating scale used in the professionalism assessment survey was; strongly disagree (1), disagree (2), agree (3), strongly agree (4). The "X" in the boxplots indicates the mean scores for each sub-scale.

Interestingly, Figures 5.29 and 5.30 show that younger evaluators were more likely to agree or strongly agree on being credentialed, shown by the elevated boxplots of the evaluators below 44 years old. On the contrary, older evaluators ages 55 or above had much variability of opinions on credentialing, as shown by the longer boxplots. It is also interesting to note that evaluators aged 45 to 54 years had the least belief in credentialing, shown by the lower positions of their boxplot and median score.

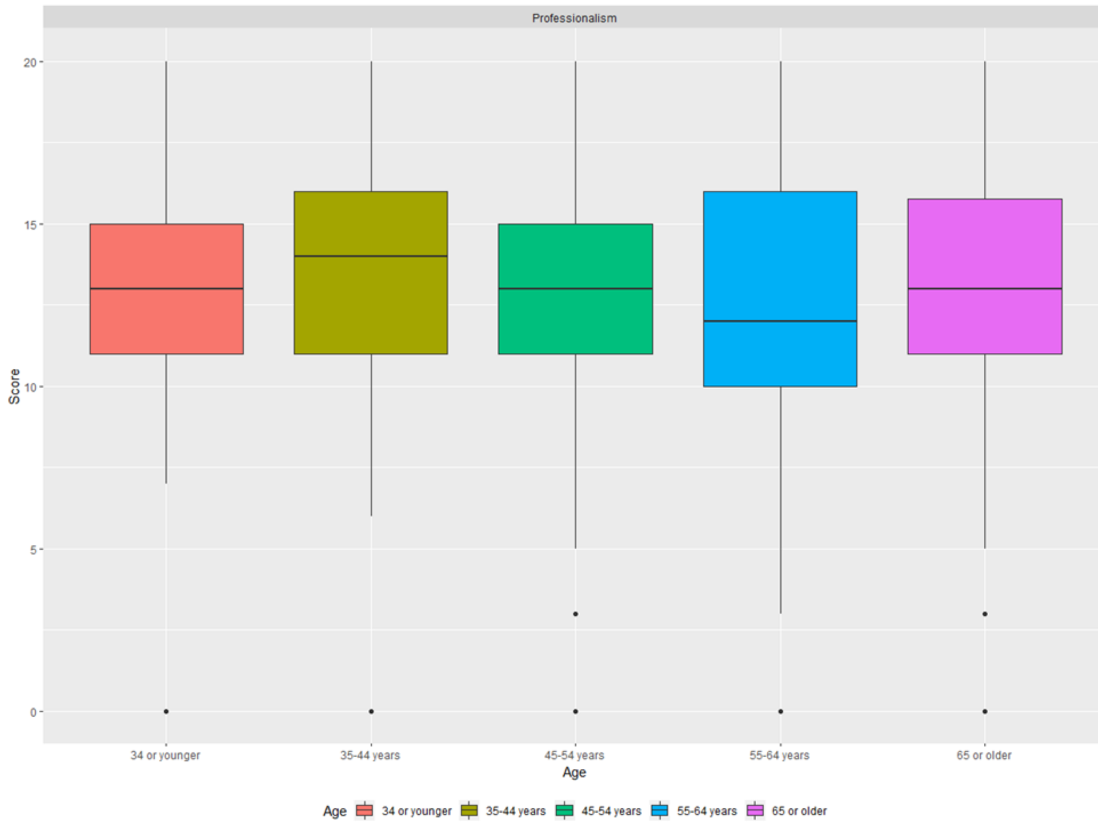

Figure 5.29. Variability in Credentialing Behavior by Seniority in Age

Note: The rating scale used in the professionalism assessment survey was; strongly disagree (1), disagree (2), agree (3), strongly agree (4).

To understand the variability of these responses by age, I plotted error bars of one standard deviation of the means depicted in Figure 5.30. The five error bars represent the five professionalism sub-scales by evaluators' age groups. The first error bar on the top level represents the Expertise sub-scale, the second bar represents the Credentialing sub-scale, the third error bar from the top represents the Professional Autonomy sub-scale, the fourth bar represents the Ethics sub-scale, and the fifth bar from the top represents the Innovation and Research sub-scale. This order is the same for all age categories.

Although evaluators of all age groups demonstrated consistent agreement on ethics, as evidenced by the shorter and elevated error bars, items on the Ethics sub-scale was more highly endorsed by the older evaluators 55 years and above than the younger evaluators, based on their means scores. This figure also confirms the variability of opinions on credentialing among evaluators aged 45 years and above, shown by their wider error bars. Evaluators in the age group 35 – 44 years old had the highest mean score in credentialing. The figure also suggests that evaluator professionalism behaviors in autonomy, innovation, and expertise were gained with seniority in years, demonstrated by the somewhat positive linear relationship.

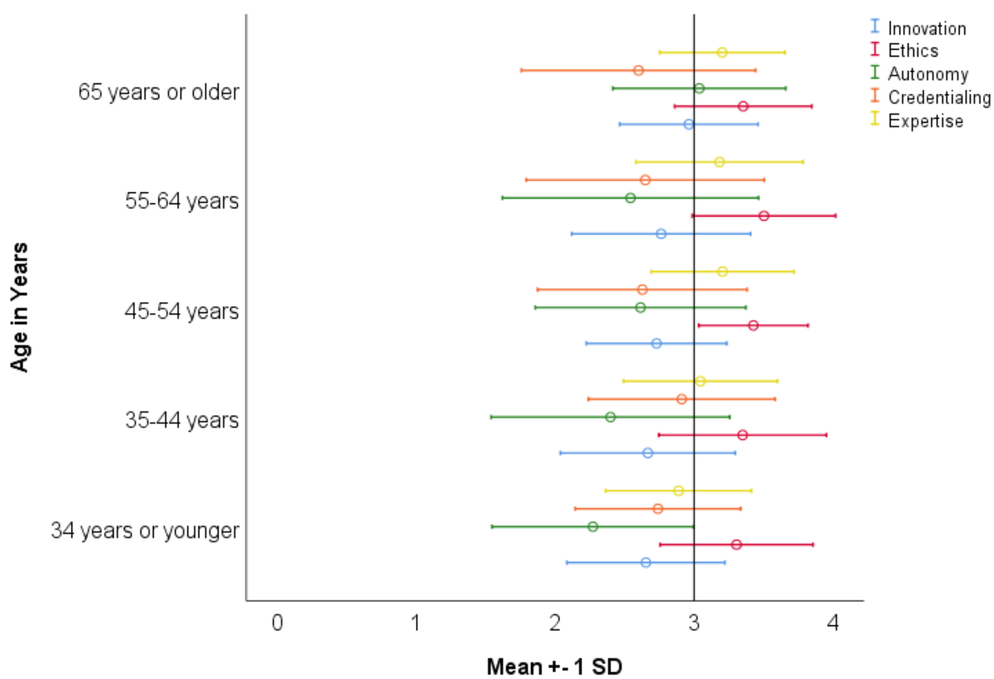

Figure 5.30. Variability in Professionalism by Age in Years

Note: The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. The vertical line is drawn on “agree” and the round dots represent the means.

I conducted a one-way MANOVA for the five evaluator age groups to understand if there was statistical significance in the differences in means for the dependent variables in the model. The assumption of independence of observations was checked and met, but the assumption of homogeneity of variance/covariance was violated as Box's M (110.88) test was significant (*p* = .001). Furthermore, the Levene's test of equality of variances was significant for Professional Autonomy (*p* = .002). For this reason, I used Pillai's trace criterion to assess the statistical significances among the evaluators' age groups on a linear combination of the five professionalism sub-scales and used Games-Howell non-parametric criterion for the post hoc tests.

Overall, the result of the one-way MANOVA was statistically significant, Pillai's $\vee = .173$, $F(20, 1480) = 3.183$, $p < .001$, and medium effect size (partial eta squared) of .043. Power to detect the effect was 1. Thus, the hypothesis that evaluator professionalism behaviors and their attitudes on credentialing was influenced by their seniority in ages was confirmed. The multivariate $\eta^2 = .043$ indicates that approximately 4% of multivariate variance of professionalism behaviors was associated with evaluators' age, and a small effect size, according to Cohen (1988).

Given the significance of the overall test, the univariate main effects were examined, based on the .01 adjusted alpha level (.05 divided by the number of ANOVAs conducted). As shown in Table 5.25, the results indicated that seniority in years was statistically significant for three professional sub-scales: the Innovation and Research ($p < .040$); Professional Autonomy ($p < .001$); and Expertise ($p < .01$). The overall

Professionalism scale ($p < .025$) and the Credentialing ($p = .047$) sub-scale were not statistically significant.

Table 5.21 shows that evaluators in the 65 years and above old age group reported higher means on the Innovation and Research sub-scale ($m = 2.96$) and the Professional Autonomy sub-scale ($m = 3.04$) on average than the younger evaluators. However, younger evaluators in the age group 34-44 years old reported higher means on the Credentialing sub-scale ($m = 2.91$) on average compared to evaluators in the youngest and older age groups. Evaluators in the age groups 45 – 54 and 65 years or older reported higher means for the Expertise sub-scale ($m = 3.20$) on average than the other age groups. Finally, evaluators in the 55 – 64 years age group reporter higher means for the Ethics sub-scale ($m = 3.50$) on average than younger and older evaluators.

I conducted a post hoc test to understand which pairs of means were significantly different. To protect against Type I error, I maintained the .01 alpha level across comparisons for the five sub-scales to be consistent with the decision for testing the five ANOVAs at the .01 level. With the Bonferroni method, I tested each comparison at the alpha level for the ANOVA divided by the number of comparisons (.01/5 = .002). Because the assumption of homogeneity of variance-covariance was violated, I used Games-Howell nonparametric post hoc criterion.

Table 5.21
Univariate Mean Differences by Age in Years per Sub-scale (at alpha = .05)

		34 years or younger		35-44 years		45-54 years		55-64 years		65 years or older					
	df	M	SD	M	SD	M	SD	M	SD	M	SD	F	*d (Eta Squared)*	*p*	*Observed Power*
Innovation	371	2.65	0.57	2.67	0.63	2.73	0.51	2.76	0.64	2.96	0.50	2.539	.027	.040	.717
Ethics	371	3.31	0.55	3.35	0.60	3.43	0.39	3.50	0.52	3.35	0.49	1.629	.017	.166	.501
Autonomy	371	2.27	0.73	2.40	0.86	2.61	0.76	2.54	0.92	3.04	0.62	7.730	.077	.001***	.998
Credentialing	371	2.74	0.60	2.91	0.67	2.63	0.75	2.65	0.86	2.60	0.84	2.435	.026	.047	.697
Expertise	371	2.89	0.52	3.05	0.55	3.20	0.51	3.18	0.60	3.20	0.45	5.157	.053	.001***	.967
Professionalism	371	2.80	0.36	2.90	0.41	2.93	0.38	2.95	0.43	3.02	0.38	2.827	.030	.025	.769

Note: * p<.05 ** p<.01 *** p<.001. The effect size (d), was computed by MANOVA. Partial Eta Squared (η^2) .000 to .003 = no effect, .010 to .039 = small effect; .060 to .110 = medium or intermediate effect, .140 to .2 = large effect, and .039 to .2 = desired effects (Cohen, 1988).

Significant age pairwise differences were obtained in the Professional Autonomy and the Expertise sub-scales at the adjusted alpha level of $p < .002$. The pairwise comparison showed statistically significantly differences between evaluators in the 65 or older age group and all four younger age groups on the Professional Autonomy sub-scale; evaluators in the 34 years or younger age group ($p = .001$), evaluators in the 35-44 years age group ($p = .001$), evaluators in the 45-54 years age group ($p = .009$), and evaluators in the 55-64 years age group ($p = .008$). Pairwise comparisons showed that the evaluators in the 65 years or older age group reported significantly higher professional autonomy behaviors on average than middle-aged and younger evaluator age groups. This suggested that older evaluators were more likely than younger evaluators to report having professional autonomy behavior.

Pairwise age comparisons also showed statistically significant differences between evaluators in the 45-54 years old and those 34 years or younger on the Expertise sub-scale ($p = .001$) and between evaluators 65 years or older and evaluators 34 years or younger on the Expertise sub-scale ($p = .005$). Interestingly, evaluators in the 45-54 age group reported significantly higher expertise behavior on average than the 34 years old or younger age group ($p = .001$). Similarly, evaluators in the 65 years or older age group reported higher expertise behaviors on average than the 34 years or younger age group ($p = .005$). Older and middle-aged evaluators were more likely than evaluators 34 years or younger to report having expertise behavior on average.

There was no statistically significant difference in the Professional Autonomy sub-scale between evaluators in the 35-44 years and 34 years or younger age group ($p =

.283) and between evaluators in the 55-64 years age group and evaluators the 35-44 years and 34 years or younger age group (p = .015). There were also no statistically significant differences in the Ethics, Innovation and Research, and Credentialing sub-scales and the overall Professionalism for evaluators in the different age groups at the adjusted alpha level of .002.

Chapter Summary

Chapter 5 provided quantitative results obtained through statistical analyses in the quantitative phase of this study. The results of the administration of PAToPE showed that the instrument provided evidence of sub-scale validity and reliability. The reliability was acceptable, and the confirmatory factor analyses provided modifications that produced a model with a reasonable fit. The statistical analyses indicated that program evaluators reported variable levels of perceived professionalism by country, workplace, gender, credentialed evaluator status, level of skills, level of education, age, and years of evaluation practice. Overall, credentialing was the main difference between Canada and the US. All evaluators endorsed ethics irrespective of the demographic characteristics defining their professionalism levels. In the next chapter, I discuss the implications, recommendations, limitations, and directions for future research.

CHAPTER 6

Discussion and Conclusion

Introduction

This study investigated how program evaluation is becoming a profession in North America by studying the reported behaviors of practicing evaluators on five traits of professionalism: Expertise, Innovation and Research, Ethics, Professional Autonomy, and Credentialing. Identifying and measuring key professionalism traits through self-report can begin to suggest the level of the maturity of evaluation as a field of professional practice. Because this framework encompasses frequently mentioned aspects of professionalism, the findings provide new insights about the current discourse among evaluators on the professionalization of the field. It focused on traits that may represent whether evaluation has matured enough to gain respect and reward in society as a profession.

In the previous chapter, I presented the results of my data analyses, which showed that the perceived professionalization of evaluation appears more advanced in Canada than in the US. Program evaluators in this study reported variable levels of professionalism with an overall endorsement of ethics irrespective of country, place of work, gender, credentialed evaluator status, level of skills, level of education, age, and years of evaluation practice. In this chapter, I present the substantive outcomes of the study and their implications for professional associations, trainers of evaluators, and professional practice.

Background

The professionalization of program evaluation gained center stage for the global evaluation community when the UN made a historic declaration that 2015 was the international year of evaluation, "EvalYear2015." This declaration came at a time when the UN's Sustainable Development Goals (SDGs) replaced MDGs that were highly criticized by the member states for not paying sufficient attention to evaluative evidence on achievements of the goals and learnings from implementation challenges ((IOCE website, n.d.). World leaders vowed to eliminate extreme poverty by 2030 with a belief that evaluation is necessary for accountability, decision making, and the improvement of member states' policies, programs, and strategies.

While this call to action raised the visibility of program evaluation as field of professional practice, the evaluation community may fail to meet global demand for high-quality evaluations if (a) there is no mechanism for regulating the supply of competent evaluators; (b) there are different versions of competencies required for judging the qualifications and competence of evaluators; (c) evaluation practitioners are neither required to exhibit competence in specialized knowledge nor conform to technical or ethical standards; and (d) there is a lack of funds for theory-based innovation and research as the engine of growth for the profession.

The international evaluation community welcomed the opportunities prompted by the United Nations' historic declaration. EvalPartners and the International Organization for Cooperation in Evaluation (IOCE) are leading a global movement to professionalize evaluation using professionalism as their organizing framework, but it is not clear what the term *professionalism* means in evaluation (Schwandt, 2017) or how to empirically assess it. As such, professionalization means different things to different people – VOPEs are taking different paths to professionalize the field, such as the voluntary credentialing program in Canada and the voluntary evaluator peer review (VEPR) system in Europe. Furthermore, existing professionalism measurement instruments do not speak to the needs and practice of program evaluation (e.g., the Professionalism Assessment Tool for pharmacists [Kelley, Stanke, Rabi, Kuba, & Janke, 2011]).

The primary purpose of this study was to investigate the extent to which program evaluation is becoming a profession in North America and to determine how to assess professionalism. This study was guided by three research questions: (a) how do program evaluators in Canada and the US describe professionalism in their everyday practice of evaluation; (b) how can professionalism in evaluators be assessed; and (c) to what extent and in what ways do program evaluators in Canada and the US demonstrate professionalism in their practice of evaluation?

The outcomes of this study are summarized by question in Tables 6.1. I shall present the summary of findings in three broad areas: (a) the development of a professionalism self-assessment tool for program evaluators, (b) the perceived status of program evaluation as a profession in Canada and the US, and (c) self-reported professionalism behaviors of the Canadian and US evaluators.

Summary of Findings

Professionalism Self-assessment Tool for Evaluators. My study used the behavioral indicators of professionalism, including Ethical Dispositions, Expertise, Professional Autonomy, Innovation and Research, and Credentialing[10] (Freidson, 1970, 1994; Picciotto, 2011). The meanings of these attributes were derived from interviews with practicing evaluators in the US and Canada to situate the traits to the field of evaluation practice. (See details in Table 6.1.) A unique contribution of this study to the field was the development of a measurement instrument for evaluators to self-assess their levels of professionalism. This instrument can be completed in six to ten minutes, and it is suitable for pre-conference or pre-workshop assessments as well as a baseline assessment in university-based courses for internships and practicum.

Previous work by Modarresi (1998) also used the sociological models of professions to assess both the structural and attitudinal measures of professionalism in sampled members of AEA. This study extended Modarresi's work, but this new instrument—probably the first of its kind in the field—differed in three important ways:

1. Foremost, the major focus of PAToPE is on the measurement of behavioral attributes of professionalism; the focus of Modarresi's instrument was on assessing program evaluators' familiarity with key theoretical approaches in evaluation and the impact of those approaches on their work. She also assessed evaluators' perceptions toward AEA meetings, the need for specialized training, and attitudes toward evaluation as a profession (pp. 97-115). However, while Modarresi's instrument assessed evaluators'

[10] Prestige and status (Picciotto's fifth indicator) are beyond the scope of this study since my primary interest was to study the behaviors of practicing evaluators, not the status of program evaluation.

attitudes toward the profession, PAToPE assesses whether evaluators reportedly possess professionalism traits or attributes. Both instruments overlap in their assessment of professional autonomy, but the items in each differ.

2. Second, the psychometric properties of PAToPE were empirically tested in order to increase confidence in the findings from this study and the potential for future use by other scholars. The results show evidence of construct validity and reliability. The exploratory factor analysis confirmed a presence of five latent factors measuring professionalism: Expertise, Innovation and Research, Ethics, Professional Autonomy, and Credentialing. These factors were validated using confirmatory factor analysis and structural equation modeling techniques. The structural equation modeling technique provided modifications that produced a model with excellent statistical fit indices and satisfactory reliability (α=.90, Omega=.95).
3. Finally, development of PAToPE extended work started with Modarresi's instrument because it included additional professionalism traits. As such, this instrument can improve our understanding of professionalism in the current discourse of professionalization of program evaluation in many ways. For example, a path analysis hypothesized four important relationships that may be relevant to the professionalization of the field: (a) a direct relationship between credentialing and expertise; (b) indirect relationships between credentialing and professional autonomy through expertise and credentialing and ethics through expertise; (c) a relationship between expertise and increased ethics and autonomy and an indication that ethics and autonomy shared unexplained causes outside the model; and (d) moderate

covariation between autonomy and ethics after controlling for the other relationships in the model.

Table 6.1 shows bi-variate relationships and the joint display of the quantitative results from the survey compared to the qualitative findings from interviews summarized. The instrument originated from the interview data and was validated with a wider population of evaluators. The table shows strong relationships between the overall Professionalism scale and all sub-scales and strong to moderate relationships within sub-scales. For example, the qualitative findings on how evaluators gained proficiency in evaluation approaches was confirmed by quantitative results. The table shows that a strong positive relationship between the Expertise sub-scale and the overall Professionalism scale, meaning expertise is an important component of evaluator professionalism. As seen in the table, the quantitative findings further support the assertion that professional autonomy was related to evaluator characteristics and agency. There is a strong relationship between the Professional Autonomy sub-scale and the overall Professionalism scale. Furthermore, the table shows confirmation of the assertions made in the qualitative findings by the quantitative findings for the Ethics, Innovation and Research, and Credentialing sub-scales.

The table shows strong correlations within these sub-scales. For example, the correlations show that evaluators who engage in innovation and research report having strong professional autonomy and expertise traits. Furthermore, people reporting strong traits in expertise and professional autonomy are more likely to report being ethical in their evaluation practice. The high association between the overall Professionalism and the Expertise sub-scale and between the Expertise and Ethics sub-scales confirms the assertion by Pavalko (1987) that a professional is a competent person who has mastery of the trade and that the person knows the trade and is good at the job and trustworthy or ethical. The relationship within the sub-scales is important in our understanding of professionalism. For example, the model hypothesis that credentialing has the potential to increased expertise suggests that people with proof of tested competence are more likely to report increased expertise. Because these traits are interrelated, it is likely that the development of professional designation programs to test the competence of members may be very important in the professionalization of program evaluation and in professional identity formation.

Table 6.1
Joint Display to Compare the Quantitative and Qualitative Findings

Qualitative Findings	Quantitative Findings: Sub-Scale	Quantitative Findings: Relationships Between Scales	Inference (Confirmation/Discordance)
Most evaluators acquired the competence or proficiency in conducting program evaluation encompassed in evaluator competencies formally and informally. Academic evaluators contended that skills can be acquired informally while expert knowledge is acquired through prolonged specialized competency-based training at the tertiary level.	Expertise M = 3.02 (SD = 0.60) (Moderate to high agreement)	Professionalism: r = .724, p<.001 Innovation: r = .584, p < .001 Ethics: r = .488, p < .001 Autonomy: r = .298, p < .001	Partially confirmed: The descriptive statistics indicated that a higher percentage of evaluators reported mastery of evaluation techniques through self-assessment, informal learning, and collegiality. Qualitative results implied that specialization came through undergoing specialized education at the tertiary level followed by substantial exposure to expert practice. Quantitative results support qualitative findings that expertise differs by personal characteristics of the evaluators. Expertise is highly correlated with professionalism, innovation, and moderately associated with ethics and autonomy
Evaluators noted that Professional Autonomy comes with personal agency, confidence, level of Expertise, years of experience and evaluator's workplace.	Autonomy M = 2.33 (1.07) (Moderate to high agreement)	Professionalism: r = .603, p<.001 Ethics: r = .374, p < .001	Confirmed: Quantitative results supported qualitative findings indicating that external evaluators with relative freedom to choose or end evaluation jobs have more autonomy than internal evaluators and evaluation managers. Autonomy also differed by age, education, and years of evaluation experience. Autonomy is highly associated with professionalism and moderately associated with ethics
Evaluators reported that their practice of program evaluation is guided by the principles of competence, integrity, respect for people, working for the common good and equity.	Ethics M = 3.35 (SD = .54) (Moderate to high agreement)	Professionalism: r = .604, p<.001 Expertise: r = .488, p < .001 Autonomy: r = .374, p < .001	Confirmed: Quantitative results support qualitative results indicating that nearly all evaluators in the sample agreed or strongly agreed on principles of ethical practice. Ethics is highly associated with professionalism and moderately associated with expertise and professional autonomy

Table 6.1 (Continued)

Qualitative Findings	Quantitative Findings: Scale	Quantitative Findings: Relationships Between Scales	Inference (Confirmation/Discordance)
Academic evaluators with access to money and links to universities are more likely to innovate novel approaches to promote the profession than non-academic evaluation practitioners.	Innovation and Research M = 2.68 (SD = .63) (Moderate agreement)	Professionalism: r = .650, p < .001 Expertise: r = .584, p < .001 Autonomy: r = .358, p < .001	Confirmed: Innovation behavior was higher for evaluators who were either academic, male, expert level, have Ph.D., or those with over 20 years of experience. There was no statistically significant difference in Innovation and Research by country, CE status, and employment status. Innovation and Research is highly correlated with the overall Professionalism scale and the Expertise sub-scale, and moderately correlated with the Professional Autonomy sub-scale
Evaluators felt that Credentialing means their peers recognize their competence and would tell people they have reached a minimum standard to be a competent evaluator.	Credentialing M = 2.72 (SD = .75) (Moderate agreement)	Professionalism: r = .440, p<.001,	Confirmed: Quantitative results supported the qualitative findings which indicated that evaluators' belief in credentialing related to professionalism. The path analyses indicated that credentialing was related to expertise and that credentialing had indirect relationships with ethics and professional autonomy through expertise. Credentialing is moderately associated with professionalism

Note: *p<.05, **p<.01, ***p<.001. The rating scale used in assessing professionalism was based on a four-point scale of strongly disagree (1) strongly disagree (2), disagree (3) agree, and (4) strongly agree. Perfect bivariate relationship is ± 1, high association is ± .05 to ± 1, moderate association is ± .30 and .49, low association is below ± .29, and no association is zero.

The Status of Program Evaluation as a Profession in Canada and the US. Professionalization of the field is perceived to be more advanced in Canada than in the US. It is important to note that due to Credentialing differences between Canada and the US, the credentialed evaluators group in this study is comprised mainly of Canadian evaluators (98%). It is not surprising that the results comparing CE and non-CE mimic those of the country comparison in which Canadian evaluators reported higher levels of overall Professionalism and Expertise and endorsement of Credentialing on average than American evaluators.

This study shows that CEs were more likely than non-CEs to define themselves as "professional evaluators." This finding is consistent with research from both Gauthier, Kishchuk, Borys, and Roy (2015) and Fierro, Galport, Hunt, Codd, and Donaldson (2016). The CES logical framework for the Professional Designation Program shows that the goal of the program is to increase identification of practitioners as professional evaluators and the recognition of evaluation as a distinct profession (CES website, n.d.).

One possible explanation for this difference in professional identity is that going through the credentialing process and maintaining active CE designation trigger professional identity formation (PIF) in CE – a process explained in the medical education literature by the conceptual change theory. Some medical scholars (Kay, Berry, & Coles, 2018) argue that changes in personal conceptualizations are initiated when cognitive disequilibrium is introduced by different experiences. These authors noted that "individuals reconcile the disequilibrium by adopting new conceptualizations that are intelligible, plausible, and fruitful" (p. 18).

Although my study did not investigate CEs' experiences of the credentialing process, it could be that the process of putting together a portfolio and responding to feedback on the portfolio from experienced and well-regarded reviewers challenge an evaluator's concept of self and trigger a new conceptualization of self as a professional.

Another incident that might have triggered CEs' professional identity relates to receiving the "CE pin." Similar to the white coat ceremony or other rites of passage for medical students, credentialed evaluators get a CE pin, a mark of accomplishment. Indeed, Gauthier noted that "[t]he CE designation has provided a tangible anchor (small things as apparently stupid as the CE pin and brand mark contribute to that dynamic)" (personal communication, March 18, 2019).

Reported Professionalism Behaviors of North American Evaluators

This study's findings identified five sub-scales used for assessing professionalism behaviors: Expertise or Competencies, Innovation and Research, Ethics, Professional Autonomy, and Credentialing. I now present the evaluator professionalism results based on people's responses to these traits. Table 6.2 provides a summary of the key study results by research question.

Overall professionalism. As seen in Table 6.2, reported overall Professionalism was higher for academic and external evaluators, Canadian evaluators, CEs, proficient and expert evaluators, and evaluators with over 21 years of experience. Evaluators who were above 45 years of age reported higher levels of overall Professionalism, although the effect size was small. There was no statistically significant difference in overall Professionalism by gender, level of education, or employment status. This study cannot answer why reported overall professionalism differed by workplace, skill level, and age, but the potential explanation about professionalism difference by CE status and country are explained in the preceding sub-section. The potential explanation about the relationship between high perception of professionalism and years of experience will be discussed later in this sub-section on the role of experience and deliberate practice in the acquisition of expert performance.

Ethical Dispositions. North America evaluators overwhelming endorsed ethics as the most important principle for conducting evaluations. Ethics was the common ground or the uniting factor in the field. Table 6.2 indicates that the endorsement was irrespective of evaluators' country, workplace, gender, CE status, level of skills, level of education, age, and years of evaluation practice. This is consistent with results of the validation of AEA evaluator competencies with members in 2018, research by Galport and Azzam (2017), and the validation of the Essential Competencies for Program Evaluators (ECPE) (Stevahn, King, Ghere, & Minnema, 2005).

There are two potential explanations for this level of endorsement. First, this could be related to self-affirmation bias. Psychologists argue that people tend to be defensive as a way of protecting their self-integrity if they perceive a threat to their self-concept (Ellemers, van der Toorn, Paunov, & van Leeuwen, 2019; Sherman & Cohen, 2006). Respondents in this study might have perceived that the items on the Ethics sub-scale questioned their sense of morality as virtuous individuals. As such, their responses may indicate that they perceive themselves as morally upright individuals with sound self-integrity.

Another potential explanation could be that the field of evaluation has truly matured, and practitioners desire change to protect the public from unethical practices. It appears that evaluation is today where medicine was roughly 200 years ago when the governing entities enforced the code of medical ethics and minimum standards of professional behavior. This move was meant to ensure that medical doctors worked in the public interest and promoted collegiality. Some participants in this study noted that the lack of enforcement may be hurting evaluation practice in North America. Evaluation standards and the AEA Guiding Principles may look good on paper, but the lack of enforcement may minimize their effect.

As shown in Table 6.2, only evaluators who self-identified as experts had a statistically significant difference on the Ethics sub-scale scores. In fact, expert evaluators reported having more ethical dispositions compared to novices and skilled or proficient evaluators. It is possible that evaluation practitioners are beginning to report conforming to ethical standards. It is not clear as to why evaluators' perceptions of ethics differed by skill level.

Credentialing. This study revealed that evaluators perceived Credentialing differently, as demonstrated by the variability of opinions on this issue of whether the evaluation professional should be credentialed or not. Of significance were the perceptions of evaluation managers, CEs, Canadian evaluators, less experienced evaluators, and novice evaluators (see Table 6.2). Although not statistically significant at the reduced alpha level of .01, evaluation managers reported a higher level of belief in Credentialing on average than the other evaluator categories (e.g., external evaluators, $M=2.57$; internal evaluators, $M=2.77$; academic evaluators, $M=2.74$; evaluation managers, M=2.93). This finding is theoretically important and is consistent with that of Cousins, Cullen, Malik, and Maicher (2009) and Fierro et al. (2016) who found that the feelings of evaluation managers and employers towards Credentialing differed from practitioners. For evaluation managers—who hire staff and are responsible for the successful execution of evaluations— this study showed that credentialing may help assure them that those who engage in evaluation conform to certain professional standards. Fierro et al. (2016) found that evaluation commissioners and employers who used CE to make their decisions felt that

> … the CE provides credibility to the work performed; having CE evaluators may increase marketability to prospective clients; using the CE can help to obtain a better pool of candidates and facilitate the development of a shortened list of individuals to get to know better; it provides an indication that an individuals' skills and knowledge have been kept "fresh"; and adopting the CE designation is simply helpful to the evaluation field. (Fierro et al., 2016, p. 33)

This finding reinforces previous reported reservations held by employers about the professionalism of evaluators, especially the concerns raised by the federal government– the largest employer of evaluators in Canada—about the relevance, quality, and rigor of evaluation (Cousins et al., 2009). The federal government in Canada contemplated starting its own credentialing program for their evaluators in order to raise their professionalism. This move by the federal government contributed to CES' development of a credentialing program in their Professional Designation Program. Interestingly, the federal government has neither officially recognized the CES Credentialing Program nor demanded that all their evaluation employees be credentialed.

Interview participants were concerned that without such support, voluntary credentialing lacks clout.

In fact, a large proportion of the online survey participants and interview respondents lamented the lack of recognition and support of credentialing from evaluation employers, despite anecdotal requirements by some Canadian employers requesting CEs in job advertisements or extra points for CEs in requests for proposals (RFPs). However, the anecdotal requirements of CEs in RFPs in Canada have triggered interest in the Credentialing Program by some individuals who are concerned about losing the market in evaluation work. As noted by Kishchuk et al. (2019), CES is receiving applications from social scientists and management consultants with less background and professional identification in evaluation. Many American and Canadian evaluators noted that they would seriously consider becoming credentialed evaluators if employers required this. In fact, 83% of Canadian and 17% of American survey respondents indicated that they "intend to apply" for CE status.

Donaldson's (2019) research yielded similar views on AEA members' interest in professional designation programs. Donaldson noted that sampled AEA members indicated a preference for "…a tiered system (beginner, intermediate, advanced, etc.)" that would account for the variability in levels of expertise and experience in the field (p. 158). In my study, several evaluators with extensive experience and Ph.D. level training felt their competencies were greater than the two years of experience required by the Canadian PDP, but others with similar educational status have embraced credentialing as a source of professional identity.

While the CES Professional Designation Program is showing promising results on how Canada is professionalizing the field, a socio-demographic analysis of the CES membership database revealed concern of the potential marginalization of younger, less experienced, and less well-educated evaluators who are beginning their careers in evaluation and need CE status the most (McDavid, Lawson, & Hunter, in press). McDavid et al. reported that

> CES members who have earned the CE designation tend to be: in the private sector; have a Ph.D.; be evaluators for a longer period of time; have evaluation as the major focus of their professional activities and be from either the Atlantic Region, the National Capital Chapter, or Western Canada. (p. 1)

As shown in Table 6.2, the findings from this study indicated that credentialing was endorsed by evaluators with fewer than 10 years of professional experience and novice evaluators or people just starting their careers in evaluation. McDavid et al. (in press) found that “[w]hile individuals aged 35 years and under comprise 31% of the CES membership, they account for only 15% of CE holders” (p. 5). A potential explanation for endorsement of credentialing by novice and less experienced evaluators might relate to their desire for affirmation by their peers and wish to gain credibility with employers. On the other hand, less endorsement of credentialing by expert and more experienced evaluators may imply that this demographic already has a niche in the field and a well-defined clientele. As such, experienced and expert evaluators may not see value addition from being credentialed.

This study highlights the need for professional associations to implement credentialing with caution and to be intentional about reaching younger evaluators and marginalized populations to ensure the inclusion of all evaluators. As noted in the qualitative analyses in Chapter 4, critics of credentialing and professionalization were concerned about the exclusion of marginalized people from practicing evaluation. They cautioned that credentialing was not a panacea leading to high-quality evaluation. Nevertheless, renewed interests in evaluative information by key employers like the US federal government may point to a need for some form of certification for evaluators to implement the Evidence Act of 2018 and the Federal Data Strategy.

Expertise. The descriptive statistics on expertise indicated that more evaluators reported mastery of evaluation techniques to make sound judgments through self-assessment and informal learning, contrary to theoretical assertions that expertise flows from access to high-quality tertiary education. This result is consistent with previous work by other scholars (LaVelle, 2014; LaVelle & Donaldson, 2015; Johnson, 2018) that evaluators gained mastery of their trade through informal training. In fact, academic evaluators differed on their perception of how evaluators acquire expertise. While most non-academic evaluators in the study reported getting expertise through informal training, academic evaluators rebutted this claim, saying that expertise is acquired through competency-based training in universities, but skills are acquired through short courses. This suggests the need to master evaluator competencies through training to become a competent evaluator. Non-academic perceptions were contrary to theoretical assertions that professionals get specialized training through completing courses in tertiary institutions.

With regards to expertise, evaluators who were more likely to identify as experts in evaluation were CEs, Canadians, and those who identified as experts, had doctoral degrees, worked full-time, had over 11 years of experience, and 45 – 54 years old (see Tables 6.2). The finding about CEs evaluators was in line with CES' intention for promoting CE designation for individual evaluators "…as a means of conveying evaluation competence, of demonstrating a commitment to the profession, of guiding professional development, and obtaining a competitive edge in evaluation-related practice fields" (McDavid et al., in press).

Nevertheless, this study is ambivalent as to *why* CEs differed on reported expertise from non-CEs. One possible explanation arises from the path analysis of the professionalism model tested in this study. As discussed in the previous chapter, path analyses showed that credentialing is directly related to expertise and innovation and research, therefore suggesting that those who were credentialed were more likely to report increased expertise. It is no surprise therefore that CE reported higher levels of expertise than non-CE on average. Furthermore, not only are CE required to show proof of their education and experience in evaluation, but their continuous reflection on evaluator competencies and using competencies for their professional development through the mandatory 40 hours of continuing education to keep current seemed to contribute to their expertise. As such, expertise was related to evaluator competencies.

The findings that evaluators with doctoral degrees reported higher levels of expertise was consistent with the theoretical assertions by Freidson (1970) and Piciotto (2011) about specialized training. These scholars argued that expertise is related to high-quality education, exposure to practice, theoretical knowledge, specialized skills, sound judgment, and the mastery of techniques. Evaluators who worked full-time, had over 11 years of experience, and were 45 – 54 years old reported higher levels of Expertise. This finding was consistent with the role of experience and deliberate practice in the acquisition of expert performance.

Psychologists have argued that deliberate practice occurs when individuals intentionally repeat an activity in order to improve their performance or when they engage in intense practice extended for a minimum of 10 years (Ericsson, 2006; Ericsson, Krampe, & Tesch-Romer, 1993). These scholars asserted that individuals can obtain expert performance from prolonged efforts to improve performance while negotiating motivational and external constraints. The theoretical framework of deliberate practice could be a potential explanation for reported expertise in evaluators with over 11 years of professional practice, especially when they worked full-time for many years. Ericsson (2006) cautions that individuals who reach the highest levels of professional mastery or expert levels of achievement are intentional about changing particular aspects of their performance. As such, experience and deliberate practice may not lead to an acceptable level of proficiency unless accompanied by deliberate efforts to change.

Innovation and Research. The quantitative analyses showed that academic evaluators reported a higher level of innovation and research behavior on average than other evaluator categories. Furthermore, the qualitative study revealed that academic evaluators with access to money and links to universities were more likely to innovate novel approaches to promote the development of the profession than were non-academic evaluation practitioners.

One possible explanation is that research is a large part of the job requirement for academic evaluators. In fact, 64.2% of academic evaluators reported participating in formal research studies on evaluation, and 63% reported purposefully generating new evaluation knowledge to benefit the field. Freidson (1994) hypothesized that universities are “shelters” where members of the professions institutionalize research and intellectual innovation to advance a discipline. He argued that research is conducted by special members of the profession serving as faculty since practitioners are heavily involved in the day-to-day activity of serving others. Starr (1988) likewise asserted that funding for research is critical for the advancement of a profession.

Aside from academic evaluators’ engagement in research, other evaluators with high reported levels of innovation and research traits were male, those with doctoral degrees, evaluators who self-identified as experts, and those with over 20 years of professional experience (see Tables 6.2). Nevertheless, very few non-academic evaluators reported conducting research on evaluation to promote the field. For example, evaluators with over 20 years of professional experience were 26%, those with doctoral degrees were 36%, and expert evaluators were 20% of those who reported higher levels of innovation. This study confirms earlier findings by Galport et al. (2017) that very few evaluators intentionally generate new knowledge to benefit the field. In fact, a participant in Galport et al.’s study noted that, ‘‘Very surprised to see that contributing to the knowledge base of evaluation was [rated] so low’’ (p. 86).

The low endorsement of the items on innovation and research by non-academic evaluators suggests that this area was not yet well-developed in evaluation (Ofir, 2016). Interestingly, some practitioners argued that they were better suited to conduct research than academics because academic evaluators lacked context. Nevertheless, most practitioners noted that they were constrained by a lack of research funds. Some stated that they must choose between bread and butter or work voluntarily when it comes to conducting research on evaluation. Another complication for non-academic evaluators was that they may not be able to publish evaluation reports due to proprietary concerns. In most cases, evaluation reports are the intellectual property of their clients.

Access to research funds and links to universities emerged as two very important factors necessary for innovation and research on evaluation in both Canada and the US. Starr's (1982) study demonstrated that substantive financial support for research and innovation from wealthy individuals such as Johns Hopkins (who founded Johns Hopkins University and Medical Center) and foundations like the Rockefeller Foundation was paramount in the advancement of medicine as a discipline.

Professional Autonomy. This study found that professional autonomy appeared to be gained by level of skills, age, gender, education, and years of evaluation experience (See Table 6.2). Although not statistically significant at the reduced .01 alpha, males were more likely than female evaluators to report having professional autonomy from clients and employers despite similar levels of education. This finding is theoretically significant considering gender inequalities endemic in most societies. Interestingly, external and academic evaluators were more likely than other evaluator groups to describe themselves as autonomous from their clients and employers. In fact, internal evaluators reported the lowest level of professional autonomy when compared with other evaluator groups. At the country level, the US evaluators were more likely than the Canadian evaluators to reportedly to end their participation in an evaluation if the evaluation was not going well. US evaluators were also more likely than Canadian evaluators to agree that they would turn down jobs where they think that the client is unwilling to hear whatever might be found.

It is interesting to speculate on possible reasons why the reported levels of professional autonomy by internal evaluators was so much lower than those expressed by external evaluators in this study. One possibility is that others determine the internal evaluators' jobs for them, and they work within a system. Specifically, internal evaluators may have the autonomy to choose methods, but they are not autonomous from their employers nor can they select what evaluation job to do.

This study's results were consistent with the theoretical assertion that the "ideal-type" professionals are practitioners with the freedom of self-employment and the ability to control their working conditions (Larson, 1977). External and academic evaluators had the perceived freedom of self-employment and the ability to control their working conditions. Larson (1997) argues that professionals who sell their labor rather than their services to bureaucracies like governments fall short of being true professionals. Although some academic evaluators perceived internal evaluators as technicians who do what they are told, internal evaluators contended that they, too, were legitimate and "ideal type" professionals who were employed based on their credentials (Gispen, 1988; Collins, 1990). It is worth noting that some academic evaluators indicated that they practiced evaluation as external consultants. As such, they did not depend on evaluation jobs for a living since their incomes from consulting jobs supplemented their regular incomes as faculty or university-based researchers.

Regarding gender difference, some possible explanations for this unexpected result may be attributed to power differentials between genders. Forbes Magazine noted that self-promotion– that is, being willing to put oneself out there and believe in one's abilities— is harder for women than men. Laura Donovan, a staff writer and editor for the

Levo (League), argued that “in a culture where women are penalized or bombarded with guilt for self-promotion, it’s easy for women to take the defeatist approach and pursue other methods for success” (Forbes website, 2011). She argued that women who are most proactive in making their achievements visible advance in their careers compared to those who do not.

A study by O’Neill and O’Reilly (2011) found that women suffer a backlash effect for behaving in a masculine way. These authors noted that women who are aggressive, assertive, and confident are more likely to get more promotions in the business world than passive men and other women. This notion of less self-promotion potentially explains why women evaluators in this study tended to endorse items on Professional Autonomy and Expertise at a lower level of the scale than men even though they had similar qualifications.

Another possible explanation of the gender difference in professional autonomy is the notion of self-esteem. Research on self-esteem gap suggests that men at every age tend to have higher levels of self-esteem than women worldwide (American Psychological Association website, 2016) and that this self-esteem gender gap is more pronounced in Western industrialized countries (Bleidorn, Arslan, Denissen, Rentfrow, Gebauer, Potter, & Gosling, 2016).

Evaluators in the sample reported acquiring Professional Autonomy through years of practice, indicating that Autonomy is not taught. This is consistent with Brown and Dinnel’s (1992) finding that university-based preservice training programs do not teach evaluators how to be autonomous in their practice of evaluation. This study showed that two-thirds of evaluators reported acquiring soft skills like Autonomy informally.

This finding is consistent with Johnson's (2018) study that found that evaluators seem not to develop soft skills or many of the essential interpersonal competencies in formal degree programs, but through other ways, including practical wisdom or life experiences and professional development. In fact, this study showed that senior evaluators were more likely than novice evaluators to report having autonomy from clients and employers (see Table 6.2). Furthermore, Brody and Doukas (2014) argue that professionals apply ethical principles and deeper attitudes requiring practical wisdom to their decisions. According to Brody and Doukas (2014), inward character translates into the values, attitudes, and beliefs that influence behaviors.

Table 6.1
Questions and Summaries of Results

Research Questions	Key Results	Comments
Question 1. How do program evaluators in Canada and the US describe professionalism in their everyday practice of evaluation?	Description of the key themes for the five traits of professionalism by interview participants: ***Expertise*** – identifying strengths, experiential learning, university courses, keeping up-to-date, collegiality, changing behaviors ***Innovation and Research*** – research on evaluation, a cross-discipline adaptation of methods, sharing reflections and insights, modifying methods, applying established methods, and innovative reporting ***Professional Autonomy*** – refusing to do something unethical, can disagree with the client's approach, can choose methods, can choose a project, turn down jobs, end participation ***Ethics*** – competence, integrity, honesty, disclosure, respect for people, doing good, confidentiality, caring for others, self-protection, common good, equity, accountability. ***"Credentialing"*** – self-reflection, competence, affirmation by peers, protect clients, professional development, align to competencies, became aware of competencies, use competencies to teach evaluation	A model of five traits of professionalism was developed from the sociological studies of occupations – Expertise, Innovation and Research, Professional Autonomy, Ethics, and Credentialing. The meanings of these professionalism behaviors were defined based on the perceptions and experiences of practicing evaluators who were interviewed in this study.
Question 2. How is professionalism in evaluators assessed?	A tested theory-based Professionalism Assessment Tool for Program Evaluators (PAToPE) grounded in quotations from interviews was developed to assess professionalism. The psychometric assessment shows that PAToPE had acceptable internal reliabilities and reasonable model fit statistics. The reliability coefficient of the current instrument was at least α=.90 for the model (Omega=.95).	A path analysis of the relationships between the five traits showed that Credentialing had a relationship with Expertise. Furthermore, Credentialing had indirect relationships with Autonomy and Ethics via Expertise.

Table 6.2 (Continued)

Research Questions	Key Results	Comments
Question 3. To what extent and in what ways do program evaluators in Canada and the US demonstrate professionalism in their practice of evaluation?	Description of the key findings from quantitative analyses of the traits of professionalism: ***Overall Professionalism scale*** There were statistically significant differences in the overall Professionalism scale by country, workplace, CE status, skill level, age, and years of experience	Program evaluators in Canada and the US differed in their levels of professionalism by workplace, credentialed evaluator status, level of skills, age, and years of evaluation practice. There was no statistically significant difference in overall Professionalism by gender, level of education, and employment status.
	The Ethics sub-scale Both Canadian and US evaluators in the sample agreed that Ethics is the most important aspect of professionalism. Expert evaluators were more likely than novice and proficient evaluators to agree that they follow a professional code of ethics to guide their work and being transparent regarding their ethical values when conducting evaluations.	All evaluators in the sample agreed on the importance of Ethics irrespective of country, place of work, gender, credentialed evaluator status, level of skills, level of education, age, and years of evaluation practice.
	The Credentialing sub-scale Overall, Canadian evaluators, novice and less experienced evaluators felt that credentialing was more important on average compared to other demographics. Canadian evaluators were more likely than US evaluators to agree that credentialing made them aware of evaluator competencies, would tell people they have reached the minimum standards to be a competent evaluator, being a credentialed evaluator would mean that their peers have recognized their expertise, and that evaluator competencies should be aligned to credentialing.	Evaluation managers had the greatest belief in credentialing, potentially relating to their hiring roles. New and younger evaluators highly endorsed credentialing items, potentially due to the benefit of being affirmed by their peers and senior evaluators on the Credentialing Board. There was no statistically significant difference in credentialing by gender, education, and employment status, and age.

Table 6.2 (Continued)

Research Questions	Key Results	Comments
Question 3. To what extent and in what ways do program evaluators in Canada and the US demonstrate professionalism in their practice of evaluation?	***The Expertise sub-scale*** There were statistically significant differences in expertise by country, CE status, skill level, education, employment status, and experience in years.	Evaluators who were more likely to report higher levels of expertise behavior were CEs, Canadians, and those who identified as experts, had doctorate degrees, worked full-time, had over 11 years of experience, and 45 – 54 years old. There was no statistically significant difference in Expertise by workplace and gender.
	The Innovation and Research sub-scale There were statistically significant differences in the Innovation and Research sub-scale by workplace, gender, skill level, education, and experience in years.	Innovation behavior was higher for evaluators who were either academic, male, expert level, have Ph.D., or those with over 20 years of experience. There was no statistically significant difference in Expertise by country, CE status, & employment status
	The Professional Autonomy sub-scale There were statistically significant differences in the Professional Autonomy sub-scale by workplace, country, gender, skill level, age, education, and experience in years.	Evaluators with Ph.D., older and more experienced evaluators reported being autonomous in their practice. Internal evaluators reported the lowest levels of professional autonomy compared to other evaluators. There was no statistically significant difference in Autonomy by credentialed status and employment status.

Implications of the Study

The concept of professionalism serves as a broad framework for studying professionalization of the field. As such, this study has implications for professional associations, educators, and evaluation practitioners in their efforts to make program evaluation a profession.

Professional Associations. My study findings have three implications for professional associations in the current discourse on professionalizing program evaluation in Canada and the US. First, while associations and global efforts are geared towards developing professional designation programs such as credentialing in Canada and VEPR in Europe, enforcing codes of ethics and minimal standards of practice has the potential of improving the reputation of the field and building trust from the public. The overwhelming endorsement of ethics by the participants in this study suggests that evaluators agree on enforcement of codes of ethics and minimum standards.

It is important for CES and AEA to develop codes of ethics from their current ethical guidelines and guiding principles respectively into three – five simple, accessible and memorable moral principles to be used by practitioners in applying values and codes of conduct to their work. For example, AEA could commission the Guiding Principles Review Task Force to engage the membership in developing a code of ethics and conduct with practical recommendations for evaluation practice and research on evaluation. The CES could do the same, but could start by requiring the CEs to take at least one or two mandatory CES ethics courses to fulfill their 40 hours of professional development. In fact, CEs might welcome a mandatory ethics course since they endorsed the items in the Ethics sub-scale the most compared to non-CEs. Enforcing codes of ethics would assure the public that those who practice evaluation conform to ethical standards usual for professional bodies and that the association holds offenders accountable for evaluation malpractice. As noted earlier, evaluation practitioners are currently neither required to exhibit competence in specialized knowledge nor conform to technical or ethical standards.

Second, access to research funds for innovation is a crucial factor in the advancement of any profession. The non-significance of the research and innovation trait of professionalism for non-academic evaluation practitioners likely indicated the lack of attention and funding to promote evaluation as a profession. Funded collaboration between professional associations and universities might ensure that members of the profession who serve as faculty can create opportunities to experiment and innovate new approaches. Such collaboration could also serve to bolster the development of practical solutions and expand the body of knowledge for practitioners. For example, one of the key objectives of the Canadian Consortium of Universities for Evaluation Education (CUEE) is to develop research initiatives and academic publications involving CUEE faculty members and students (CUEE website, n.d.). Such noble initiatives could be extended to non-academic members of CES to challenge and give members opportunities to collaborate on research. In fact, collaboration among academics, non-academics, and professional associations could produce research products relevant to practice. Some of the non-academic participants were critical of academics because they perceived that academic researchers lacked practice and nuance in determining what works on the ground.

Some participants in this study indicated interest in engaging in research to promote innovation, but also noted being constrained by lack of funds for such work. As noted earlier, the medical field advanced due to substantive research funding from individuals and foundations. Professional associations are better placed to mobilize research funds accessible to both practitioners and faculty members of the associations through large-scale research proposals. For example, recent studies show that AEA members preferred a tiered Credentialing system (Donaldson, 2019). As AEA considers the next steps post-adoption of their evaluator competencies, the association could approach foundations and governments with fundable proposals for developing an evidence-based tiered Professional Designation Program and roll out a more plausible option after testing different options. In fact, AEA might consider grandfathering evaluators with several years of experience and those who identify as experts to avoid tension in the designation process. These evaluators endorsed Credentialing the least, but indicated having practical wisdom based on their responses to having professional autonomy, ethics, and expertise traits compared to novice and evaluators with the least experience who endorsed the items in the Credentialing sub-scale the most.

Third, this study shows that Credentialing was the main difference between Canada and the US. The high endorsement of Credentialing by managers and the evidence of perceived higher professionalism behaviors among CEs based on differences in means are hints that implementing Professional Designation Programs for its members could lead to the maturity of evaluation as a profession. These programs could provide evaluators opportunities to get credentials from their associations as proof of competence and signal to employers that associations have a mechanism for supplying high-quality evaluators to safeguard the public. The path analysis indicated that earning a designation may trigger professional identity formation whereby evaluators begin identifying with the field of evaluation as a profession and defining themselves as professional evaluators. Professional associations should promote credentialing and other forms of designation so that evaluators can access offices in bureaucratic settings like government offices without fear of being labeled inferior. As mentioned earlier in the problem statement, the field currently lacks a mechanism for regulating the supply of competent evaluators, including evaluation education.

Trainers of Evaluators. The findings from this study have two implications for trainers of evaluators. Foremost, the professionalism framework offers a frame of reference in the design of pre-service training of evaluators. This study's findings are consistent with prior research showing that key aspects of professionalism, including professional autonomy and ethics, are not consistently taught in university or non-university settings. Trainers of evaluators should intentionally and deliberately design professionalism curriculum or incorporate professionalism content in formal and informal curricula for training evaluators. Professionalism is foundational in university and non-university-based evaluation training programs and could provide topics in discussions, especially for students in internship programs or practicums. The professionalism traits covered in this study were consistent with the domains of the AEA and CES evaluator competencies, ethical guidelines, and evaluation standards. The qualitative findings indicated that preservice training programs based on competencies could provide an ethical grounding in the field of practice. Professionalism is part of the formal curriculum for training medical doctors, pharmacists, engineers, dentists, and social workers, and this could serve as a model for the field of evaluation.

Professionalism standards can guide students' training as well. The PAToPE instrument is suitable for administration at the beginning of relevant evaluation courses to assess prior knowledge and levels of professionalism. This could ensure that the training is tailored to the needs of students, and the results could serve as a method of tailoring mentoring and coaching of the students. As pointed out earlier, it takes fewer than ten minutes to complete this professionalism assessment instrument. Students can use their professionalism scores to guide personal study or to identify areas for more support from the instructors or their peers. Instructors and advisors could use the results to focus their efforts with individual students.

Professional Practice. This study also has implications for professional practice in Canada and the US. While participation in credentialing in Canada gave CEs a sense of professional identity, the question about how participation in credentialing triggers professional identity formation needs further exploration. This finding is critical considering the current debate on professionalization of the field. As mentioned in the problem statement in Chapter Two, professionalization means different things to different people. Furthermore, critics are concerned about the potential disadvantages, including potential marginalization of historically disadvantaged people and some new evaluators as highlighted by McDavid et al. (in press) in their study of the demographic characteristics of CEs.

It also appears that many evaluators do not understand the significance of the credentialing beyond the thought that it is a "mere" piece of paper. While some see the designation as not required by employers, others feel they already received their credentials from academic institutions and see no need for a credential from an association. Furthermore, other evaluators question whether those with the CE designation do a better job than non-CEs. These concerns are legitimate and require further investigation.

This current study also suggests that there was a strong relationship between the Credentialing and Expertise sub-scales and between the Expertise and Innovation and Research sub-scales. This suggests that credentialing may help evaluators improve their expertise through additional training and professional development, and this in turn may help evaluators define themselves as professional evaluators. Furthermore, the model suggests that reported expertise was related to increased ethics and increased autonomy. The model hypothesizes an indirect relationship between credentialing and autonomy through expertise and an indirect relationship between credentialing and ethics through expertise. It may be fruitful to explore these linkages and hypotheses to provide evidence of the perceived benefits of credentialing.

Additionally, of the 456 individuals who completed the online survey, nearly one-third (n=126) expressed interest in knowing their professionalism scores. This might suggest that these evaluators wanted to identify areas for improvement through professional development. Going through credentialing or similar designation programs has the potential of helping evaluators identify their strengths and weaknesses and guiding their professional development based on evaluator competencies. This finding may also be consistent with the notion that credentialing triggers identity formation.

Finally, this study showed that evaluators are developing collegiality with their peers, such as through referrals. Professionalism demands that evaluators become more dependent on one another for referrals and access to jobs instead of competing with one another. The medical profession advanced this approach when doctors started supporting each other through referrals, which in turn promoted specialization for better service delivery. Professionalism may serve as a source of meaning in work and a basis for solidarity so that evaluation gains respectability and recognition in the occupational ladders.

Strengths of This Study

This study has several strengths, as well as unique contributions to the literature on evaluator professionalism. First, it advances the meaning and the empirical measurement of professionalism in evaluators, an area that has received little attention in the current discourse on professionalization of the field (Schwandt, 2017). This study provides operational definitions of theory-based professionalism traits (e.g., Expertise, Ethics, Autonomy, Innovation and Research, and Credentialing) based on the language of evaluation practitioners in Canada and the US. Language and items in the measurement instrument were checked for face and content validity by well-known evaluation scholars and practitioners in Canada, Europe, and the US. As such, PAToPE is authentic and original.

Second, as noted earlier, PAToPE has desirable psychometric properties, including relatively high levels of reliability and convergent, discriminant, and construct validity. These qualities were rigorously tested using confirmatory factor analyses and structural equation modeling. The statistical fit indices were excellent, and it is possible that the instrument measured Canadian and Americans in the same way. (A measurement invariance was not conducted due to sample size.) The instrument has acceptable reliability indices. Thus, the outcomes of the current study are credible and provide reliable information about the reported professionalism of evaluators in Canada and the US and the status of professionalization in Canada and the US based on the study samples.

Another strength of this study is that it gives insights into the use of professionalism as an organizing framework for understanding and possibly implementing the professionalization of evaluation in Canada and the US. Previous studies on professionalization focused primarily on one or two traits of professions without giving a comprehensive picture of professionalization. The current study provided the big picture of professionalization as a process, along with key traits of professionalism, which is the desired outcome for professionals. It is worth noting that the professionalization debate stagnated for over two decades when Worthen (1994) concluded that evaluation fell short on three of the nine criteria for judging professions (i.e., exclusion of unqualified, accreditation of pre-service preparation programs, and certification and licensure).

The current trend in professionalization in evaluation scholarship is piecemeal, focusing on individual components, such as promoting credentialing, peer review systems, and accreditation of pre-service training without adequate focus on the big picture. Still, this study unraveled various professionalism traits, and these findings suggest that professional associations could advance the field by incrementally implementing professional designation programs alongside enforcement of ethics and minimum standards. In fact, a path analysis of relationships between professionalism traits showed that implementing one trait could indirectly lead to increases in other traits.

An incremental approach would take advantage of the current shift in perspectives and positions among evaluators in favor of designation, especially in the US. It has been almost a quarter century since Bickman (1997) urged AEA to professionalize the field in his presidential address. His position was that evaluators could secure the franchise as truth tellers of the effectiveness of public action like accountants who secured the franchise as the truth tellers on money matters. This approach also worked in promoting the medical profession and could be less threatening to people opposed to professionalization.

Finally, another strength of the current study is that it validated a theory-based descriptive model of professionalism to guide practice and professionalization of evaluation. As noted earlier, the study draws from the sociology of professions theories (Wilensky, 1964; Forsyth & Danisiewicz, 1985) and the logic of professionalism (Freidson, 1970; Picciotto, 2011). These theorists provided key traits of professions and professionalism as the desired measurable outcome. A theory-based descriptive model is traditionally used in most social science disciplines to guide practice (Stevahn et al., 2005).

EvalPartners/IOCE noted that evaluation is a “new kid” on the block among the social sciences because evaluation has not yet met all the prerequisites of a professional occupation (IOCE website, n.d.). The current discourse on professionalization would benefit from this empirically validated, theory-based descriptive professionalism model rather than solely relying on the representation of the authors' opinions that currently dominates the literature on professionalization.

Limitations of the Study

Although this study offers insights to inform the current debate on professionalization of the field, it is not without limitations. First, the sample for this study was restricted to members of AEA and CES who were practicing program evaluation in Canada and the US. Subsequently, the results are not generalizable to members of other professional associations who did not have an equal opportunity to participate in the current study. The results would probably differ for evaluators operating in different socio-economic and cultural contexts. The analyses excluded responses for international members of AEA. Furthermore, the CES sample was limited only to individual Anglophone members who were in good standing and agreed to third-party research and surveys. It is possible there would have been important differences had Francophone members participated in the study. Their participation would have necessitated translation of the survey into French and adapting the instrument to reflect their socio-cultural contexts. As such, the instrument would have required validation and perhaps adaptation to reflect the language and cultural differences.

Second, participants in the online survey self-reported their professionalism behaviors and beliefs in credentialing. One key weakness of relying on participants' subjective evaluations of their behaviors and aptitudes concerns social desirability bias.[11] Participants may have tended to report "good behaviors," especially ethical behaviors out

[11] "Social desirability is the tendency of some respondents to report an answer in a way they deem to be more socially acceptable than would be their "true" answer" (Lavrakas, 2008, p. 825).

of a desire to make a good impression in social interaction (or avoid a negative one) (Dillman, Smyth, & Christian, 2014). This may also relate to the self-affirmation bias discussed earlier in this chapter. Such biases interfere with interpreting averages as well as individual differences.

Third, the instrument itself has certain limitations. Foremost, it was not possible to pilot it. The initial plan was to test the instrument with the members of the Minnesota Evaluation Association, but it was impossible to get the desired sample of 600 for the piloting of a 59-item instrument, approximately ten participants per item (Kline, 2016). Therefore, although I conducted think alouds with two senior evaluators in Minnesota to test the clarity of items, the scale was not pre-tested.

For example, I discovered when conducting confirmatory factor analyses that a frequency scale was more suitable for measuring professional behavior compared to the use of an agree-disagree ordinal scale. As such, I revised the instrument in fall 2018 and validated it with a second sample of 2000 people from the AEA 2018 registry. However, data from this validation study were not included in this study. Future research should consider validating this revised instrument with members of CES and other evaluation populations to assess professionalism behaviors. Validation of a new instrument is a continuous process and cannot be accomplished through one study or by one researcher (Benson & Clark, 1982).

Fourth, this instrument was developed at the time when the AEA Evaluator Competencies were being formulated and the Guiding Principles for Evaluators were under review. Furthermore, the CES Evaluator Competencies were also being reviewed at the time of this study. It would be desirable to crosswalk the items of this instrument to the recently adopted AEA Evaluator Competencies, AEA's Cultural Competence Statement, and the updated Guiding Principles for Evaluators as well as the revised CES Evaluator Competencies to reflect the current thinking. Future research should consider these changes when validating the present instrument.

Finally, the sample size was too small to conduct a measurement invariance analysis. As such, I was unable to assess whether the instrument measured Canadians and American evaluators the same way. I recommend that future research could work with larger samples and to thoroughly investigate my assumption of measurement invariance for the instrument.

Recommendations for Future Research

The results of this study suggest several notable recommendations aimed at understanding professionalism and professionalization of the field of evaluation. First, professionalization of program evaluation is currently implemented across the globe, and professionalization means different things to different people. Future research should explore validating the professionalism assessment tool for evaluators with VOPEs in ethnically and culturally diverse settings to examine professionalism in evaluators in those contexts. The outcome of such validation would be useful in assessing the maturity of the field in those regions to (a) inform the choice of a relevant professional designation programs to improve evaluation practice, (b) improve the measurement instrument to suit other contexts around the globe, (c) inform pre-service training of evaluators, and (d) provide a tool for evaluators to self-assess their levels of professionalism to inform their professional development.

As noted earlier, a second area of future study could be to crosswalk items in the PAToPE with the competencies, standards, statement of cultural competence, and guiding principles to situate professionalism assessment in these frameworks approved by professional associations. The AEA Guiding Principles for Evaluators are used and modified by other VOPEs in other contexts (Picciotto, 2011). This would promote ownership and the use of the outcomes of the instrument by AEA and other professional associations to advance the field and evaluation practice globally. In addition, grounding the instrument in current language would ensure that it is kept up-to-date and promote research on evaluation, an area that currently lacks sufficient attention, a confirmation of the conclusions from other studies (King & Stevahn, 2013).

Third, it would be desirable to administer the professionalism survey to the alumni of AEA Graduate Education Diversity Internship (GEDI). The GEDI program has been in existence for nearly 10 years, and such a study would be an opportunity to assess the impact of the program on evaluators' practice and program effectiveness. This research should aim to determine whether the internship experience served to enhance professional identity formation of interns and achievement of the goals of the program. The outcome of such a study with GEDI alumni would also generate necessary feedback to strengthen the elements of cultural competence in the professionalism measurement instrument.

Finally, another important area of study is assessing the professional autonomy of VOPEs in advancing evaluation as a field of professional practice. The findings of this study indicated that evaluation practitioners reportedly lament the non-enforcement of ethics and evaluation standards by the professional associations. Future research should explore how VOPE’s professional autonomy is associated with occupational authority and prestige via controlling strategic accessories such as enforcing codes of ethics and controlling recruitment, quality of pre-service preparation programs, and disciplinary processes. As noted earlier, associations do not currently perform these roles of self-regulation and image-building, which in turn may affect the legitimacy of the profession and professional autonomy of evaluation practitioners. Prestige and status were beyond the scope of this study. It would be desirable to assess the professional status of program evaluation relative to other relevant professions in the division of expert labor.

Chapter Summary

The purpose of this study was to inquire about how program evaluation is becoming a profession in Canada and the US using sociology of professions theories. This study found that evaluation practice was reportedly more advanced in Canada than in the US based on mean differences for overall professionalism, although evaluators in this study felt that enforcing codes of ethics and minimum standards of practice would advance evaluation practice in North America. Enforcing codes of ethics and minimum standards of practice alongside competency-based professional development programs and preservice training would legitimize evaluation practice and likely make evaluation more reputable and trusted by the public. The study also found that expertise and professional autonomy came with many years of evaluation practice, the level of education, and some form of certification like credentialing. This implies that novice and evaluation students could learn professionalism traits if intentionally integrated into the training programs like practicums. Furthermore, using professionalism as an organizing framework for professionalization is feasible since it can be empirically assessed by studying the behaviors of practitioners. The professionalism framework showed that innovation and research was the area that was least developed and needing proper attention to advance the field through forming strategic relationships by the associations. A study of the medical profession (Starr, 2008) showed that innovation and research are the engine of growth for the any profession, especially when the profession is deemed to be a public good.

This study has provided evidence of the importance of using a professionalism framework in assessing how evaluation is becoming a profession in Canada and the US. Further research is needed to ground the professionalization efforts in theory to improve our understanding of making evaluation a profession. A further understanding of the role of the professional associations in self-regulation and image-building is crucial in legitimizing evaluation practice so that evaluation can secure the franchise as truth tellers of the effectiveness of public action. It would be regrettable if the field failed to maximize the current global interests in professionalization and the demand from the world governments for high-quality evaluation and to amplify gains made by CES in the evaluator designation program to improve evaluator professionalism. As we envision program evaluation as an emerging profession, we must plan for ways to position program evaluation as a public good by holding practitioners accountable for their actions to protect the public through codes of ethics and to improve evaluator professionalism.

REFERENCES

Abbott, A. (1988). *The system of professions. An essay on the expert division of labor*. Chicago, IL: University of Chicago Press.

Abbott, A. (1991). The order of professionalization: an empirical analysis. *Work and Occupations*, *18*(4), 355-384.

Abbott, A. (2014). *The system of professions: An essay on the division of expert labor*. Chicago, IL: University of Chicago Press.

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, *33*(9), 1289-1312. https://doi.org/10.1080/09500693.2010.512369

Altschuld, J. W. (1999a). 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. https://doi.org/10.1177/109821409902000307

Altschuld, J. W. (1999b). The case for a voluntary system for credentialing evaluators. *American Journal of Evaluation*, *20*(3), 507-517. https://doi.org/10.1177/109821409902000309

Altschuld, J. W., & Engle, M. (2015). The inexorable historical press of the developing evaluation profession. *New Directions for Evaluation, 145*, 5-19. https://doi.org/10.1002/ev.20108

American Evaluation Association (2015). *A draft of proposed evaluator competencies for the American Evaluation Association*. Paper presented at the American Evaluation Association Conference, Chicago, IL.

American Evaluation Association (n.d.). http://www.evaluationconference.org. Accessed on February 20, 2016.

Barrington, G. V., Frank, C., Gauthier, B., Hicks, K. (2015). View from the Credentialing Board: Where we've been and where we're going. *Canadian Journal of Program Evaluation/La Revue canadiaenne d'évaluation de programme, 29*(3), 86-97. https://doi.org/10.3138/cjpe.29.3.86

Bazeley, P. (2018). *Integrating analyses in mixed methods research.* London, UK: Sage Publications Ltd.

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. *American Journal of Occupational Therapy*, *36*(12), 789-800. https://doi.org/10.5014/ajot.36.12.789

Bickman, L. (1997). Presidential aAddress: Evaluating evaluation: Where do we go from here? *Evaluation Practice*, *18*, 1-16. https://doi.org/10.1177/109821409701800101

Bickman, L. (1999). AEA, bold or timid? *American Journal of Evaluation, 20*(3), 519-520.

Bleidorn, W., Arslan, R. C., Denissen, J. J., Rentfrow, P. J., Gebauer, J. E., Potter, J., & Gosling, S. D. (2016). Age and gender differences in self-esteem—A cross-cultural window. *Journal of personality and social psychology*, *111*(3), 396-410. http://dx.doi.org/10.1037/pspp0000078

Branch Jr, W. T. (2010). The road to professionalism: reflective practice and reflective learning. *Patient Education and Counseling*, *80*(3), 327-332. https://doi.org/10.1016/j.pec.2010.04.022

Brody, H., & Doukas, D. (2014). Professionalism: A framework to guide medical education. *Medical education*, *48*(10), 980-987. https://doi.org/10.1111/medu.12520

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. https://doi.org/10.1177/0193841X9201600102

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K. A. & Long, J. S. (Eds.), *Testing structural equation models*. *Beverly Hills, CA: Sage*, 111-135.

Buchanan, H. (2015). A made-in-Canada credential: Developing an evaluation professional designation. *Canadian Journal of Program Evaluation*, *29*(3), 33-53. https://doi.org/10.3138/cjpe.29.3.33

Byrne, B. M. (2016). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY:, Taylor & Francis Group: Routledge.

Canada" *in Evaluation Globe*, Americas, Palgrave Macmillan

Canadian Evaluation Society (n.d.). *CES Strategic Plan 2015-2018*. http://evaluationcanada.ca/2015-2018-strategic-plan. Accessed on August 10, 2016.

Caplow, T. (1966). The sequence of professionalization. *Professionalization*, 19-21.

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions: An empirical examination. *American Journal of Evaluation*, *28*(1), 8-25. https://doi.org/10.1177/1098214006298065

Christie, C. A., Quiñones, P., & Fierro, L. (2014). Informing the discussion on evaluator training: A look at evaluators' course taking and professional practice. *American Journal of Evaluation*, *35*(2), 274-290. https://doi.org/10.1177/1098214013503697

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: The New York Academy of Sciences.

Collins, R. (1990). Changing conceptions in the sociology of the professions. In Torstendahl, R. & Burrage, M. (Eds.) *The formation of professions: Knowledge, state and strategy*, 11-23. London, United Kingdom: Sage.

Consortium of Universities for Evaluation Education (CUEE) website (n.d.). Accessed on February 20, 2016.

Corbin, J., & Strauss, A. (2008). Strategies for qualitative data analysis. In: *Basics of Qualitative Research (3rd ed.): Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: SAGE Publications, Inc.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: SAGE Publications, Inc.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Coryn, C. L., Ozeki, S., Wilson, L. N., Greenman, G. D., Schröter, D. C., Hobson, K. A., ... & Vo, A. 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. https://doi.org/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.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, *10*(7), 1-9.

Cousins, J. B., Cullen, J., Malik, S., & Maicher, B. (2009). Debating professional designations for evaluators. *Journal of Multidisciplinary Evaluation*, *6*(11), 71-82.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Third edition. Thousand Oaks, CA: Sage publications, Inc., 200-224.

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage publications, Inc.

Creswell, J. W. (2016). Revisiting mixed methods and advancing scientific practices. In Hesse-Biber, S. N. & Johnson, R. B. (Eds.), *The Oxford handbook of multimethod and mixed methods research inquiry*, (pp. 1-24). London, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199933624.013.39

Creswell, J. W. (2016). *30 essential skills for the qualitative researcher*. Thousand Oaks, CA: SAGE Publications, Inc.

Creswell, J. W., & Clark, V. L. P. (2018). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297-334. https://doi.org/10.1007/BF02310555

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, *34*(4), 4-9. https://doi.org/10.1111/emip.12095

Davies, R., & MacKay, K. (2014). Evaluator training content and topic valuation in university evaluation courses. *American Journal of Evaluation*, *35*(3), 419-429. https://doi.org/10.1177/1098214013520066

DeVellis, R. F. (2012). Chapter 6: Factor Analysis. *Scale development: Theory and applications,* (Vol. 26). Sage Publications, New York, NY (2012).

Dewey, J. D., Montrosse, B. E., Schröter, D. C., Sullins, C. D., & Mattox, J. R. (2008). Evaluator competencies: What's taught versus what's sought. *American Journal of Evaluation*, *29*(3), 268-287. https://doi.org/10.1177/1098214008321152

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: John Wiley & Sons Inc.

Donaldson, S. I. (2019). Where do we stand? Recent AEA member views on professionalization. *Evaluation and program planning*, *72*, 152-161. https://doi.org/10.1016/j.evalprogplan.2018.10.009

Donaldson, S. I., & Christie, C. A. (2006). Emerging career opportunities in the transdiscipline of evaluation science. *Applied psychology: New frontiers and rewarding careers*, 243-260. https://doi.org/10.4324/9780203837603-18

Dumaine, F. (2015). A point of no return finally reached: The journey ahead. *The Canadian Journal of Program Evaluation*, *29*(3), 154-159. https://doi.org/10.3138/cjpe.29.3.154

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology*, *105*(3), 399-412. https://doi.org/10.1111/bjop.12046

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, *23*(4), 332-366. https://doi.org/10.1177/1088868318811759

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge handbook of expertise and expert performance*, *38*, 685-705. https://doi.org/10.1017/CBO9780511816796

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363-406. https://doi.org/10.1037/0033-295X.100.3.363

Etzioni, A. (1969). Preface. In A. Etzioni (ed.), The Semi-Professions and Their Organisation, v-xviii, New York, NY: Macmillan.

European Evaluation Society (2013). *Towards a Voluntary Evaluator Peer Review (VEPR) system for EES—Concept summary*. Retrieved from http://europeanevaluation.org/community/thematic-working-groups/twg-4-professionalization-evaluation.

EvalAgenda2020 & EvalPartners (2015). Global Evaluation Agenda 2016-2020: Developed by a global, multi-stakeholder consultative process. 15 January 2016.

Evetts, J. (2003). The sociological analysis of professionalism occupational change in the modern world. *International Sociology, 18*(2), 395-415. https://doi.org/10.1177/0268580903018002005

Feinberg, W., & Soltis, J. F. (2009). *School and society*. New York, NY: Teachers College Press.

Fetters, M. D., & Molina-Azorin, J. F. (2017). The Journal of Mixed Methods Research starts a new decade: The mixed methods research integration trilogy and its dimensions. *Journal of Mixed Methods Research,*11(3) 291-307. https://doi.org/10.1177/1558689817714066

Fierro, L. A., Galport, N., Hunt, A., Codd, H., & Donaldson, S. I. (2016). Canadian Evaluation Society credentialed evaluator designation program: Evaluation report, 1-74.

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *49*(2), 78-88. https://doi.org/10.1037/cbs0000069

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology*, 3(55), 1-21. https://doi.org/10.3389/fpsyg.2012.00055

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it. *European Journal of Psychological Assessment*, 33(6), 399–402. https://doi.org/10.1027/1015-5759/a000460

Forsyth, P. B., & Danisiewicz, T. J. (1985). Toward a theory of professionalization. *Work and occupations*, *12*(1), 59-76. Newbury Park, CA: Sage Publications, Inc. https://doi.org/10.1177/0730888485012001004

Freidson, E. (1970). *Profession of medicine*. New York, NY: Dodd, Mead.

Freidson, E. (1984). The changing nature of professional control. *Annual Review of Sociology*, 10:1-20. https://doi.org/10.1146/annurev.so.10.080184.000245

Freidson, E. (1994). *Professionalism reborn: Theory, prophecy, and policy*. Chicago, IL University of Chicago Press.

Freidson, E. (2001). *Professionalism, the third logic: On the practice of knowledge*. Chicago, IL: University of Chicago Press.

Frumkin, P., & Francis, K. (2008). Constructing Effectiveness: The Emergence of the Evaluation Research Industry. *LBJ School of Public Affairs, University of Texas*, 1-26.

Fullerton, K., Maloney, J. (2017). Evaluation for the public good or as a public good? Retrieved from https://www.senseandsustainability.net/2017/11/30/evaluation-for-the-public-good-or-as-a-public-good/.

Galport, N., & Azzam, T. (2017). Evaluator training needs and competencies: A gap analysis. *American Journal of Evaluation*, *38*(1), 80-100. https://doi.org/10.1177/1098214016643183

Gauger, P. G., Gruppen, L. D., Minter, R. M., Colletti, L. M., & Stern, D. T. (2005). Initial use of a novel instrument to measure professionalism in surgical residents. *The American Journal of Surgery*, *189*(4), 479-487. https://doi.org/10.1016/j.amjsurg.2004.09.020

Gauthier, B., Borys, S., Kishchuk, N., & Roy, S. (2006). Evaluation practice in Canada: Results of a national survey. *The Canadian Journal of Program Evaluation*, *21*(3), 1-42.

Gauthier, B., Kishchuk, N., Borys, S., & Roy, S. N. (2015). The CES professional designations program: Views from members. *Canadian Journal of Program Evaluation*, *29*(3), 98–133. https://doi.org/10.3138/cjpe.29.3.98

Gauthier, B., Lahey, R.E., & Jacob, S. (forthcoming) “The institutionalization of evaluation in
Gispen, C. W. R. (1988). German engineers and American social theory: historical perspectives on professionalization. *Comparative Studies in Society and History*, *30*(3), 550-574.
Goode, W. J. (1960). Encroachment, charlatanism, and the emerging professions: Psychology, medicine, and sociology. *American Sociological Review*, *25(6)*, 902-914.
Goode, W. J. (1969). The theoretical limits of professionalization. In A. Etzioni (Ed.), *The semi-professions and their organization*, 266-313, New York, NY: Macmillan.
Groß, C., & Kieser, A. (2006). Are consultants moving towards professionalization? *Research in the Sociology of Organizations*, *24*(1), 69-100. https://doi.org/10.1016/S0733-558X(06)24004-7
Gullickson, A. M., King, J. A., LaVelle, J. M., & Clinton, J. M. (2019). The current state of evaluator education: A situation analysis and call to action. *Evaluation and program planning*, *75*, 20-30. https://doi.org/10.1016/j.evalprogplan.2019.02.012
Hall, R. (1968). Professionalization and Bureaucratization. *American Sociological Review*, *33*(1), 92-104. Retrieved April 15, 2020, from www.jstor.org/stable/2092242
Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. New York, NY: Guilford Press.
Hirschfeld, G., & Von Brachel, R. (2014). Improving Multiple-Group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation*, *19*(1), 1-12. https://doi.org/10.7275/qazy-2946
Hogan, R. L. (2007). The historical development of program evaluation: Exploring past and present. *The Online Journal for Workforce Education and Development.*, *2*(4), 1-14.
Hoskins, K. (1979). The examination, disciplinary power and rational schooling. History of Education, 8(1), 135-146. https://doi.org/10.1080/0046760790080205
House, E. R. (1990). Trends in evaluation. *Educational Researcher*, *19*(3), 24-28. https://doi.org/ 10.3102/0013189X019003024
House, E. R. (1994). The future perfect of evaluation. *Evaluation Practice*, *15*(3), 239-247. https://doi.org/10.1177/109821409401500303
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118
Hunter, T. and McDavid J.C. (In press) *Inventory of Canadian Graduate Evaluation Education*, Consortium of Universities for Evaluation Education.
International Organization for Cooperation in Evaluation (IOCE) website (n.d.). https://www.ioce.net/. Accessed on February 20, 2016.
Irby, D. M., & Hamstra, S. J. (2016). Parting the clouds: three professionalism frameworks in medical education. *Academic Medicine*, *91*(12), 1606-1611. https://doi.org/10.1097/ACM.0000000000001190

Jacob, S., & Boisvert, Y. (2010). To be or not to be a profession: Pros, cons and challenges for evaluation. *Evaluation*, *16*(4), 349-369. https://doi.org/10.1177/1356389010380001

Johnson, S. M. (2018). *The development and use of interpersonal competencies by evaluators* (Doctoral dissertation, University of Minnesota). http://hdl.handle.net/11299/194558.

Johnson, T. J. (1972). *Professions and Power*. London, United Kingdom: Macmillan

Johnson, T. J. (1972). *The rise of professionalism: A sociological analysis*. Berkeley, CA: University of California Press.

Kay, D., Berry, A., & Coles, N. A. (2019). What Experiences in Medical School Trigger Professional Identity Development? *Teaching and learning in medicine*, *31*(1), 17-25. https://doi.org/10.1080/10401334.2018.1444487

Kelley, K. A., Stanke, L. D., Rabi, S. M., Kuba, S. E., & Janke, K. K. (2011). Cross-validation of an instrument for measuring professionalism behaviors. *American Journal of Pharmaceutical Education*, *75*(9), 1-10. https://doi.org/10.5688/ajpe759179

King, J. A. (2003). Evaluating educational programs and projects in the USA. In T. Kellaghan, & D. L. Stufflebeam (Eds.), *International handbook on educational evaluation*, Dordrecht: Kluwer Academic Publishers, https://doi.org/721-732. 10.1007/978-94-010-0309-4_41

King, J. A. (2013, August). *Essential Competencies for Program Evaluators*. Workshop presented at the Summer Institute of the American Evaluation Association, Atlanta, GA.

King, J. A. (2015). From the outside, looking in with a smile: A summary and discussion of CES's Credentialed Evaluator Designation. *The Canadian Journal of Program Evaluation*, *29*(3), 134–153. https://doi.org/10.3138/cjpe.29.3.134.

King, J. A., & Stevahn, L. (2015). Competencies for program evaluators in light of adaptive action: What? So wWhat? Now wWhat? *New Directions for Evaluation*, *2015*(145), 21-37. https://doi.org/10.1002/ev.20109

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. https://doi.org/10.1177/109821400102200206

Kirkhart, K. (1981). Defining evaluator competencies: New light on an old issue. *American Journal of Evaluation*, *2*(2), 188-192. https://doi.org/10.1177/109821408100200225

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford publications.

Kuji-Shikatani, K. (2015). Credentialed evaluator designation program, the Canadian experience. *New Directions for Evaluation*, *2015*(145), 71-85. https://doi.org/10.1002/ev.20112

Kuji-Shikatani, K., Thompson, M., & Matthew, M. (2015). Launching the credentialed evaluator (CE) designation. *Canadian Journal of Program Evaluation*, *29*(3), 70-85. https://doi.org/10.3138/cjpe.29.3.33

Larson, M. S. (1977). *The rise of professionalism: A sociological analysis*. Berkeley, CA: University of California Press.

LaVelle, J. M. (2014). *An examination of evaluation education programs and evaluator skills across the world* (Doctoral dissertation, The Claremont Graduate University), Claremont, CA.

LaVelle, J. M., & Donaldson, S. I. (2015). The state of preparing evaluators. *New Directions for Evaluation*, 145, 39-52. https://doi.org/10.1002/ev.20110

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods* Thousand Oaks, CA: Sage Publications, Inc. https://doi.org/10.4135/9781412963947.

Leach, N., Barrett, K., & Morgan, G. (2011). *IBM SPSS for intermediate statistics: Uses and interpretation*. New York, NY: Routledge, Taylor & Francis Group.

Leavy, P. (2017). *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. New York, NY: Guilford Publications.

Lee, Y. F., Altschuld, J. W., & Lee, L. S. (2012). Essential competencies for program evaluators in a diverse cultural context. *Evaluation and Program Planning*, *35*(4), 439-444. https://doi.org/10.1016/j.evalprogplan.2012.01.005

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Love, A. J. (1994). Should evaluators be certified? *New Directions for Program Evaluation*, 62, 29-40. https://doi.org/10.1002/ev.1674

Madaus, G. F., Scriven, M. S., & Stufflebeam, D. L. (1983). *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*. Dordrecht, Netherlands: Springer Netherlands.

Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (Eds.). (2012). *Evaluation models: Viewpoints on educational and human services evaluation* (Vol. 6). Springer Science & Business Media. Boston: Kluwet-Nijhoff Publishing.

Maicher, B., & Frank, C. (2015). The development and initial validation of competencies and descriptors for Canadian evaluation practice. *Canadian Journal of Program Evaluation*, *29*(3), 54-69. https://doi.org/10.3138/cjpe.29.3.54

Matsunaga, M. (2010). How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-To's. *International journal of psychological research*, *3*(1), 97-110. https://doi.org/10.21500/20112084.854

McDonald, R. P. (1999). Test theory: a unified treatment. Lawrence Earlbaum Associates. *Inc., Mahwah, NJ*, 142-145.

McGuire, M., & Zorzi, R. (2005). Evaluator competencies and performance development. *Canadian Journal of Program Evaluation*, *20*(2), 73-99.

McNeish, D. (2017). Exploratory factor analysis with small samples and missing data. *Journal of personality assessment*, *99*(6), 637-652.

Mertens, D. M. (1994). Training evaluators: Unique skills and knowledge. In J. W. Altschuld & M. Engle (Eds.), *New Directions for Program Evaluation: No. 62. The preparation of professional evaluators: Issues, perspectives, and program* (pp. 17–28). San Francisco, CA: Jossey-Bass. https://doi.org/10.1002/ev.1673

Mertens, D. M. (2015). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Los Angeles, CA: Sage Publications, Inc.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Los Angeles, CA: Sage Publications, Inc.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. (3rd ed). Thousand Oaks, CA: Sage Publications, Inc.

Millerson, G. (1964). Dilemmas of professionalism. *New Society*, *4*, 15-18.

Modarresi, S. (1998). The profession of program evaluation: theory and practice. (Unpublished doctoral dissertation submitted to the University at Albany). State University of New York, New York, NY.

Modarresi, S., Newman, D. L., & Abolafia, M. Y. (2001). Academic evaluators versus practitioners: Alternative experiences of professionalism. *Evaluation and Program Planning*, *24*(1), 1-11. https://doi.org/10.1016/S0149-7189(00)00041-0

Moore, T. S., Lapan, S. D., & Quartaroli, M. T. (2012). Case study research. In Lapan, S. D., Quartaroli, M. T., & Riemer, F. J. (Eds.). *Qualitative research: An introduction to methods and designs* (Vol. 37), 243-270. San Francisco, CA: John Wiley & Sons.

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. https://doi.org/10.1016/0149-7189(78)90003-4

Muthén, L. K., & Muthén, B. O. (2017). Mplus (Version 8) [computer software]. (1998–2017). *Los Angeles, CA: Muthén & Muthén*.

Muthén, L. K., & Muthén, B. O. (2019). Mplus home page.

Newsom, J. (2012). Practical approaches to dealing with non-normal and categorical variables. Viewed January 2018, Psy 523/623 Structural Equation Modeling.

Nunnally, J. C., & Bernstein, I. H. (1994). Validity. *Psychometric theory*, *3*, 99-132.

O'Neill, O. A., & O'Reilly III, C. A. (2011). Reducing the backlash effect: Self-monitoring and women's promotions. *Journal of Occupational and Organizational Psychology*, *84*(4), 825-832. https://doi.org/10.1111/j.2044-8325.2010.02008.x

OECD, D. (2015). DAC evaluation quality standards: Draft for test application phase. *DAC Network on Development Evaluation*.

Ofir, Z. (2016). Grand Challenges for Evaluation? Part 2. http://zendaofir.com/grand-challenges-evaluation-part-2/. Evaluation for Development Blog.

Osborne, J. W. (2008). *Best practices in quantitative methods*. Thousand Oaks, CA: Sage Publications.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. *Best practices in quantitative methods*, 86–99. Los Angeles, CA: Sage Publications

Patton, L. D., & Bondi, S. (2015). Nice white men or social justice allies? Using critical race theory to examine how white male faculty and administrators engage in ally work. *Race Ethnicity and Education*, *18*(4), 488-514. https://doi.org/10.1080/13613324.2014.1000289

Patton, M. Q. (1990). The challenge of being a profession. *Evaluation Practice*, *11*(1), 45-51. https://doi.org/10.1177/109821409001100107

Patton, M. Q. (2013). The future of evaluation in society: Top ten trends plus one. In S. Donaldson (Ed.), *The future of evaluation in society: A tribute to Michael Scriven*, 45–62. Charlotte, NC: Information Age Publishing.

Patton, M. Q. (2015). Qualitative research and methods: Integrating theory and practice. *Thousand Oaks, CA: SAGE Publications*.

Pavalko, R. M. (1988). *Sociology of occupations and professions*, (2nd ed.). Itasca, IL: F. E. Peacock Publishers, Inc.

Peersman, G. & Rogers, P., (2017). *Pathways to advance professionalisation within the context of the Australasian Evaluation Society.* Melbourne: ANZSOG/BetterEvaluation.

Picciotto, R. (2011). The logic of evaluation professionalism. *Evaluation, 17*(2), 165-180. https://doi.org/10.1177/1356389011403362

Podems, D. (2014). Evaluator competencies and professionalizing the field: Where are we now? *Canadian Journal of Program Evaluation*, *28*(3), 127-136.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in nursing & health*, *30*(4), 459-467. https://doi.org/10.1002/nur.20199

Pornprasertmanit, S., Miller, P., Schoemann, A., Rosseel, Y., Quick, C., Garnier-Villarreal, M., Longo, Y. (2016). semTools: Useful tools for structural equation modeling. R package version 0.4-11.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*(3), 354. https://doi.org/10.1037/a0029315

Richey, R. C., Fields, D. C., & Foxon, M. (2001). *Instructional design competencies: The standards* (3rd ed.). Syracuse, NY: ERIC Clearinghouse on Information & Technology.

Ritzer, G. (1975). Professionalization, bureaucratization and rationalization: The views of Max Weber. *Social Forces*, *53*(4), 627-634.

Rodenberg, C., Kuznicki, J. T., & Yiu, G. H. (2012). Instrument Development and Validation. In Chow Shein-Chung (Ed.), *Encyclopedia of biopharmaceutical statistics*, 1, 635-645. New York, NY: Informa Healthcare

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Rossi, P. H., & Freeman, H. E. (1985). Evaluation: A systematic approach, (3rd ed.). Beverley Hills, CA: Sage Publications, Inc.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach*. Newbury Park, CA: Sage Publications, Inc.

Rugh, J., & Segone, M. (2013). *Voluntary Organizations for Professional Evaluation (VOPEs): Learning from Africa, Americas, Asia, Australasia, Europe and Middle East*. EvalPatners and IOCE.

Russ-Eft, D., & Preskill, H. (2009). *Evaluation in organizations a systematic approach: A systematic approach to enhancing learning, performance, and change*. New York: Basic Books.

Schulberg, H. C., & Perloff, R. (1979). Academia and the training of human service delivery program evaluators. *American Psychologist*, *34*(3), 247-254. https://doi.org/10.1037/0003-066X.34.3.247

Schwandt, T. A. (2017). Professionalization, ethics, and fidelity to an evaluation ethos. *American Journal of Evaluation*, *38*(4), 546-553. https://doi.org/10.1177/1098214017728578

Sciulli, D. (2005). Continental sociology of professions today: Conceptual contributions. *Current sociology*, *53*(6), 915-942. https://doi.org/10.1177/0011392105057155

Scriven, M. (1996). Types of evaluation and types of evaluator. *American Journal of Evaluation*, *17*(2), 151-161. https://doi.org/10.1177/109821409601700207

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage Publications, Inc.

Secretariat, T. B. (2005). Improving the Professionalism of Evaluation. http://www.tbs-sct.gc.ca/cee/prof-eval-eng.asp.

Seidling, M. B. (2015). Evaluator certification and credentialing revisited: A survey of American Evaluation Association members in the United States. *New Directions for Evaluation*, *2015*(145), 87-102. https://doi.org/10.1002/ev.20113

Shackman, G. (2015). Accreditation, Certification, Credentialing: Does It Help? *New Directions for Evaluation*, *2015*(145), 103-113. https://doi.org/10.1002/ev.20114

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

Shafer, H. B. (1936). *The American Medical Profession: 1783 to 1850* (No. 417). New York, NY: Columbia University Press.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in experimental social psychology*, *38*, 183-242. https://doi.org/10.1016/S0065-2601(06)38004-5

Smelser, N. J., & Baltes, P. B. (Eds.). (2001). *International encyclopedia of the social & behavioral sciences* (Vol. 11). Amsterdam, the Netherlands: Elsevier.

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. https://doi.org/10.1177/109821409902000311

Sou, G. (2008). Historical Review of Program Evaluation for Educational Quality Assurance. School of Graduate Studies Asia International Open University Macau Special Administrative Region, People's Republic of China.

Starr, P. (2008). *The social transformation of American medicine: The rise of a sovereign profession and the making of a vast industry*. New York, NY: Basic books, Inc., Publishers.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation*, *26*(1), 43-59. https://doi.org/10.1177/1098214004273180

Stockmann, R., & Meyer, W. (2016). The Future of Evaluation: Global Trends, New Challenges and Shared Perspectives. In *The Future of Evaluation* (pp. 9-19). London, United Kingdom: Palgrave Macmillan. https://doi.org/10.1057/9781137376374_1

Stufflebeam, D. L. (2001). Interdisciplinary Ph. D. programming in evaluation. *American Journal of Evaluation*, *22*(3), 445-455. https://doi.org/10.1177/109821400102200323

Stufflebeam, D.L. (2000). The CIPP model for evaluation. In D.L. Stufflebeam, G.F. Madaus, & T. Kellaghan (Eds.), Evaluation models: *Viewpoints on educational and human services evaluation* (2nd ed., pp. 279-318). Boston, MA: Kluwer Academic.

Tashakkori, A., & Creswell, J. W. (2007a). Editorial: The new era of mixed methods. Journal of Mixed Methods Research, 1(3), 3-7. https://doi.org/ 10.1177/2345678906293042

Tashakkori, A., & Creswell, J. W. (2007b). Exploring the nature of research questions in mixed methods research. *Journal of Mixed Methods Research*, 1(3), 207-211. https://doi.org/10.1177/1558689807302814

Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and evaluation in psychology and education* (8^{th} ed.). Boston, MA: Pearson Education, Inc.

Trizano-Hermosilla, I., & Alvarado, JM (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in psychology*, 7(769), 1-8. https://doi.org/10.3389/fpsyg.2016.00769

Trotter II, R. T. (2012). Qualitative research sample design and sample size: Resolving and unresolved issues and inferential imperatives. *Preventive medicine*, *55*(5), 398-400. 10.1016/j.ypmed.2012.07.003

United Nations General Assembly (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development.* Resolution adopted by the General Assembly on 25 September 2015. New York: United Nations, http://www.un.org/ga/search/view_doc.asp. E [Accessed 2016-12-01].

Van de Ven, A. H. (2007). *Engaged scholarship: A guide for organizational and social research*. Oxford, United Kingdom: Oxford University Press on Demand.

Van der Krogt, T. (2015). Professionals and their work. Blog book posted to https://profqual.wordpress.com. Accessed on November 2, 2016

Verma, J. P., & Abdel-Salam, A. S. G. (2019). *Testing statistical assumptions in research*. Hoboken, NJ: John Wiley & Sons.

Waddington, I. (1990). The movement towards the professionalization of medicine. *BMJ: British Medical Journal*, *301*(6754), 688-690. https//doi.org/10.1136/bmj.301.6754.688

Wilensky, H. L. (1964). The professionalization of everyone? *American journal of sociology*, *70*(2), 137-158. https://doi.org/10.1086/223790

Worthen B.R. (2003). How can we call Evaluation a Profession if there are no Qualifications for Practice? In: Kellaghan T., Stufflebeam D.L. (eds) *International Handbook of Educational Evaluation*. Kluwer International Handbooks of Education, vol 9. Dordrecht, Netherlands: Springer Netherlands. https://doi.org/10.1007/978-94-010-0309-4_20

Worthen, B. R. (1994). Is evaluation a mature profession that warrants the preparation of evaluation professionals? *New Directions for Program Evaluation*, 62, 3-15. https://doi.org/10.1002/ev.1672

Worthen, B. R. (2001). Whither evaluation? That all depends. *American Journal of Evaluation*, *22*(3), 409-418. https://doi.org/10.1177/109821400102200319

Worthen, B. R. (2001). Whither evaluation? That all depends. *American Journal of Evaluation*, *22*(3), 409-418. https://doi.org/10.1177/109821400102200319

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

Yin, R. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Yin, R. K. (2009). *Case study research: Design and methods*. Thousand Oaks, CA: Sage Publications, Inc.

Zeraatkar, D., Obeda, M., Ginsberg, J. S., & Hirsh, J. (2017). The development and validation of an instrument to measure the quality of health research reports in the lay media. *BMC Public Health*, *17*(343), 1-10. https://doi.org/10.1186/s12889-017-4259-y

APPENDICES

Appendix A: Interview Consent Form

for
Dissertation Research
University of Minnesota, College of Education
Organizational Leadership and Policy Development
Doctoral Candidate: Sandra Ayoo

Study: Professionalization of Program Evaluation: A Comparative Mixed Methods Case Study of Canada and the United States

What is the Research?
You have been asked to take part in a research study that is being conducted by Sandra Ayoo, a Doctoral Candidate at the University of Minnesota. This study will be used to complete the doctoral requirements for the research component of the Evaluation Studies curriculum. This information will also be beneficial to the field of evaluation practice. The IRB has determined that my STUDY00000797 is not human research.

Why have I been asked to take part?
You have been invited to participate because you are a practicing evaluation professional and a member of a professional association of evaluators.

Voluntary participation
This discussion is voluntary, and you do not have to take part if you do not want to. This project is not related to or sponsored by any national, state or professional organization, including the American Evaluation Association and the Canadian Evaluation Society. If you do not take part, it will have no effect on your relationships with your professional association. You may stop responding to the interview questions at any time, and you are not obligated to answer any questions.

Risks and Benefits
There are no known risks involved in taking part in this interview. There are no personal risks or benefits. Your insights are helpful to this topic, and it could contribute to how the field is evolving as a mature profession.

Audio Recording
With your permission, this interview will be audio recorded to ensure that your comments are accurately captured. Your privacy will be protected, and your name will not be used in any report. This discussion will be kept strictly confidential, and this recording will only be available to the interviewer. The recordings will be stored securely and erased when the analysis is completed.

If you agree to this consent form, please check yes and sign
_______ Yes, I agree to take part in this interview.
Name ____
Signature ____
Date ____

Appendix B: Semi-structured Interview Protocol

Name of Study: Professionalization of Program Evaluation: A Comparative Mixed Methods Case Study of Canada and the United States

Date:
Time:
Location: Minneapolis
Type of Participant: Evaluation Practitioner
Interviewer Name: Sandra Ayoo

Agenda

Welcome: *Good morning/afternoon and welcome. Thank you for taking the time out of your busy schedule to participate in this interview. My name is Sandra Ayoo, and I am a doctoral candidate at the University of Minnesota. You will be providing me information necessary in completing my doctoral requirements*.

Overview of the topic: *My study is investigating the development of program evaluation as a profession in Canada and in the United States. The questions are designed to get your insights, perceptions, experiences, and opinions on the key constructs in professionalism in program evaluation. The purpose of this interview is to get conceptual clarity on professionalism constructs and to generate essential information that I will use for modifying an instrument for measuring professionalism in program evaluation. I shall administer the proposed instrument as an online survey to professional evaluators who are members of American Evaluation Association (AEA) and the Canadian Evaluation Society (CES).*

Introduction

(Probes: Tell me more. How is that? In what ways? Anything else?)

Q1. How long have you worked as an evaluator and in what capacity (internal/external, educator)?
Q2. Tell me about how you became an evaluator? What was the turning point in your career pathway to become an evaluator, if any?
Q3. Think back to the time you first started as an evaluator. What changes do you see in the field of program evaluation today? What has changed?

Professionalism in Evaluation

Q4. *Professional Autonomy or Professionals Controlling their Work*: The sociology of professions literature proposes that professionals should have autonomy in making technical decisions. It proposes that evaluators could claim authority to insulate themselves from undue interferences in making technical decisions like in deciding on appropriate methods, developing evaluation plans, data analysis, and making recommendations. Professional autonomy i.e. self-management of the occupation:

collective controls over designation, behavioral standards, quality assurance criteria buttressed by control mechanisms and disciplinary processes.

a. To what extent do you feel that you have autonomy in your professional role as an evaluator?
b. Did you always have autonomy in your practice or did it emerge, if at all?
c. Are there barriers or challenges limiting your professional autonomy as an evaluator?

[Probe for ways of interaction with clients and/or employers]

Q5.*Expertise or Professional Development is Linked with Specialization*: "*No one would willingly agree to have one's heart surgery performed by a plumber*." This implies program evaluation is not ordinary work. Theory proposes that professionals are expected to get expertise through undergoing specialized education at the tertiary level, followed by substantial exposure to expert practice, and keeping up-to-date with new knowledge to help them make sound judgements. Demonstrated expertise and judgment that flow from access to high quality tertiary education; extended exposure to practice; theoretical knowledge; mastery of techniques;

a. What types of formal training were you involved in to develop expertise in program evaluation?
b. What types of continuing education have you engaged in to keep current and/or develop new knowledge and skills? Please explain or give examples.
c. What are some of the challenges in keeping up-to-date with new knowledge and developments in the field, if any?

Q6.*Ethical Dispositions Entails Orientation Towards the Public Interest*: Theory proposes that professionals should have commitment to a life-long career, collegial behavior, occupational solidarity, peer review, and responsibility for the quality of one's work. Ethical dispositions which implies assuming responsibility for the quality of one's work, orientation to the public interest; loyalty to the group; collegial behavior, compliance with professional guidelines and principles.

Traditional professionals like physicians, engineers, and clergy have ethical standards to guide their practice and to protect the public.

a. In your opinion, what are examples of ethical standards that an evaluator should have?
b. What is an example of the most difficult ethical challenge or ethical dilemma you faced as an evaluator, if any, and how did you resolve it?

Q7.*Innovation, Research and Professional Engagement:* Theory proposes institutionalization of research and intellectual innovation by members of the profession to develop new knowledge, skills, ideas, and new practical solutions to advance the discipline. For example, EvalAgenda2020 is inviting innovations from evaluators worldwide to institutionalize expert knowledge in the areas of evaluation methods and capacity-building, peer-to-peer exchanges, use of technology for communications, scale-up and involvement of new partners.

a. Over the course of your career, have you identified a new method or approach for conducting evaluation? (Yes/No). (E.g., use of technology in data collection, data management, engaging stakeholders)
b. If yes, please describe ways you think your method was innovative. How might it benefit evaluators or change the quality of evaluations?
c. In what ways are you engaging in research to promote the field of evaluation?
d. In what ways are you professionally engaged with the field of program evaluation?

Q8. *Credentialing of Program Evaluators is Linked with Tested Performance*: Credentialed designation is earned by a person qualified to perform a job or task by a professional body to safeguard the public interest. Credentialing confirms proof of completion of specified training and experience requirements.

a. Are you familiar with CES' current work on credentialing for Canadian evaluators?
b. If yes, what did it take to start the credentialing effort in Canada?
c. In your opinion, what are the possible pros and cons of credentialing evaluators?
d. What influence does credentialing have on (a) evaluators and (b) the clients they serve?

Conclusion

Q9. In thinking about the changes in the field, what do you see as the future of program evaluation in Canada?

Q10. Is there anything else you would like to share from your experience or insight about how the field of program evaluation is evolving in Canada?

Thank you very much for your time today.

Appendix C: Professionalism Survey

I have read, understood, and printed a copy of the above invitation and consent letter and desire of my own free will to participate in this study.

○ Yes

○ No

Skip To: End of Survey If I have read, understood, and printed a copy of the above invitation and consent letter and desire... = No

End of Block: Consent Form

Start of Block: Expertise

Q1 Expertise: This domain focuses on how you acquire and keep up-to-date with knowledge to help practice program evaluation as a field of professional practice, for example, through specialized graduate education in evaluation, substantial exposure to expert practice, mastery of techniques, and theoretical knowledge. How strongly do you agree or disagree, if at all, with each statement?

	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
I self-assess to identify my professional strengths as an evaluator.	○	○	○	○	○
I learn from evaluation thought leaders in areas of interest.	○	○	○	○	○
My approach to professional development is to learn by doing.	○	○	○	○	○
I improve my skills in evaluation through reading on my own.	○	○	○	○	○
I developed my skills by working with other professional evaluators as an apprentice (e.g., internship, practicum).	○	○	○	○	○
I strengthen my expertise in program evaluation via on-the-job training.	○	○	○	○	○

I keep up-to-date informally by meeting occasionally with other evaluators.	◯	◯	◯	◯	◯
I keep up-to-date formally by attending evaluation conferences.	◯	◯	◯	◯	◯
I keep up-to-date formally by attending professional development workshops.	◯	◯	◯	◯	◯
I keep up-to-date formally by integrating empirical research on evaluation.	◯	◯	◯	◯	◯
I keep current by reading evaluation-related journals.	◯	◯	◯	◯	◯
I seek feedback on my work from evaluation colleagues.	◯	◯	◯	◯	◯
I incorporate feedback from my colleagues to make changes in my professional behavior.	◯	◯	◯	◯	◯
I have people I can ask if I have questions about my practice.	◯	◯	◯	◯	◯
I am part of a group of evaluators that meets periodically to learn from each other.	◯	◯	◯	◯	◯
I refer clients to someone who is an expert in an evaluation approach if I am not an expert in it.	◯	◯	◯	◯	◯
I submit manuscripts to evaluation journals.	◯	◯	◯	◯	◯

Please provide comments you have on any of the above items (for example changes, revisions, deletions, additions). Please refer to each item by number.

____

____

____

End of Block: Expertise
Start of Block: Innovation and Research.

Q2 **Innovation and Research**: This domain focuses on how you are engaged in developing new theoretical and practical solutions to advance the field of evaluation, for example, in developing new evaluation methods, peer-to-peer exchanges, interdisciplinary networking, and use of technology. How strongly do you agree or disagree, if at all, with each statement?

	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
I purposefully generate new evaluation knowledge to benefit the field.	○	○	○	○	○
I apply knowledge from other disciplines to my evaluation work.	○	○	○	○	○
I advance the profession by sharing reflections on my own practice.	○	○	○	○	○
I modify methods created by other people for use in my own practice.	○	○	○	○	○
I share my insights at professional conferences.	○	○	○	○	○
I apply established methods to novel situations.	○	○	○	○	○
I participate in formal research studies on evaluation.	○	○	○	○	○
I rarely prepare traditional evaluation reports.	○	○	○	○	○

Please provide comments you have on any of the above items (for example changes, revisions, deletions, additions). Please refer to each item by number.

____
____
____

End of Block: Innovation and Research.

Start of Block: Ethical Dispositions

Q3 **Ethical Dispositions**: This domain focuses on how you are taking responsibility for the quality of your work as an evaluator, for example, your commitment to a life-long career, collegial behavior, loyalty to your community of practice, peer review, and compliance with professional guidelines and guiding principles. How strongly do you agree or disagree, if at all, with each statement?

	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
I provide stakeholders with the best possible information to improve their programs.	○	○	○	○	○
I disclose my personal interests when they may influence an evaluation.	○	○	○	○	○
I am upfront about my own values when people hire me.	○	○	○	○	○
I tell my clients that I follow a professional code of ethics to guide my work.	○	○	○	○	○
I respect the autonomy of program participants to freely provide evaluation data.	○	○	○	○	○
I am concerned about how evaluation decisions affect program staff.	○	○	○	○	○
I am concerned about how evaluation decisions affect program participants.	○	○	○	○	○
I provide respondents with information on how participating in evaluation could be beneficial to them.	○	○	○	○	○
I believe that the most important ethical standard is the integrity of the evaluation process.	○	○	○	○	○
I believe that transparency in terms of ethical values is important.	○	○	○	○	○
I minimize respondents' discomfort in participating in evaluation.	○	○	○	○	○

My first ethical standard is to do no harm to those who participate in an evaluation.	○	○	○	○	○
I use the program evaluation standards as a safeguard in every evaluation.	○	○	○	○	○
I turn down jobs where I think that the client is unwilling to hear whatever might be found.	○	○	○	○	○

Please provide comments you have on any of the above items (for example changes, revisions, deletions, additions). Please refer to each item by number.

____

____

____

____

____

End of Block: Ethical Dispositions

Start of Block: Professional Autonomy

Q4 **Autonomy**: This domain focuses on how you self-manage to prevent undue interferences in making discretionary judgment in evaluation to ensure high-quality evaluations, for example, when interacting with clients, commissioners of evaluators, employers, and your fellow evaluators. How strongly do you agree or disagree, if at all, with each statement?

	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
If a client tries to get me to do something that I don't believe in, I won't do it.	○	○	○	○	○
I would turn down evaluation work when I don't agree with my client's intended approach.	○	○	○	○	○
I must decide which battles to fight if I want to be in the evaluation business.	○	○	○	○	○
I have autonomy to decide what evaluation projects to take on.	○	○	○	○	○
I can decide what sort of evaluation method to employ.	○	○	○	○	○
I perform evaluation activities based on my client's consent with the evaluation methods.	○	○	○	○	○
Ultimately my concern with my clients is to make methodologically-sound rather than popular decisions.	○	○	○	○	○
I can end my participation in a project if the evaluation is not going well.	○	○	○	○	○
Because my clients are knowledgeable about their program, I believe they should participate in evaluation decisions.	○	○	○	○	○
I always seek the broader context in framing an evaluation.	○	○	○	○	○
I can tell people that, based on my professional training and experience, the approach I	○	○	○	○	○

identify is the best for a given
evaluation.

Please provide comments you have on any of the above items (for example changes, revisions, deletions, additions). Please refer to each item by number.

____

____

____

____

____

End of Block: Professional Autonomy

Start of Block: Credentialing

Q5 **Credentialing**: This section is on your perceptions on credentialing. A credentialed designation is earned by a person deemed to be qualified to perform a job or task by a professional body to safeguard the public. How strongly do you agree or disagree, if at all, with each statement?

	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
Being credentialed would encourage me to reflect on my skills.	○	○	○	○	○
Credentialing would direct my professional development.	○	○	○	○	○
Credentialing would tell people I have reached the minimum standards to be a competent evaluator.	○	○	○	○	○
Evaluator competencies should be aligned to credentialing.	○	○	○	○	○
Credentialing has made me aware of evaluator competencies.	○	○	○	○	○
Credentialing protects stakeholders from hiring someone who does not have enough of an evaluation background to do a good job.	○	○	○	○	○
Being a credentialed evaluator would mean that my peers have recognized my expertise.	○	○	○	○	○
Credentialing assures me that those who engage in evaluation conform to certain professional standards.	○	○	○	○	○
Evaluator competencies should provide a framework for teaching evaluation.	○	○	○	○	○

Please provide comments you have on any of the above items (for example changes, revisions, deletions, additions). Please refer to each item by number.

____

____

____

____

____

End of Block: Credentialing

Start of Block: Your Demographic Information

Q1 **How many years have you been involved in program evaluation work (for example, planning studies, collecting data, analyzing data, interpreting results, providing consultation, assisting in evaluations)?** *Select one*.

○ Not applicable (I do not consider myself to be an evaluator)

○ Less than 1 year

○ 1 - 5 years

○ 6-10 years

○ 11-15 years

○ 16 - 20 years

○ 21 - 25 years

○ 26 - 30 years

○ More than 30 years

Q2 **Do you currently practice evaluation full-time?** *Select one*.

○ Yes

○ No

Q3 **If yes, please select your area of expertise in evaluation.** *Select all that apply.*

▢ Developmental

▢ Empowerment

▢ Impact

▢ Meta-evaluation

▢ Utilization focused

- [ ] Theory driven
- [ ] Realist
- [ ] Collaborative
- [ ] Responsive
- [ ] Culturally responsive
- [ ] Ripple effect mapping
- [ ] Other (please specify) ____

Q4 **In what professional contexts does your evaluation work currently take place?** *Select all that apply*.

- [ ] PreK-12 education
- [ ] Higher education
- [ ] Business/industry
- [ ] Health care/public health
- [ ] Foundations
- [ ] Government (federal/state/provincial/municipal/local)
- [ ] Nonprofit/social service
- [ ] Private sector/consultant
- [ ] International development
- [ ] Other (please specify) ____

Q5 **Where does most of your work related to program evaluation take place?** *Select one*.

○ Mostly within Canada

○ Mostly outside Canada

○ Equally within and outside Canada

○ Mostly within the United States of America

○ Mostly outside the United States of America

○ Equally within and outside the United States of America

○ Other (please specify) ____

Q6 **How would you describe your skill level as an evaluator?** *Select one*.

○ Not applicable (I do not consider myself to be an evaluator)

○ Novice/entry-level evaluator

○ Proficient/skilled evaluator

○ Expert/master evaluator

○ Other (please specify) ____

Q7 **How have you gained professional knowledge as an evaluator?** *Select all that apply*.

☐ University training/courses

☐ National/international professional association conferences/training/workshops

☐ Regional/local professional association conferences/training/workshops

☐ Job-embedded professional development

☐ Informal on-the-job learning

☐ Professional literature (e.g., books, journals, online resources)

☐ Other (please specify) ____

Q8 **Do you currently hold a CES Credentialed Evaluator designation?**

○ Yes

○ No

○ I intend to apply

○ I prefer not to respond

Display This Question:
If Do you currently hold a CES Credentialed Evaluator designation? = No

Q9 If no, why not?

____
____
____
____
____

Q10 **What is your highest degree earned?** *Select one*.

○ High School Diploma

○ Associate's degree

○ Bachelor's degree

○ Master's degree

○ Doctoral degree

○ Other (please specify) ____

Q11 **Which best describes your current job title/position?** *Select all that apply*.

▢ Director of an organization/institution

▢ Director of a program in an organization/institution

▢ Program provider/staff within an organization/institution

▢ Director of evaluation/assessment/research in an organization/institution

▢ Staff assisting in evaluation/assessment/research in an organization/institution

▢ Consultant providing services in evaluation/assessment/research

▢ University educator teaching evaluation/assessment/research

▢ Scholar/Theorist/Researcher in evaluation/assessment/research

☐ Graduate student in evaluation/assessment/research

☐ Other (please specify) ____

Q12 **Which Evaluation Association do you most identify with?** *Select one*.

○ Canadian Evaluation Society

○ American Evaluation Association

○ Canadian Evaluation Society and American Evaluation Association

○ Lapsed Membership/Non-member

○ Other (please specify) ____

Display This Question:
If Which Evaluation Association do you most identify with? Select one. = Lapsed Membership/Non-member

Q13 If lapsed membership or non-member, please give reasons for your non-renewal or non-membership status.

____
____
____
____
____

Q14 **What is your age?** *Select one***.**

○ 24 years or younger

○ 25-34 years

○ 35-44 years

○ 45-54 years

○ 55-64 years

○ 65 years or older

○ I prefer not to respond

Q15 **How do you self-identify?** *Select one*.

○ Male

○ Female

○ I self-identify as (please specify) ____

○ I prefer not to respond

Q16 What is your nationality?

○ American

○ Canadian

○ Other (Please specify) ____

Display This Question:
If What is your nationality? = American

Q17 If American, what is your racial/ethnic background? *Select one*.

○ African American, Black

○ American Indian, Native American, or Alaska Native

○ Asian

○ Caribbean Islander

○ European American, White

○ Latino or Hispanic

○ Native Hawaiian or Pacific Islander

○ Middle Eastern or Arab

○ Multiracial/Multicultural

○ I self-identify as (please specify) ____

○ I prefer not to respond

Display This Question:
If What is your nationality? = Canadian

Q18 If Canadian, what is your racial/ethnic group? Select one.

○ Aboriginal

- Arab
- Black
- Chinese
- Filipino
- First Nations
- Inuk
- Japanese
- Korean
- Japanese
- Latin American
- Metis
- South Asian
- Southeast Asian
- West Asian
- White
- I self-identify as (please specify) ____
- I prefer not to respond

End of Block: Your Demographic Information

Start of Block: Conclusion
Display This Question:
If HIDE = 1

Q19 **Are you interested in knowing how you compare with other evaluators on the professionalism measure?**

- Yes
- No

Display This Question:
If HIDE = 1

Q20 If yes, please provide your name and email address.

- ☐ Name: ____
- ☐ Email Address: ____

Display This Question:
If HIDE = 1

Q35 If yes, please provide your name and email address.

- ☐ Name: ____
- ☐ Email Address: ____

Q21 If you would like to clarify any of your responses or give feedback on any of the questions, please share below.

____
____
____
____
____

Q22 Congratulations! You have reached the end of the survey. You may now review or click submit. Please note that you will not be able to edit the survey after submission. Thank you for completing it. I appreciate your help.

End of Block: Conclusion

Appendix D: Survey Respondents' Demographic Characteristics

Table D1

Survey Respondents Demographic Characteristics

Demographic Variable	**n**	**%**
Gender		
Male	103	26%
Female	282	71%
Other	6	2%
Decline to state	4	1%
Nationality		
American	248	64%
Canadian	118	30%
Other (Please specify)	21	5%
American racial/ethnic background		
African American, Black	15	6%
American Indian, Native American, or Alaska Native	3	1%
Asian	8	3%
Caribbean Islander	2	1%
European American, White	187	77%
Latino or Hispanic	11	5%
Native Hawaiian or Pacific Islander	1	0%
Middle Eastern or Arab	0	0%
Multiracial/Multicultural	5	2%
I self-identify as (please specify)	5	2%
I prefer not to respond	7	3%

Table D1. (Continued)

Demographic Variable	n	%
Canadian racial/ethnic background		
Aboriginal	0	0%
Arab	0	0%
Black	5	4%
Chinese	1	1%
Filipino	0	0%
First Nations	0	0%
Inuk	0	0%
Japanese	0	0%
Korean	0	0%
Japanese	0	0%
Latin American	1	1%
Metis	1	1%
South Asian	2	2%
Southeast Asian	1	1%
West Asian	0	0%
White	91	78%
I self-identify as (please specify)	12	10%
I prefer not to respond	2	2%
Interested in knowing the professionalism measure		
Yes	128	46%
No	153	54%
Professional contexts		
PreK-12 education	106	10%
Higher education	108	10%
Business/industry	34	3%
Health care/public health	144	14%

Table D1. (Continued)

Demographic Variable	n	%
Foundations	86	8%
Government (federal/state/provincial/municipal/local)	196	19%
Nonprofit/social service	213	20%
Private sector/consultant	75	7%
International development	65	6%
Other (please specify)	25	2%
Place where most of the program evaluation take place		
Mostly within Canada	108	27%
Mostly outside Canada	13	3%
Equally within and outside Canada	3	1%
Mostly within the United States of America	220	56%
Mostly outside the United States of America	14	4%
Equally within and outside the United States of America	16	4%
Other (please specify)	22	6%
Skill level as an evaluator		
Not applicable (I do not consider myself to be an evaluator)	4	1%
Novice/entry-level evaluator	64	16%
Proficient/skilled evaluator	245	62%
Expert/master evaluator	78	20%
Other (please specify)	4	1%
gained professional knowledge as an evaluator		
University training/courses	307	17%
National/international professional association conferences/training/workshops	297	16%
Regional/local professional association conferences/training/workshops	257	14%
Job-embedded professional development	255	14%
Informal on-the-job learning	331	18%
Professional literature (e.g., books, journals, online resources)	344	19%

Table D1. (Continued)

Demographic Variable	n	%
Other (please specify)	40	2%
CES Credentialed Evaluator designation		
Yes	44	11%
No	298	75%
Highest degree earned		
High School Diploma	0	0%
Associate's degree	1	0%
Bachelor's degree	19	5%
Master's degree	217	55%
Doctoral degree	137	35%
Other (please specify)	21	5%
Current job title/position		
Director of an organization/institution	19	3%
Director of a program in an organization/institution	15	3%
Program provider/staff within an organization/institution	39	7%
Director of evaluation/assessment/research in an organization/institution	61	11%
Staff assisting in evaluation/assessment/research in an organization/institution	115	20%
Consultant providing services in evaluation/assessment/research	167	30%
University educator teaching evaluation/assessment/research	33	6%
Scholar/Theorist/Researcher in evaluation/assessment/research	30	5%
Graduate student in evaluation/assessment/research	34	6%
Other (please specify)	53	9%
Professional Association Membership		
Canadian Evaluation Society	98	25%
American Evaluation Association	244	62%
Canadian Evaluation Society and American Evaluation Association	25	6%

Table D1. (Continued)

Demographic Variable	n	%
Age		
24 years or younger	3	1%
25-34 years	87	22%
35-44 years	101	25%
45-54 years	85	21%
55-64 years	69	17%
65 years or older	46	12%
I prefer not to respond	7	2%