Title
Toward a Definition of Evaluative Thinking

Permalink
https://escholarship.org/uc/item/26t5x5f6

Author
Vo, Anne T.

Publication Date
2013

Peer reviewed|Thesis/dissertation
Toward a Definition of Evaluative Thinking

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Education

by

Anne Dao Thanh Vo

2013
ABSTRACT OF THE DISSERTATION

Toward a Definition of Evaluative Thinking

by

Anne Dao Thanh Vo

Doctor of Philosophy in Education
University of California, Los Angeles, 2013

Professor Marvin C. Alkin, Chair

The field of evaluation is at a critical juncture as it faces new scrutiny and questions about what constitutes good research and good practice. I argue in this study that if the discipline is to be rooted in a sound empirical foundation, we need a clear understanding of key terms employed by scholars and practitioners alike. In particular, greater clarity concerning the term “evaluative thinking” will allow evaluators to engage in deeper, more meaningful dialogue about their work, thereby advancing and strengthening the field.

This study empirically articulates an operational definition of evaluative thinking by systematically soliciting and analyzing opinion data from 28 evaluation experts using
the Delphi technique, an iterative survey method developed by the RAND Corporation. Results across three rounds of survey administration indicate that evaluative thinking is primarily linked to one’s use of data and evidence in argumentation and secondarily focused on reasoning and practice in the face of contextual constraints. Thinking evaluatively also requires striking a balance between objectivity, professional judgment, and personal conviction.

With these findings in mind, the study leads to a working definition for evaluative thinking that recognizes it as a particular kind of critical thinking and problem-solving approach that is germane to the evaluation field. Specifically, it is the process by which one marshals evaluative data and evidence to construct arguments that allow one to arrive at contextualized value judgments in a transparent fashion.

In light of these findings, this investigation challenges the idea that evaluation is strictly about determining an evaluand’s merit and worth. Rather, it is more productive to recognize that evaluators create knowledge during the evaluative process through the ways in which they address context. As such, the evaluative act and the thinking that accompanies it can—and should—be extended to include considerations for other dimensions that provide a more nuanced understanding of the evaluand and enable one to make evaluative claims about it. Understood in this way, the notion of evaluative thinking anchors the field’s sense of professional identity in the goal of solving social problems and in fulfilling an educative function.
The dissertation of Anne Dao Thanh Vo is approved.

Christina A. Christie
Noreen M. Webb
Todd M. Franke

Marvin C. Alkin, Committee Chair

University of California, Los Angeles
2013
THIS PAGE INTENTIONALLY LEFT BLANK.
Kính tặng Ba Má – thầy dẫu đời của tôi –

với tâm lòng yêu thương và quý trọng nhất.

For mom and dad – my first teachers in life –

with all my love and respect.
# TABLE OF CONTENTS

List of Tables........................................................................................................... x

List of Figures............................................................................................................ xi

Acknowledgements.................................................................................................. xii

Vita........................................................................................................................... xxiv

## CHAPTER 1: Introduction....................................................................................... 1
  - Statement of the Problem.................................................................................. 1
  - Conceptual Framework..................................................................................... 8
  - Study Purpose & Research Questions............................................................. 13
  - Study Significance & Implications.................................................................. 14
  - Manuscript Organization................................................................................ 16

## CHAPTER 2: Review of Relevant Literature....................................................... 17
  - Introduction....................................................................................................... 17
  - Reasoning Within Other Practice-Based Fields.............................................. 18
  - Reasoning Across Other Practice-Based Fields.............................................. 28
  - Logic & Reasoning in Evaluation.................................................................. 33
  - Research on Reasoning in Evaluation............................................................ 45
  - The Current Study.......................................................................................... 46

## CHAPTER 3: Research Methods............................................................................ 49
  - Sampling Frame............................................................................................... 49
  - Participants....................................................................................................... 50
  - Study Design.................................................................................................... 53
  - Procedures & Instruments............................................................................... 58
  - Research Tools.................................................................................................. 61
  - Definition of Terms.......................................................................................... 61
  - Analyses........................................................................................................... 64

## CHAPTER 4: Results............................................................................................ 73
  - Round 1 Results............................................................................................ 73
  - Round 2 Results............................................................................................ 78
  - Round 3 Results............................................................................................ 81
  - Cumulative Results......................................................................................... 84
  - Summary of Results....................................................................................... 101
LIST OF TABLES

Table 3.1  Study Participants by Theoretical Orientation................................. 52
Table 3.2  Summary of Procedures for a Delphi Commencing with a Closed-Ended Survey.............................................................. 57
Table 3.3  Inventory of 28 Descriptive Statements by Domain.......................... 70
Table 4.1  Summary Statistics for 20 Descriptive Statements Rated in Round 1. 74
Table 4.2  Reduced List of 33 Suggested Statements Respondents Provided During Round 1............................................................ 77
Table 4.3  Summary Statistics for 20 Descriptive Statements Rated in Round 2. 79
Table 4.4  Summary Statistics for 14 Descriptive Statements Rated in Round 3. 82
Table 4.5  Descriptive Statistics for 28 Statements Rated During the Delphi with Domain Labels, In Order of Importance to Evaluative Thinking.......................................................................................... 87
Table 4.6  Summary Statistics for Seven Statements Classified as “Very Important” for Evaluative Thinking, by Domain................................. 91
Table 4.7  Summary Statistics for 13 Statements Classified as “Important” for Evaluative Thinking, by Domain................................................ 93
Table 4.8  Summary Statistics for Seven Statements Classified as “Moderately Important” for Evaluative Thinking, by Domain................................. 94
Table 4.9  Classification of Eight Descriptive Statements for Which Consensus was Reached in Round 1, by Domain............................................. 97
Table 4.10 Classification of Six Descriptive Statements for Which Consensus was Reached in Round 2, by Domain................................................ 98
Table 4.11 Classification of Four Descriptive Statements for Which Consensus was Reached in Round 3, by Domain............................................. 99
Table 4.12 Classification of 10 Descriptive Statements Rated for Which Dissensus Remained in Round 3, by Domain................................. 100
LIST OF FIGURES

Figure 1.1 Conceptual Model of the Development of Evaluative Thinking as an Outcome......................................................... 11

Figure 1.2 Conceptual Model of Evaluative Thinking as a Necessary Input for Social Change.................................................. 12

Figure 1.3 Conceptual Model of Evaluative Thinking as an Individual Outcome and Input for Social Change.......................... 13

Figure 3.1 Possible Categories of Statements with Respect to Averaged Means and Variability.................................................... 62

Figure 3.2 Determining Importance Level from Four Possible Categories of Items in Terms of Average Rating and Variability of Rating........ 68

Figure 4.1 Scattergram of Mean Ratings and Variances of 20 Statements Rated with Respect to Averaged Mean and Variance (Round 1)........ 75

Figure 4.2 Scattergram of Mean Ratings and Variances of 20 Statements Rated with Respect to Averaged Mean and Variance (Round 2)........ 81

Figure 4.3 Scattergram of Mean Ratings and Variances of 14 Statements Rated with Respect to Averaged Mean and Variance (Round 3)........ 84

Figure 4.4 Scattergram of 28 Statements' Mean Ratings and 95% Confidence Intervals, by Domain..................................................... 86

Figure 4.5 Scattergram of 28 Statements' Mean Ratings and 95% Confidence Intervals, by Domain and Grouping................................. 90

Figure 4.6 Scattergram of 28 Statements' Mean Ratings and 95% Confidence Intervals, by Domain and Round in Which Consensus was Reached................................................. 96
ACKNOWLEDGEMENTS

As I reflect on the experiences that lead up to this moment, I am reminded of the many blessings that I have received and how fortunate I have been throughout this journey. Words of thanks cannot fully capture the profound gratitude that I feel towards those who have affected the trajectory of this personal and professional endeavor. Nonetheless, I rely on them here in an attempt to express my appreciation.

First, I thank my advisor, Professor Marvin Alkin. Marv has played a number of critical roles throughout my tenure at UCLA and has been an important part of my professional life these past six years. However, I will remember him most fondly as my teacher, trusted colleague and friend. Marv’s unrelenting support and unwavering confidence in my abilities has led to amazing opportunities that would have otherwise been difficult to come by. I am also grateful for all of the productive and thought-provoking discussions that we have had. Though mostly impromptu, it was during these conversations that I truly learned about evaluation and grew as an intellectual. Perhaps most importantly, I deeply appreciate Marv challenging me to be a better version of my scholarly self. Thank you, Marv, for all of these things and more.

The process of producing this document would have been much longer and unnecessarily arduous if not for the rest of my dissertation committee’s invaluable insights, commitment, and guidance. Although Tina Christie did not serve as my formal advisor, the energy and devotion that she has dedicated to my professional development
in the time that I have known her is awe-inspiring. I am deeply grateful for the guidance that Tina has provided, for the friendship that we have developed, and much more. Reenie Webb and Todd Franke’s encouragement to think carefully about my methodological decisions throughout this project, and others, as well as their gift for posing thoughtful, challenging questions not only pushed on, but also enhanced this investigation’s quality and reach. Together, this committee exemplified what it means to be good teachers and to be invested in their students. Each committee member has served as a role model for achieving a meaningful and successful career.

I extend my thanks to the 28 esteemed scholars who participated in my study. Each person’s involvement in this investigation contributed to it in unique and important ways. Eleanor Chelimsky, Lois-ellin Datta, Deborah Fournier, Jennifer Greene, Linda Mabry, Robin Miller, Jonny Morell, Sharon Rallis, and Tom Schwandt engaged with me in ways that far exceeded what I could ever imagined and hoped for. Our exchanges were deep and meaningful and will continue to shape my thinking for years to come. Jody Fitzpatrick, Rodney Hopson, Jean King, Mel Mark, Michael Patton, Laurie Stevahn, Will Shadish, and the late Carol Weiss provided encouragement at critical junctures. Bob Boruch, Brad Cousins, Stewart Donaldson, Gary Henry, Stafford Hood, Ernie House, George Julnes, Donna Mertens, Hallie Preskill, Debra Rog, and Joe Wholey were especially generous with their time despite incredibly hectic schedules. Everyone provided thoughtful comments throughout the investigation that shaped this document’s development.
Others within the UCLA School of Education in general and the Social Research Methodology (SRM) Division, in particular, have also played roles during this process. SRM faculty—Li Cai, Fred Erickson, Felipe Martinez, Mike Rose, and Mike Seltzer—pushed on my research abilities and expanded my repertoire of research methods in critical ways. In particular, Mike Rose’s informal mentorship and guidance with respect to writing and mixed methods research have simply been inspiring.

Peers and colleagues within the SRM Division, notably, Cecilia Henriquez Fernandez, Regina Richter Lagha, and Ji Seung Yang were kind, supportive, and constructive at all of the right times. Tarek Azzam, Eric Barela, and Karen Jarsky, alumni, are also recognized here for their mentorship and friendship throughout this journey. Eric deserves special thanks for turning me on to evaluation seven years ago when we worked together at the Los Angeles Unified School District.

Colleagues from the Academic Preparation and Educational Partnership (APEP) are also recognized here for encouraging me to think creatively, critically, and reflectively about my practice. They have each welcomed me into their uniquely productive spaces and granted me the privilege of working with and learning from them. Our collaborative exchanges have undoubtedly contributed to the development of my professional identity as an evaluator. Most importantly, I thank them for the sense of community that they have provided over the years—Aimée Dorr and Janina Montero (APEP Co-Chairs through 2012); Alfred Herrera (Center for Community College Partnerships); Merle Price and Jody Priselac (Center X); Justyn Patterson and Debbe
Pounds (Early Academic Outreach Program); Tony Tolbert and Leo Trujillo-Cox (Law Fellows Program); Natasha Saelua (Student Initiated Access Center and Community Programs Office); Christine Shen (Together-in-Education in Neighborhood Schools); and Jack Sutton (APEP Executive Coordinator).

I have also been fortunate to be in the company of other notable teachers and friends whom I met while supported by the Foreign Language and Area Studies (FLAS) Fellowship. Professors Quyen Di Chuc Bui, George Dutton, Thu-Ba Nguyen, Thu-Huong Nguyen-Vo helped me to develop a deeper understanding of and appreciation for the beauty and richness that characterizes the roots of my cultural heritage. Peers Lilly Nguyen, Kaitlyn Tram Nguyen, and Tam Lai are also thanked here for their genuine friendship, support, and for graciously and patiently putting up with me as I often struggled to find the right words to express myself in Vietnamese (or, more accurately, my broken version of it).

From the time that I met them, Ronald Scott, Vilaloy Phomsouvanh Warlick and Scott Warlick, and the Toapanta family welcomed me into their lives. I thank them for their generosity and treating me as one of their own. Victor Toapanta, Jr., in particular, is credited for his endless encouragement and uncanny ability to make me laugh. The time that I have had the pleasure of spending with my Angelino family has enabled me to maintain a more balanced life and contributed immensely to my general well-being.

Although I am attempting the task of thanking my parents—Nelson and Anita Vo—I continue to struggle because finding the words that accurately convey both the
depth and breadth of the gratitude, respect, and admiration that I feel towards them is the greatest challenge that I have encountered in this entire journey.

My parents loved me even before I was born. They endured hardships, made sacrifices, and toiled more than half their lives so that I can, among other things, attain an education, live in a free and democratic country, and become a contributing member of society. Though they always stressed the value of education, my parents placed greater emphasis on being a person of faith, integrity, and humility. These lessons and values are difficult to impart on others, but they have provided me with an unrivaled foundation upon which I can continue to develop these attributes and to become a better person.

I can neither thank my parents enough for their unconditional love nor can I express my deepest gratitude for their uncompromising support. As such, I only hope that my pursuits and accomplishments will be some of the sources of their comfort and joy and that my endeavors are worthy of the sacrifices that they have made over the years. From where I stand, it is an incredible honor to call them Mom and Dad and nothing makes me more proud than being their daughter.

Lastly, I acknowledge the funding support received from UCLA’s Graduate Division, the Graduate School of Education & Information Studies (GSE&IS), and the Social Research Methodology Division within the GSE&IS while enrolled as a doctoral student at the university. The views expressed here are mine alone and do not reflect the views or policies of the funding agencies or its grantees.
THIS PAGE INTENTIONALLY LEFT BLANK.
LỜI CẢM ON

Khi nghĩ về những kinh-nghiệm dẫn đến thời-diểm này, tôi cảm thấy mình rất may mắn vì đã nhận được nhiều phương tiện trong suốt cuộc hành trình giáo dục của mình. Lời cảm ơn không thể nói lên hết lòng biết ơn của tôi đối với những người đã ảnh hưởng đến phương hướng của sự nỗ lực cá nhân và chuyên nghiệp này. Mặc dù vậy, tôi vẫn dựa vào nó để bày tỏ sự cảm kích của mình.


Nếu hiểu sự hiểu biết sâu sắc và hướng dẫn tận tâm và vô giá của những thành viên ủy ban luận án thì quá trình soạn bài chức cần sẽ dài hơn và khó khăn hơn rất nhiều. Mặc dù Giáo Sư Tina Christie không phải có văn chính thức của tôi, sự tích cực và

Nói chung thì mỗi một người trong ủy ban này đã minh họa ý nghĩa và tiêu chuẩn của một vị giáo viên tốt qua cách đào tạo và đầu tư vào sinh viên của mình. Mỗi thảnh viên trong ủy ban là gương cho những sinh viên muốn đạt được thành công trong ngành nghiên cứu và quan trọng hơn nữa là sự nghiệp có ý nghĩa và chiều sâu.

House, George Julnes, Donna Mertens, Hallie Preskill, Debra Rog và Joe Wholey cùng nhiệt tình tham gia vào dự án này, mặc dù lịch trình của họ vẫn bận rộn. Tất cả mọi người đều có phần đóng góp ý kiến chính churn trong cuộc điều tra này.

Tôi cũng chân thành cảm ơn một số giảng viên, bạn học và đồng nghiệp trong Trường Đại Học Sử Phạm tại UCLA nói chung và Bộ Phướ Pháp Nghỉn Cựu xã Hội nói riêng. Giáo Sư Li Cai, Fred Erickson, Felipe Martinez, Mike Rose và Mike Seltzer đã thách thức khả năng nghiên cứu và mở rộng vốn kiến thức trong phân cụ của tôi một cách quan trọng. Đặc biệt nhất là sự đắn đo và hết lòng hướng dẫn của Giáo Sư Mike Rose về những vấn đề liên quan đến kỹ thuật sáng tác và phương pháp nghiên cứu hồn học.


Tôi cũng cảm ơn các đồng nghiệp từ Hội Đự Bị Học Tạp và Đội Tác Giáo Dục. Mỗi một người đã từng khuyễn khích tôi hãy nhìn vào những việc làm của mình với tâm
mắt sáng tạo và phê bình, mặc dù chẳng ai hay mình đã làm vậy. Những cơ hội được công sự, trao đổi cùng như học hỏi từ các dòng nghiệp trong Hội thật đáng quý. Sự hợp tác của chúng tôi đã góp phần vào sự phát triển đặc tín thẩm định chuyên nghiệp của tôi. Điều tôi cần cảm ơn nhất là bao không khí và những ý thức công đồng mà các dòng nghiệp đã tạo cho tôi trong những năm qua. Tôi chân thành gửi lời tri ân đến: Bà Aimée Dorr và Janina Montero (Đồng Chủ Tịch đến năm 2012); Ông Alfred Herrera (Trung Tâm Đối Tác Trường Cao Đẳng); Ông Merle Price và Bà Jody Priselac (Trung Tâm Sự Phạm/Trung Tâm X); Ông Justyn Patterson và Bà Debbe Pounds (Chương Trình Tiếp Cận Học Thuật); Ông Tony Tolbert và Leo Trujillo-Cox (Chương Trình Nghiên Cứu Pháp Luật); Cô Natasha Saelua (Trung Tâm Truy Cập do Sinh Viên Khởi Xướng và Văn Phòng Chương Trình Cộng Đồng); Cô Christine Shen (Chương Trình Cùng Nhau Trong Giáo Dục) và Ông Jack Sutton (Điều Phối Viên Điều Hành của Hội Dư Bị Học Tập và Đối Tiếp Giáo Dục).

Tôi cũng rất may mắn đã được sự đào tạo của một số giáo viên trong Bộ Văn Hóa Á Châu tại Trường Đại Học UCLA, nhất là trong thời gian tôi được học bổng nghiên cứu về văn hoá. Giáo Sư Bùi Quyên Di, George Dutton, Nguyễn Thu-Ba và Nguyễn-Vơ Thu-Hương đã giúp tôi mở rộng tầm hiểu biết về nguồn gốc của mình. Các thầy cô cũng đã giúp tôi cảm nhận được về cao đẹp và sự phong phú đặc trưng của lịch sử và văn hoá Việt Nam. Tôi cũng rất vui khi nhớ đến những chuỗi ngày được đến lớp, làm việc và gần
gửi chị Nguyễn Lilly, em Nguyễn Trâm Kaitlyn và em Lai Tâm. Điều tôi cần cảm ơn nhất là sự thông cảm và kiên nhẫn của các bạn đối với tôi khi đang cố gắng tìm những từ ngữ vững chắc hợp để thể hiện bản thân bằng tiếng Việt. Cho dù chân dung chúng tôi đi chung không dài, tôi vẫn cảm thấy mình rất may mắn khi được sự hỗ trợ, tình thương từ chúng như tình bạn chân thành của mọi người.


Mặc dù tôi đang cố gắng thực hiện, việc nói lên lời cảm ơn đến Ba Má—Ông Bà Nelson và Anita Huỳnh Vô—thật rất khó đối với tôi. Việc tìm kiếm những từ ngữ có thể truyền đạt chính xác cả chiều sâu lẫn chiều rộng của lòng biết ơn, tôn trọng và ngưỡng mộ tôi dành cho Ba Má là sự thách thức lớn nhất tôi gặp phải trong toàn bộ cuộc hành trình này.

Ba Má đã thường tôi ngay cả trước khi tôi ra đời. Hai Ông Bà chấp nhận mọi khó khăn, chịu đựng mọi hy sinh và làm việc nhọc nhằn vật và hồn nã cuộc đời. Mục đích là
nuôi con lớn lên, để rồi có thể lớn lên trên đất nước tự do dân chủ, trở thành người hữu dụng và đồng góp cho xã hội. Mặc dù lúc nào Ba Má cũng nhận mảnh giá trị của nền giáo dục, hai Ông Bà chú trong đến vấn đề đức tin và cách làm người nhiều hơn—nất là một người có tinh chính trực và biết kiếm tốn. Truyền đạt những bài học ở đời này rất khó, nhưng Ba Má đã tạo cho tôi một nền tảng vững chắc, ít ai có thể sánh được. Nhờ đó mà tôi có thể tiếp tục phát triển những đức tính này và trở thành một người tốt.

Không có một lời cảm ơn nào có thể nói lên hết lòng biết ơn sâu sắc đối với tình thương vô điều kiện và sự ủng hộ kiên quyết mà Ba Má đã dành cho tôi. Vì vậy tôi chỉ hy vọng rằng những thành tích của tôi luôn sẽ là nguồn an ủi và niềm vui của Ba Má trong đời sống còn lại. Tôi cũng mong nổ lực của mình sẽ xứng đáng với những sự hy sinh của Ba Má trong những năm qua. Theo tôi thì không có gì bằng được làm con của Ba Má.

Cuối cùng, tôi cảm ơn sự hỗ trợ tài chánh đã nhận được từ Văn Phòng Cao Học tại Trường Đại Học UCLA, Trường Đại Học Sư Phạm và Nghiên Củ Thông Tin và Bộ Phương Pháp Nghiên Củ Xã Hội khi còn là sinh viên tiến sĩ tại trường. Toàn bộ quan điểm được thể hiện ở đây là của một mình tôi và không phản ánh quan điểm hay chính sách của các nhà tài trợ.
VITA

EDUCATION

2005  B.A., Psychology and English
      University of California, Los Angeles

2008  M.A., Education
      University of California, Los Angeles

PROFESSIONAL EXPERIENCE

2005-2006  Research Assistant
            Program Evaluation and Research Branch
            Los Angeles Unified School District

2006-2007  Consultant
            Student Assignment Center
            Oakland Unified School District

2008-2013  Teaching Assistant/Graduate Student Researcher
            Department of Education
            University of California, Los Angeles

2012-2013  Lecturer
            Charter College of Education
            California State University, Los Angeles

SELECT PUBLICATIONS


SELECT PRESENTATIONS

2009  

2010  

2011  


SELECT HONORS AND AWARDS

2012-2013  
Dissertation Year Fellowship  
Graduate Division  
University of California, Los Angeles

2010-2011  
Graduate Research Mentorship Fellowship  
Graduate Division  
University of California, Los Angeles

2009-2010  
Foreign Language and Area Studies Fellowship – Vietnamese  
Graduate Division  
University of California, Los Angeles
CHAPTER 1
INTRODUCTION

Before we start talking, let us decide what we are talking about.

— Socrates, in Plato’s Phaedrus

The field of evaluation is at a critical juncture, as it faces new scrutiny and questions about what constitutes good research and good practice. This introductory chapter describes this context and argues that if the discipline is to be rooted in a sound empirical foundation, we need a clear understanding of key terms employed by scholars and practitioners alike. In particular, greater clarity concerning the term “evaluative thinking” will allow evaluators to engage in deeper, more meaningful dialogue about their work, thereby advancing and strengthening the field.

Statement of the Problem

Evaluation is an intrinsic aspect of human life. Generally speaking, people engage in some form of evaluation on a daily basis. Deciding which type of car to purchase, which political candidate to vote for, or which charitable organization to donate to all require an evaluative journey of one sort or another. These types of evaluations, however, are quite different from professional evaluation.

Professional evaluation, of which there are many kinds (e.g., personnel, policy, product, program, student, etc.), involve value judgments made in a systematic fashion. Some writers have traced the history of evaluation to Biblical times, citing comparisons
of the effects of a Hebrew versus Babylonian diet on health as one early example
(Shadish & Luellen, 2005), while others have pointed to Chinese civil service exams
dating back to 2000 B.C. as the origin (Fitzpatrick, Sanders, & Worthen, 2004). Many
evaluators have credited Ralph Tyler’s much more recent “Eight Year Study” as the
point at which modern professional evaluation was born. This curriculum and
instruction evaluation study, conducted from 1933 to 1941 with 30 elementary and
secondary schools as well as 300 colleges and universities, indicates that evaluation as
we now know it has its modern roots in the education field (Tyler, 1942).

Since Tyler’s hallmark contribution approximately 80 years ago, the evaluation
field has grown in a number of areas. For example, evaluation methods, tools, and
analytic techniques have broadened from primarily quantitative approaches during the
War on Poverty and Great Society legislations in the 1960s and 1970s to now include
qualitative and mixed methods approaches (Rossi, Lipsey, & Freeman, 2004; Shadish,
Cook, & Campbell, 2001). Also, the number of professional evaluation associations—and
the size of their membership—has increased from about five in 1990 to over 90
worldwide (Donaldson & Christie, 2006; Mertens, 2003).

Developments such as these have given way to the proliferation of ideas,
concepts, language, and models that now make up the rich intellectual fabric of
evaluation. The field’s general growth could be attributed to a number of factors,
including funders’ increased interest in return on investment and the importance placed
on accountability. Demand for evaluation in the United States, in particular, is likely
driven by the pressure on new and established social programs to demonstrate effectiveness in the face of dwindling federal resources to finance them. In this context, the high stakes nature of funding decisions, coupled with justified concerns about equity and social justice in programmatic and policy decision-making, have led the policy, research, and evaluation communities to pause and reassess.

A series of federal initiatives at the start of the 21st century, for example, forced the evaluation discipline to reflect critically on long-standing issues related to rigor, relevance, and quality of research and practice. The Elementary and Secondary Education Act, reauthorized by the No Child Left Behind Act of 2001, shaped and continues to affect how educational evaluators address issues of accountability and credible evidence. A year later, the establishment of the What Works Clearinghouse reignited debates about what counts as empirical research (Institute of Education Sciences, 2002). In 2003, the U.S. Department of Education released the Scientifically Based Evaluation Methods (SBEM) statement, which clearly defined rigorous research as studies that determine causality through experimental, quasi-experimental, regression discontinuity, and single-subject designs. And most recently, in the 2011–2012 and 2012–2013 fiscal years, the Office of Management and Budget (OMB) allocated financial resources for research on evaluation capacity (Orszag, 2009a, 2009b, 2010).

While each of these is interesting as a standalone event, together they have far-reaching implications for the continued development and trajectory of the evaluation
field. In particular, they call for us to revisit and re-examine questions such as: What counts as credible evidence? How should standards of rigor be determined? How do we know that the evaluations we are conducting are rigorous? How do we know that our evaluations are of high quality?

The evaluation community has responded to these initiatives—directly and otherwise—in a variety of ways, including: the American Evaluation Association’s (AEA) answer to the Department of Education’s SBEM statement, released in 2004; revisions to AEA’s Guiding Principles for Evaluators, also published in 2004; the establishment of the Evaluation Policy Task Force within AEA in 2007; an edited text by Donaldson, Christie, and Mark (2009) that includes diverse responses to the question, “What counts as credible evidence?”; the establishment of AEA’s Evaluation Policy Topical Interest Group in 2010; and revisions to the widely accepted Program Evaluation Standards (Yarbrough, Shulha, Hopson, & Caruthers, 2010).

Engagement with policymakers, scholarly exchanges, clearly-articulated guiding principles for practice, and standards for judging evaluation quality are important not only to individual evaluators, but also more broadly to evaluation as a practice and a discipline. But the field also needs an active and thriving empirical knowledge base, in part because it will allow evaluators to continue responding to the types of external forces previously described. However, obstacles to building this foundation persist; evaluators have been challenged in their efforts to support productive discussions with those unfamiliar with evaluation, to identify and develop evidence-based practices, and
to continue to progress in the establishment of evaluation as both an academic and practice-based discipline. The paucity of existing research on evaluation is due, at least in part, to an ongoing lack of resources to support such an enterprise, as well as to the slow-growing number of academic programs in evaluation (LaVelle & Donaldson, 2010). These factors have hindered growth in the design and implementation of new studies, leaving unfilled gaps in the literature in spite of a plethora of calls to address the issue (Christie, 2011; Henry & Mark, 2003; Mark, 2008; Mark & Henry, 2004; Miller, 2010).

One of the more notable gaps relates to the basic concepts employed by the evaluation field. Nearly 20 years ago, in a reflective article about positivist philosophy, Shadish (1995) argued that contemporary perceptions about and ideas associated with positivism were often being misused and misunderstood by evaluation scholars and practitioners. Moreover, those who considered themselves positivists often did not fully understand what positivism was, leading to ill-informed debates about research methods that subsequently drove activities within the field and policymaking in consequential ways. It is still essential that evaluation researchers and practitioners reach consensus on the terminology used within the field, or at the very least understand the sources of disagreement. Thus, scholars have begun to unpack key concepts. Miller and Campbell (2006) and Fierro (2012), for example, have taken the lead with “empowerment evaluation” and “evaluation capacity building,” respectively. The refinement of these ideas is key to the future success not only of research on evaluation, but of the evaluation field more broadly. Specifically, we must clarify
constructs that have made their way into the evaluation lexicon and witnessed popularized use before we delve into more studies designed to offer causal explanations and identify causal mechanisms.

In this spirit, let us consider more carefully the term “evaluative thinking,” which has found its way into the evaluation literature both explicitly and implicitly, yet with little consensus about its meaning. The challenge in the field with respect to this lack of consensus relates to the under-developed conceptual, theoretical, and empirical knowledge bases pertaining to the construct and to a general lack of clarity concerning what “evaluative thinking” is. Varied uses by those within and outside of the evaluation community further exacerbate the issue.

If, for example, evaluative thinking refers to how one develops the reasoning skills used prior to or during the conduct of professional evaluation, then the bodies of literature on prescriptive evaluation theory (i.e., guiding frameworks that specify what a good or proper evaluation is and how it should be done), the teaching of evaluation (i.e., methods and approaches that are used to train and educate evaluators and program administrators), and research on evaluation (i.e., the systematic examination of the field’s theories, methods, and practices) are most relevant for understanding the construct. But, if evaluative thinking is a referent for substantive contemplation about the potential effects of evaluation on society, then the field’s vast theoretical and conceptual literature bases (i.e., evaluation’s role in upholding democracy and democratic ideals) are particularly relevant. A third possibility (of many)—one that is
explored to a lesser extent in the evaluation literature—is that evaluative thinking might signal the ways in which individuals make decisions and determine value in daily life. Which is it? The answer to this question has significant implications for the field’s common understanding of a principle that is fundamental to both theory and practice.

It is reasonable to anticipate multifarious understandings of emerging ideas that stem from a field as nascent as evaluation. In this context, it is also reasonable to expect scholars and practitioners to adapt tools and appropriate language from other disciplines in order to fit evaluative purposes. Moreover, variations in evaluators’ interpretive lenses are likely responsible, at least in part, for the different ways in which the same ideas, concepts, and tools are used within the field. However, we need some agreement about the core elements or defining features of often used but seldom defined ideas like evaluative thinking. This agreement will allow us to make informed decisions about how it can be studied, measured, and taught.

If we do not take up this task of unpacking the notion of evaluative thinking, we risk perpetuating the production of unexamined ideas, concepts, and models that do not contribute any new knowledge to theory or practice. This could result in increased misunderstandings that impede progress towards the development of a broad and deep empirical knowledge base and a strengthened field. In other words, it is critical for individuals to agree on what it is they are arguing about before taking strong positions on either side of the aisle. The ability to share ideas and debate issues in an intelligible manner is essential to the field-building endeavor because this is how productive
conversations—scholarly and otherwise—take place and stagnancy is combatted.

The present study addresses these issues by contributing a systematic, descriptive investigation of consensus and dissensus among evaluation experts regarding ideas that are at the core of evaluative reasoning. This is accomplished through the Delphi Technique—an iterative survey method developed by the RAND Corporation—and a content analysis of participants’ feedback. Defining and refining the construct adds clarity to what scholars and practitioners mean when they say “evaluative thinking” and, in so doing, not only contributes to the foundation from which future research on evaluation studies can be conducted, but also to the growth of the field.

**Conceptual Framework**

Three potential ways in which evaluative thinking can be understood were outlined in the previous section: 1) how one develops the reasoning skills used during the conduct of professional evaluation; 2) substantive contemplation about the potential effects of evaluation on society; and 3) the way in which individuals make decisions and determine value in daily life. The last of these three referents is not emphasized in the remainder of this manuscript because it does not deal directly with professional evaluation. The two referents that remain, however, are important elements of this study’s conceptual framework and will be further discussed in this section of the chapter.

The body of literature on evaluative thinking may be considered underdeveloped, on the one hand, and diffuse on the other. As I will discuss in greater detail in the next
chapter, the idea is usually not addressed explicitly in the relatively few publications in which it appears explicitly (Patton, 2002a; Schwandt, 2002, 2008a; Scriven, 1995). And, when discussed implicitly, it is found in a broad range of areas in the evaluation literature not limited to *prescriptive evaluation theory* (Alkin & Christie, 2004; Alkin & Christie, 2005; Christie & Alkin, 2013; Christie & Azzam, 2005; Cousins & Earl, 1992; Greene, 2005; Mark, Henry, & Julnes, 1998; Shadish, Cook, & Leviton, 1991); *teaching of evaluation* (Christie & Rose, 2003; Lee, LeBaron Wallace, & Alkin, 2007; Levin-Rozalis & Rosenstein, 2003; Patton, 2005), including sub-topics on *evaluator competencies* (Ghere, King, Stevahn, & Minnema, 2006; Stevahn, King, Ghere, & Minnema, 2005), *evaluation capacity building* (Arnold, 2006; Bourgeois, Chouinard, & Cousins, 2008; Preskill & Boyle, 2008), and *organizational learning* (Preskill & Torres, 1999; Rogers & Williams, 2006); *research on evaluation* (Alkin, 1991; Azzam, 2011; Azzam & Szanyi, 2011; Fitzpatrick, Christie & Mark, 2008); and evaluation’s *potential role in and impact on society* (Chelimsky & Shadish, 1997; Cronbach, 1980; Dahler-Larsen, 2012; Schwandt, 2002), which includes discussions about *evaluation logic* (Fournier, 1995a, 1995b, 2005; House, 1977, 1995; Rog, 1995; Scriven, 1995, 2005).

Given this lack of detailed and focused attention, trying to understand what scholars and practitioners mean when they use the term “evaluative thinking” poses a great challenge. To better understand how this construct has been addressed in the field, I developed a conceptual framework based on a review of the evaluation literature. Specifically, I explored publications that implicitly and/or explicitly addressed
evaluative thinking and organized the ideas according to whether the emphasis was on
1) how one develops the reasoning skills used during the conduct of professional
evaluation or 2) the potential ways in which thinking evaluatively contributes to an
improved society. These categorizations are represented in Figures 1.1 and 1.2,
respectively.

Figure 1.1 describes three distinct yet interrelated elements: inputs, activities, and
individual outcomes. Towards the left side of the figure are inputs and activities that
could lead to the development of evaluative thinking as an ability at the individual
level—e.g., a skill acquired as the result of completing certain activities, such as formal
education and training, mentorship, and practical experience (gained, perhaps, by
observing or participating in an evaluation). Bodies of evaluation literature that directly
map onto this figure have to do with the process of becoming an evaluator or building an
evaluation skill set, and include the teaching of evaluation, evaluation capacity building,
and evaluator competencies (e.g., Christie & Rose, 2003; Ghere, King, Stevahn, &
Minnema, 2006; Lee, LeBaron Wallace, & Alkin, 2007; Levin-Rozalis & Rosenstein,
2003; Preskill & Boyle, 2008).
Figure 1.1
*Conceptual Model of the Development of Evaluative Thinking as an Outcome*

![Diagram showing inputs, activities, and evaluative thinking]

Figure 1.2 addresses the anticipated effects of individuals on the environments in which they are nested, specifically as a result of those individuals having undergone the transformative process depicted in Figure 1.1. The expected effects may occur at the programmatic, organizational, or societal level, and can be observed through shifts in social norms and culture through changes in values, beliefs, and practices. Thus, Figure 1.2 depicts evaluative thinking not as an individual outcome, but rather as a necessary ingredient (or, input) for social change on a broader level. Bodies of evaluation literature that directly map onto this figure typically focus more broadly than individual development to include the effects of evaluation on a group or community. The literature includes prescriptive evaluation theory, organizational learning, and capacity building (e.g., Alkin, 1991; Alkin & Christie, 2004; Alkin & Christie, 2005; Christie & Alkin, 2013; Preskill & Torres, 1999; Rogers & Williams, 2006).
It is often implied in the literature that engaging in certain activities will lead to an individual’s ability to think evaluatively. On the other hand, it is also implied that these shifts—usually demonstrated as or evidenced by changes in group-level processes—should contribute positively to the settings in which an individual who has evaluation knowledge and skills functions. With this in mind, Figure 1.3 represents the intersection between these two understandings of evaluative thinking. Specifically, it depicts the connections between individual change (as shown in Figure 1.1) and broader impact (as shown in Figure 1.2) and speaks to the complex role of evaluative thinking as both an outcome on the individual level and an input on the group level. In other words, the center of Figure 1.3 highlights the way that evaluative thinking can potentially be transformed from an outcome (such as a skill) that lies within an individual evaluator to an input that leads to broader social change.

Figure 1.3 is especially useful because it visually depicts the areas in which evaluative thinking has and has not been addressed in the evaluation field, providing a broad strokes representation of the construct’s place within the literature. As noted,
evaluate thinking as either an outcome or an input has previously been taken up in a number of areas within evaluation, albeit mostly in a high-level manner (Patton, 2002a, 2002b; Schwandt, 2002, 2008a; Scriven, 1995). A more comprehensive interpretation of the construct—one that takes into account the concept’s role as an individual outcome, an input for social change, and a combination of the two—has yet to be studied in any systematic fashion. Thus, the framework reveals the gaps in the existing knowledge base and highlights this investigation’s unique contribution to the field by providing a sense of where the study’s focus and scope fit within the broader evaluation literature.

Figure 1.3
Conceptual Model of Evaluative Thinking as an Individual Outcome and Input for Social Change

Study Purpose & Research Questions

The purposes of this study were two-fold: (1) to identify ideas that experts consider central to the notion of evaluative thinking, and (2) to empirically derive a working definition of evaluative thinking that can be used to guide future research. To accomplish these goals, I was guided by the following research questions:
1. What do evaluation experts consider important to evaluative thinking?
2. What is the nature of consensus and dissensus among experts?

**Study Significance & Implications**

By answering these two research questions, this investigation makes a contribution to the evaluation community and those outside of it on several levels. First, from a methodological perspective, this study provides a roadmap for how to systematically study a vague construct. It also offers suggested indicators around which tools can be developed to measure the extent to which evaluative thinking is occurring. This insight is valuable to scholars and practitioners not only in the evaluation field, but in other fields and disciplines as well.

Second, this study contributes a working definition for thinking about evaluative thinking. This is of primary interest to evaluation researchers who can use it as a starting point to design studies intended to better understand how evaluative reasoning develops and plays out at the individual, programmatic, organizational, and societal levels. Researchers in other disciplines might adapt the framework for the purposes of studying related constructs in their respective areas of inquiry. These research endeavors can stimulate additional empirical work on the one hand and inform and improve practice in evaluation as well as other fields on the other hand.

Additionally, because the term has witnessed such broad use, it is often linked to myriad aspects of the organization, from learning and change to capacity and culture. Unfortunately, its disparate and varied implications pose a number of practical and
highly probable challenges, including the lack of a shared frame of reference that can be used to communicate about evaluative thinking. Thus, while grantor agencies, for instance, might release calls for proposals to examine the extent to which an organization thinks evaluatively or has the capacity to do so, responding evaluators can only make educated guesses about what grantors mean. As such, the resulting products and findings may have limited use and relevance. By suggesting what indicators should be considered in determining the extent to which an organization or individual is thinking evaluatively, it is anticipated that the findings of the study will lead to the generation of new knowledge about the intricacies of organizational life, particularly when evaluation is involved, and improved decision-making at the organizational level.

Further, the study’s findings have implications for evaluator training and education. Postsecondary institutions—both degree-granting and certificate-oriented—that offer training in evaluation might use the results to develop and refine their programs. University faculty, in particular, can use the information to focus their course content and improve the ways that instruction is delivered. Doing so will ensure that students complete their coursework ready and able to think evaluatively.

Perhaps most importantly, this study will serve as a catalyst for the evaluation community to consider how the discipline can continue to progress towards the goals of building a clear and sound body of empirical research and of translating theoretical and conceptual ideas into researchable constructs, actionable agendas for research and practice, and practical tools.
On the whole, this investigation fills a critical gap in the evaluation knowledge base and encourages critical dialogue about what evaluative thinking is within the evaluation research community, which can lead to more reflective and rigorous practice among evaluators.

**Manuscript Organization**

The remainder of this manuscript consists of four chapters that provide more context for the research, as well as summarize the study procedures and findings. In Chapter 2, I present a review of the relevant literature on cognition and reasoning in evaluation and other practice-based fields. Chapter 3 describes the study’s methods, including the procedures of the Delphi Technique and the descriptive and content analyses that were conducted. In Chapter 4, I outline the study findings in terms of the research questions presented above. Finally, in Chapter 5, I present conclusions drawn from the study’s results and describe the implications of the findings for the evaluation field. In this final chapter I also reflect on the study’s limitations and make suggestions for future research.
CHAPTER 2

REVIEW OF RELEVANT LITERATURE

Evaluative thinking, or evaluative reasoning, is an emerging construct in the evaluation field. Only a limited number of texts directly address it, and so the body of literature on this topic may be considered underdeveloped. Therefore, to effectively frame the current investigation, this literature review draws from a broad range of areas. Specifically, I will review what is known concerning comparable concepts in other practice-based fields, including education, law, and medicine. I will then turn my focus to research on thinking and reasoning specifically in the field of evaluation. Finally, I conclude with a description of the state of evaluative thinking in evaluation.

**Introduction**

In general science, cognition refers to a broad range of mental processes such as memory, reasoning, perception, learning, and language production. The study of these various processes and the mind as a whole has led to substantial understandings of its inner workings and, in particular, how people go about solving various complex problems. Ideas developed around inductive, deductive, and abductive thinking—examples of problem solving approaches that I will describe in detail later in the chapter—are considered contributions and advances that stem from research on cognition in general and thinking and reasoning more specifically. The present study focuses on one type of thinking and reasoning—evaluative thinking—in general terms.
Thus, the remainder of this literature review will hone in on the question of how thinking and reasoning have been studied and are understood in other fields. Although there are distinctions that could be made between “reasoning” and “thinking,” they are not germane to this research and I therefore use the terms interchangeably. The discussion that follows will serve as a backdrop for reflecting on and understanding the present state of and gaps in knowledge within the evaluation discipline.

**Reasoning Within Other Practice-Based Fields**

Every human choice—from low-stakes, everyday decisions to those requiring an ethical or moral stance—requires one to engage in reasoning and thought. It is a familiar process and activity that we drift into—often involuntarily—like breathing. As such, stepping back from what we mean when we refer to “thinking” and talking about it objectively can be quite a difficult task to accomplish. This challenge is perhaps amplified during discussions about evaluative reasoning because not very much is known about it and so we lack a frame of reference. For this reason, an examination of thinking and reasoning as it is described in other fields is useful. As we create distance from the object of interest—here, evaluative thinking—what is familiar becomes unfamiliar, and a more effective examination from different angles can be achieved.

In this section of the literature review, I describe how thinking and reasoning within the practice-based fields of education, law, and medicine have been described. Each of these disciplines has a long, complex history and has been influenced by a number of different schools of thought. As such, a comprehensive review of the
literature in each field is not feasible. However, a number of key insights drawn from work that specifically addresses reasoning and thinking in each field are valuable because they highlight areas of overlap and distinction between evaluative thinking and comparable concepts in other professions and areas of study. This contributes to a deepened understanding of the same construct’s developmental trajectory across different fields.

**Educational Thinking**

In the education field, thinking and reasoning conjure—among other things—notions related to intelligence, assessment, achievement, aptitude, setting standards, instructional time, and pacing plans. All of these topics are quite broad and can be unwieldy. Additionally, some of these topics are more politically charged than others, making a coherent discussion of them a bit challenging because the point of view taken greatly influences the scope and direction of the conversation. Thus, let us consider reasoning in this area from two different, but intimately related, perspectives—that of the student and that of the instructor.

Critical thinking often comes to mind in discussions about students’ reasoning and intellectual abilities. Discourse around this issue has been shaped by several generations of educators, educational philosophers, and scholars not limited to Aristotle, Bloom, Dewey, Piaget, and Vygotsky. Of the various theories and frameworks for learning that have been developed in this area (e.g., Cole & Engestrom, 1996; Piaget, 1971), Bloom’s (1956) *Taxonomy of Educational Objectives*, and its many subsequent
iterations, is perhaps one of the most notable for our present purposes. Its early influence on the trajectory of student assessment and how the education field came to think about measuring student learning is inarguable.

Bloom’s framework organized learning objectives into three domains—cognitive, affective, and psychomotor. The cognitive domain received the most attention and consisted of knowledge definition, comprehension, application, analysis, synthesis, and evaluation. Each category within the cognitive domain was broken down further and described in terms of the skills and abilities that demonstrate the level of complexity in thinking. Bloom’s work led the education field not only to carefully define constructs such as intelligence and achievement, but also to differentiate between different kinds of knowledge and different ways that students might demonstrate that they are thinking critically. Clearly, critical thinking involves demonstrating the ability to handle complexity and going beyond procedural knowledge, but perhaps more importantly, it involves solving problems across a range of content areas (e.g., literature, mathematics, science).

Reasoning from the instructor’s point of view, then, should involve contemplation about how to educate students in a way that will help them develop and attain higher order skills such as critical thinking. Within the teacher education community, this kind of reasoning is generally referred to as pedagogical reasoning, and it typically pertains to instructors’ content expertise (i.e., content knowledge) and their ability to consider issues and concerns that are germane to the teaching profession (i.e.,

Pedagogical reasoning is traditionally thought of in two ways. First, it can be observed through the instructor’s understanding of the way in which knowledge is organized in a given subject (i.e., structure of knowledge; Bruner, 1977) and his/her ability to communicate core principles, theories, and concepts that are instrumental to students’ uptake of the lesson being taught. Thus, the process by which instructors take in, organize, retrieve, and present information is especially pertinent when considering pedagogical reasoning from the perspective of content knowledge.

Alternatively, pedagogical reasoning can be observed in how teachers navigate the ever-shifting classroom context. Because classrooms are familiar to many for a variety of different reasons, it is understandably easy to take what might be considered “everyday classroom activity” at face value and miss the level of complexity and nuance that are masked by the buzz and hum in a typical classroom. For example, it is quite reasonable to surmise that differences in teacher practice could be explained, at least partially, by contextual factors including students’ grade levels, whether they are considered English Language Learners, the amount of experience the instructor has, and the subject matter that is being taught. Thus, the mental processes that are activated to negotiate these issues individually as they arise, as well as on the whole, along with the cognitive power called upon to decide how instructional content will be
delivered (Chen & Ennis, 1995) can also be considered pedagogical reasoning, albeit of another ilk.

**Legal Thinking**

Legal reasoning or thinking is used to describe thought processes specific to the field of law. As with reasoning in education, this type of reasoning will vary depending on the perspective of the individual who undertakes it. This section of the literature review will focus on conceptualizations of reasoning undertaken by those who have been legally trained—legal counsel and jurists—primarily in the context of criminal law. While it is reasonable to argue that jurors also engage in some form of legal thinking, it mostly involves efforts to remain objective so that prior knowledge and experiences do not influence their verdicts. Thus, jurors’ reasoning is different and is not under consideration here (Spellman & Schauer, 2012).

When one traces the roots of legal reasoning, two common schools of thought emerge. On the one hand, the “traditional view” contends that legal reasoning is different from other forms of thinking. Specifically, it can refer to ethical judgments that have to be made in light of the moral principles that have been laid out and set by the law (Spellman & Schauer, 2012). Legal decisions addressed by the traditional view typically involve sentence duration for felony (e.g., murder, aggravated assault) and misdemeanor (e.g., driving under the influence, simple battery) crimes. On the other hand, the “realist view” posits that legal reasoning is subject to confirmation bias and motivated reasoning. That is, it leads to results consistent with one’s personal beliefs.
and allows one to serve as “the master of [his/her] own ideas of what is good or right” (Scharffs, 2004, p. 733). Examples include rulings regarding the legality of school segregation, abortion, and same-sex marriage. These philosophical stances can be translated into two very practical questions about legal reasoning, namely, how is a case argued and how is judgment reached?

The practice of law involves a variety of different activities including representing clients, drawing up wills and contracts, and engaging in negotiations. However, legal reasoning is primarily understood in the context of the thinking that is demonstrated in courtrooms, particularly as lawyers prepare for and engage in trials and appeals (Coke, 1628). In this sense, it is quite different from pedagogical reasoning, which is more broadly defined and touches on a number of different areas within education that relate to a range of other individuals.

Attorneys’ reasoning has been described in a number of different ways in the law literature, most often through binary frameworks. One such framework describes the process by which a legal argument is constructed—deductively or inductively—and the form that it takes when presented (Lamond, 2008; Rips, 2001; Walker, 2009). Building an argument deductively through rule-based reasoning requires the lawyer to identify the legal issue, the statute that has been violated, the ways in which the statute has been violated, and the evidence relevant to the case. Deductive legal reasoning focuses on demonstrating how the law has been violated using the evidence that is available.
In contrast, arguments that are constructed inductively are done through reasoning by analogy. This involves examination of the legal issue, all relevant facts, the ways in which outcomes of previous cases might or might not be applicable to the present case, and the law(s) that have been broken. Inductive legal reasoning is interested in proving how the law has or has not been broken by examining the outcomes of previously tried cases. Thus, this approach capitalizes more heavily and explicitly on the rule of law, which requires that “similar cases be decided similarly” (Walker, 2007, p. 1687). The attorney forms a conclusion about the issue based on how these various elements fall with respect to each other and presents it for the judge’s consideration, regardless of the approach that was used when forming the argument. Further, while presented in a dichotomous fashion here, it is important to note that there is fluidity between the extremes of this continuum, as counsel is likely to call upon both approaches in practice.

While the prosecutor and defense are focused on building and presenting their cases, the judge is responsible for determining the extent to which the underlying logic of their arguments is sound. This is accomplished through cognitive processes similar to those described above, but is differentiated with an added element—namely, judicial discretion. In the legal arena, discretion is just as much about personal bias as it is about one’s interpretation of the law (Ferguson, Babcock, & Shane, 2008; Guthrie, Rachlinski, & Wistrich, 2007). Further, discretion creates a problem for the judge—one that requires the present judgment to reflect “how an old equity can be preserved and better
and how new justice be done” (Levi, 1965, p. 397; Schauer, 2008). In effect, the judge must legitimize his or her own ruling while also maintaining consistency with previous judgments. Certainly, the cognitive push and pull that one experiences during the process of arriving at conclusions of the legal variety is demanding on a number of levels and reflects a skill that is honed over time through socialization and professional preparation. These issues and concerns have recently begun to receive attention within the field of law. Hartung and George (2009) and O’Neill (2008), for example, have called for additional empirical examinations of how lawyers are trained and specifically how their legal reasoning develops while they are attending law school.

**Medical Thinking**

Similar to education and law, the history of modern medicine—and likewise, reasoning and thought within the field—can be traced to ancient times. The notion of medical reasoning, or clinical reasoning, is a relatively nascent idea, however, when compared to the medical field’s rich and long history. Broadly used, it refers to the ways in which medical problems are solved. Examples include how clinicians apply their medical training in real-life settings and the cognitive processes by which diagnoses are reached (Patel, Arocha, & Zhang, 2012).

While still a strong influence on medical reasoning, hypothetico-deductive reasoning, or the hypothesis testing tradition (Elstein, Shulman, Sprafka, 1978), has recently been challenged and alternative ways of describing thinking in medicine have been presented in the literature. Hypothesis-directed and data-directed reasoning have
often been juxtaposed, as have ideas related to the vertical and horizontal organization of medical knowledge. Each of these ways of considering medical reasoning is explored in turn.

As in so many other practices and disciplines that are rooted in science, the opinion that clinical reasoning is hypothesis-directed (i.e., driven by deductive processes) is broadly held. In medicine, this mode of thinking involves reasoning that generates and tests hypotheses so that potential causes of symptoms (clinical data) can be individually ruled out, a correct diagnosis can be reached, and a treatment plan can be decided upon. Hence, this approach uses basic science knowledge—e.g., knowledge about biochemistry, anatomy, physiology—to eliminate potentially irrelevant disease processes from consideration (reasoning from hypothesis to data). In contrast, data directed reasoning works in the opposite direction. Specifically, clinicians use patient data such as symptom presentation and history to reach a complete diagnosis. Thus, they engage in an inductive thinking process and use data to formulate a probable hypothesis (Ramoni, Stefanelli, Magnani, & Barosi, 1992).

Other attempts to capture the complexity and nuances that underlie medical reasoning are reflected in the writings of Joseph and Patel (1990), Magnani (1992), and Peirce (1955). These authors have urged the field to consider clinical reasoning from an abductive perspective. That is, they stress thinking that involves the intermixing of deductive and inductive processes and “in the medical context is illustrated by generating a plausible explanatory hypothesis through a process of heuristic rule
utilization [e.g., classification, differentiation, testing]” (Patel, Arocha, & Zhang, 2012, p. 739).

While the direction of reasoning is important, several authors have argued that the field’s attention should be on how medical knowledge is actually organized because that has direct implications for how clinicians store, retrieve, and use information in practice. Blois (1988), in particular, has ruminated about what he refers to as the “horizontal” and “vertical” organization of medical knowledge. On the one hand, Blois (1988) argues, medical knowledge is domain-specific and therefore horizontally organized. Domain-specific knowledge in medicine is analogous to content knowledge in education and law; it is related to a specific subject. In medicine, domain-specific knowledge includes what one has learned about subjects such as physics, chemistry, and anatomy. This kind of knowledge is characterized as having a horizontal orientation because causal explanations can be constructed by drawing on principles from a broad range of areas to test specific hypotheses. As such, it is tied to the hypothetico-deductive process described above. This connection is also the reason that Blois (1998) argued horizontal concepts of medical knowledge were an inappropriate way of thinking about clinical reasoning. For Blois, it was an oversimplified view of medical practice; there are too many other variables involved for the process to be so clean.

Rather, Blois suggested—and Schaffner (1986) agreed—that medical knowledge is in actuality vertically organized. Both scholars argued that some clinicians arrive at diagnoses by comparing a group of symptoms in one patient to a group of symptoms in
another patient and they use that information to reach conclusions about the particulars of disease causation and mechanisms. In other words, rather than reason from principles—explaining particular cases using ideas that can be generalized to other, similar medical cases—some physicians reason by analogy. They also claim that other medical practitioners tend to use the hypothetico-deductive process described above. These observations have been empirically supported through studies conducted by various researchers. For example, Coderre, Mandin, Harasym, and Fick (2003) and Patel and Groen (1991) found that changing one’s diagnostic strategy—from hypothesis-driven to data-driven, in particular—improves accuracy of diagnoses because it is linked to how knowledge is cognitively stored and acquired. Patel and Groen (1991), in particular, also noted that novice clinicians tend to reason using the former approach while expert physicians tend to use the latter approach.

Reasoning Across Other Practice-Based Fields

Up to this point, this literature review has discussed thinking as it has been conceptualized and presented in other disciplines, chiefly, education, law, and medicine. These different perspectives on types of thinking in practice-based fields are useful because they provide a basis for understanding the trajectory and development of evaluative thinking. Specifically, an examination of how other fields have studied and understood the topic of thinking and reasoning generates a comparative lens through which evaluative reasoning can be viewed and understood. It also provides a sense of how a body of research around evaluative thinking could possibly be developed.
As an additional step towards that goal, this section of the literature review offers a comparative overview of themes in the preceding section. Specifically, these themes relate to the types of problems being addressed in other fields, the domains in which reasoning occurs, and how reasoning has been studied. This discussion will serve as a bridge into a more careful examination of the ways in which evaluative thinking has been discussed in the evaluation field.

**General and Specific Types of Problems**

One of the themes that cuts across discussions of thinking in education, law, and medicine relates to the kinds of problems each field seeks to solve and whether these problems are conceptualized generally or specifically. Solving mathematical problems in the education field, for instance, requires transferrable knowledge of basic, computational math as well as knowledge of properties, proofs, and theorems specific to subjects like algebra, trigonometry, and calculus. Likewise, using legal reasoning to settle different kinds of disputes may require expertise in different systems of law, including civil, corporate, educational, family and juvenile, and tax, as well as knowledge of particular statutes (e.g., landlord-tenant, custody, etc.). Similar observations can be made about problem-solving in medicine where both general and specialized clinical knowledge are necessary.

Taking the kinds of problems that have to be solved into account focuses our attention on the question of central interest to each field. For instance, learning is the primary outcome of interest in education, while the sentencing that follows one’s course
of actions is of chief interest in many types of law. In contrast, reaching the correct
diagnosis is instrumental in medicine. Each is a focal issue within its respective field,
but each is only implicitly addressed in discussions about reasoning. When evaluation is
compared with these fields, the contrast is quite stark; there is little agreement about the
defining question or focal unit of interest in evaluation. For example, some evaluation
writers hold fast to the idea that decisions about value and worth are of central interest
(Eisner, 1991; Scriven, 1986). Others contend that evaluation serves other masters;
namely, learning, accountability, and knowledge production (Campbell, 1969, 1973;
Preskill, 2008). Assuming that evaluative thinking is analogous to the kinds of thinking
described in other fields, then this investigation will shed light on what exactly the
central question of interest in evaluation is.

**Where Reasoning Occurs**

The literature that describes thinking, summarized earlier in this review, seems
to assume that reasoning happens within the mind, that it is an unobservable, invisible
process governed by neurological transmissions and synaptic connections. However,
this oversimplified view does not account for the ways that reasoning can be represented
or consider where it can be demonstrated. In the case of pedagogical thinking, for
instance, evidence of instructors’ reasoning can be found in lesson plans, syllabi, class
activities, and lectures. Evidence of students’ critical thinking can be located in their
writing assignments, in the work that is shown when solving mathematical problems, or
through class participation. The planning and doing of these kinds of activities happens
as much inside the classroom as it does outside of it. Similarly, while legal reasoning in
the context of criminal law is mostly discussed in terms of activities that take place
during court trials and appeals, it can be found just as frequently outside of the
courtroom (e.g., when counsel prepares witnesses for the arguments that will be
presented). Likewise, medical reasoning does not only occur in a laboratory setting;
doctor-client interactions serve as valuable sites for observing clinical reasoning,
particularly during the course of a history review or a physical exam.

Activities that unfold in front of a classroom, during a hearing, or in an exam
room might be considered a kind of performance art and the mental preparation and
cognitive work that must be done in advance of the performance provides a more
nuanced view of thinking in these areas. Thus, Newell and Simon’s (1972) work on the
classic notion of the problem space is particularly helpful and pertinent here. The
problem space refers to the contexts in which individuals must solve problems and how
tasks that need solving are represented. Under this framework, problem-solving is
represented as both a mental and a physical activity, as individuals take in contextual
information, transform it into decision information, and engage in behaviors that are
physically recognized as problem-solving (Newell & Simon, 1976). This process is
analogous to what is described in the evaluation literature as being attentive to the
social, historical, political, and organizational contexts of the evaluand (Alkin, 2011;
Patton, 2008). While the implications of viewing problem-solving as a physical
representation of a mental process are clearer for the fields of education, law, and
medicine, they are less clear for evaluation because it has yet to be explicitly and empirically addressed in the literature. This study hopes to address this lack of clarity.

**Efforts to Understand Reasoning**

The preceding discussion of how thinking is understood in other fields begs the question of what sorts of research methods have been used to develop conceptual frameworks and arrive at empirical data that substantiate or refute those frameworks. Similar to research on thinking in other areas not explicitly mentioned in this review (e.g., business and music), research on reasoning in general and problem solving in particular has been conducted using a variety of methods, designs, tools, and techniques.

Every scientific endeavor, by definition, must begin with an observation. The approach to confirming or disconfirming that initial observation diverges from there, taking off in any number of directions. The landscape of educational research—and, in particular, research on reasoning in education—has a history that is dotted with action research studies, case studies, and more recently, comparative qualitative studies (Swanson, O'Connor, & Cooney, 1990; Westerman, 1991). With the exception of a few descriptive studies (Fiedler, 2011; Galanter, 2004), legal reasoning has been fairly understudied, making it a target for criticism within the field of law. Walker (2007), in particular, has launched a research agenda targeting the study of legal reasoning grounded in principles of artificial/computational intelligence. Research on clinical reasoning in the medical field, on the other hand, has witnessed much development and
progress with studies that use decision analysis (Leaper, Horrocks, Staniland, & DeDombal, 1972), protocol analysis (Ericcson & Simon, 1993), and human information processing principles (Elstein, Shulman, & Sprafka, 1978).

The absence of more sophisticated quantitative designs, and experimental techniques in particular, is quite telling concerning the views, conceptualizations, and understandings of reasoning in each of these disciplines. Studies of cognition and reasoning in these areas have highlighted the value of process-oriented approaches to understanding decision- or conclusion-focused outcomes. This is significant for several reasons. First, it suggests that perhaps we do not know enough yet about reasoning and thinking to justify using more sophisticated research methods and tools; descriptive, exploratory methods may be more suitable at this time. Second, it re-affirms the importance of “looking inside the black box” and unpacking the complexity that underlies what may at first seem like a straightforward task—in this case, studying thinking and reasoning in the evaluation field. The implication for this study, in particular, is the justification of the use of a descriptive rather than causal approach for understanding evaluative thinking—a construct that is not only similar to the others, but also emergent and vague in nature.

**Logic & Reasoning in Evaluation**

Evaluation, like education, law, and medicine, is a field rooted in practice. Its identification as a science and an academic area of study is fairly recent (Donaldson, 2007), and only a few writers have explicitly argued for its reputation as such (Rossi &
Freeman, 1985; Scriven, 1994, 2008). As a result, the literature on reasoning and thinking in evaluation is relatively thin when compared to the work that has been done in the other fields described earlier in this review. This is not to say, however, that reasoning is not important or of interest in evaluation. It is simply that discussions in the literature to date have primarily focused on the logic of evaluation and, to a lesser extent, the reasoning that occurs during evaluation.

This portion of the literature review summarizes existing perspectives on evaluation logic as it pertains to reasoning within the field. I will discuss areas in evaluation where evaluative reasoning has been described, and provide an overview of methods that have been previously used to study similar emergent constructs. This section of the literature review will highlight areas of scholarship within evaluation that are well-developed and I will explore how they achieved that status. This section also identifies the meaningful ways in which the current study will contribute to the field.

**Evaluation Logic**

With the exception of a *New Directions for Evaluation (NDE)* volume (Fournier, 1995a) and a few isolated publications dedicated to the topic (e.g., Fournier, 2005; Fournier & Smith, 1993; House, 1977), issues pertaining to logic have received relatively little explicit attention in evaluation. Fournier (1995a) noted in the *NDE* volume’s opening remarks that the motivation for publishing it stemmed from the observation that “professionalized evaluation has spent much of its time and effort on developing methodological sophistication, and less so on logical sophistication,” even though it “has
to be the field’s greatest unmet challenge” (p. 1). As such, Fournier enlisted authors, including Blair, House, Rog, Scriven, and Smith, to contribute their insights concerning the nature of the problem and how the field might collectively address it.

Fournier (1995b) argued early in her own chapter that, “the general logic of evaluation overarches all the various approaches and models within evaluation” (p. 2). She staked this claim on the notions of general and working logic. General logic, as defined by Fournier in this volume and informed by Scriven’s (1980, 1981, 1990, 1993) previous work, captures the procedures for engaging in the conduct of evaluation: (1) establish dimensions against which merit will be judged; (2) construct standards of merit; (3) measure performance and compare it against the outlined standards; and (4) establish a claim about the merit of the entity being evaluated (i.e., the evaluand) based on results of the comparison. These procedures can be broadly applied to various kinds of evaluands (e.g., personnel, product, program, etc.). To Fournier, all evaluation approaches (e.g., empowerment evaluation, realist evaluation, utilization focused evaluation, etc.) could be subsumed under the umbrella term of general logic because they share the same purpose—to determine an entity’s merit or worth—even though each approach’s path to this end goal is qualitatively different. Thus, evaluation approaches (i.e., prescriptive evaluation theories) are considered specific examples of working logic. Evaluators use general and working logic to guide their practice.

While Fournier’s work has been validated and confirmed by other literature, the seemingly contradictory arguments within her article could benefit from additional
clarification. Specifically, she described the parameters that define an evaluation’s scope and boundaries—namely, the problem, phenomenon, question, and claim. These parameters were presented as sources of variation between working and general logic, but they could also be considered synonymous with the four steps of general evaluation logic presented above. Additionally, while she argued that differences in how general logic is followed can be explained by the particularities of these parameters, the argument is unintuitive and seems to be in direct conflict with her previous claim that general evaluation logic is immutable while working evaluation logic is not. Further, although Fournier (1995b) did not explicitly connect these parameters to the evaluation’s context, examination of other authors’ works (e.g., Alkin, 2011; Conner, Fitzpatrick, & Rog, 2012; Vo, 2013) suggests that they could be analogous. Additional clarity regarding these issues is important for the current investigation because logic and reasoning are hand-in-glove. That is, understanding one can lead to an improved understanding of the other. As such, clarification of these issues could potentially guide the interpretation of this study’s findings in the context of existing scholarship.

A number of evaluation scholars have built on Fournier’s perspectives without directly challenging them. Perhaps due to the connections between Fournier’s and Scriven’s work on logic and, in particular, logic in evaluation, extending and relating their ideas to logicians and the field of logic seemed a natural step in the appropriate direction. When addressing the logic of evaluation, for example, Scriven (1995) described it as the cognitive principles that underlie the process of arriving at inferences
and, by extension, decisions about the entity being evaluated. He highlighted the importance of being able to separate poorly constructed evaluative conclusions from those that are not: “Seeing that the dismissive argument and view are unsound is one thing, seeing why they are unsound takes a little more work” (Scriven, 1995, p. 64). The “work” to which Scriven (1995) referred may involve meticulous analysis of how a claim was developed and the evidence called upon to build that claim. This process seems suggestive of what evaluative thinking could entail. However, whether this is the case remains unclear.

Blair (1995), however, took the discussion in a slightly different direction and highlighted a question that logicians and practicing evaluators had in common—namely, is good, non-deductive reasoning (i.e., inductive reasoning) possible? He noted that the consensus among informal logicians is that “there can be logically good, but nonvalid, reasoning” (p. 73). That is, conclusions can be logical and well-reasoned but still come across as either sound or unsound. The implications for evaluation and evaluators are two-fold: (1) how to best corroborate conclusions that have been reached using non-deductive reasoning with deductive reasoning and (2) how to ensure correct and credible conclusions are reached when using non-deductive reasoning alone.

House (1977, 1995) similarly addressed the logic of evaluative arguments, and offered a view that slightly contrasts with Blair’s. Specifically, he posited that at their core the conclusions drawn in evaluation “can be no more than acts of persuasion” because contexts for judgment are dynamic (p. 5). They are in a constant state of
change. Evaluative conclusions, for this reason, require credibility and coherence in presentation.

While House (1977, 1995) takes a different position than Blair (1995), both authors’ ideas converge with Fournier’s (1995b) thinking as it pertains to the notion of working logic as a reasoning pattern, which she described in terms of six features that are common to all inquiry processes: claims, evidence, warrants (i.e., the authority figure that legitimates the conclusions drawn from the given evidence), backings (i.e., other authority figures who support the warrant), exceptions (i.e., conditions under which the warrant would not hold), and qualifiers (i.e., indicators of the claim’s strengths). This is an extension of Toulmin’s (1964) previous work on the structure of claims, which is inextricably tied to how evaluative conclusions are organized and configured. When considered in the broader context of thinking as a kind of cognitive work in other fields, this view of evaluation logic most closely resembles the kind of reasoning needed to solve problems unique to the evaluation discipline. In terms of the present study, this view of evaluation logic and reasoning is the most likely so far to inform the way in which study findings can be understood and related back to the literature.

Together, the ideas represented in this body of scholarship constitute the most explicit views of what evaluative reasoning might mean when it is discussed from and informed by the perspective of logic. The connections between logic and reasoning are clear and easy to understand. What remains ambiguous, however, is the extent to which
these connections can be extended to how evaluative thinking is used in the broader evaluation community. As such, an examination of the extant literature in other areas of evaluation will provide a fuller understanding of this matter and how the field could move forward.

**Evaluative Thinking as Skill**

Being able to apply the logic of evaluation and, perhaps, to “think like an evaluator” implies that some skill set has to be developed and honed. The business of training evaluators has been integral to many evaluation scholars’ work, particularly as they describe the kinds of “hard” technical and “soft” interpersonal skills that evaluators should have. However, this discussion has primarily been framed in terms of general prescriptions for how to conduct evaluations and less often in terms of activities that could be implemented to help novice evaluators hone those skills. In fact, the ways in which actual teaching could occur did not feature prominently in the broader evaluation literature until the 1980s and did not experience a renewed interest until the early 2000s.

An increasing number of scholars have contributed to the teaching evaluation literature following Preskill’s (2000) call for a renewed commitment to the teaching of evaluation, research on teaching and training, and evaluator training. Trevisan’s (2004) review of literature published between 1965 and 2003 indicates that simulations, role-play, discussion groups, single course projects, and practicum experience were most often used to socialize and apprentice novice evaluators into the field. He found that
instructors tended to rely only on learning models found in the adult education literature, and also that, by and large, empirical research on evaluation training was lacking.

A more recent review of the literature was conducted to pinpoint possible nuances in the themes that Trevisan (2004) identified and to identify any new developments in this area between 2004 and 2010 (Vo, 2010). A few salient themes were found, including increased attention to mentoring through experiential learning (Christie & Rose, 2003; Lee, LeBaron Wallace, & Alkin, 2007; Levin-Rozalis & Rosenstein, 2003), evaluator competencies, and the acquisition of evaluation skills (Stevahn, King, Ghere, & Minnema, 2005; Ghere, King, Stevahn, & Minnema, 2006).

The process of acquiring evaluation skills has primarily been discussed in terms of teaching methods, approaches, and activities that help evaluators acquire mostly technical skills. Evaluators also need soft skills, which do have a place within the evaluation literature, but the amount of space and attention given to this topic is minimal and the focus is mainly on the development of interpersonal skills. This is particularly true within the competencies literature.

Thinking and reasoning skills are also a part of this literature base, but they are discussed in terms of decision-making or implied through what and how such skills are taught. Patton (2002b), for example, noted that making decisions based on data is insufficient; rather, people must also be able “to use information, that is, to weigh evidence, consider inevitable contradictions and inconsistencies, articulate values,
interpret findings and examine assumptions” (p. 127). This suggests that evaluative thinking can be expected as an outcome of participating in the kinds of learning activities described above as well as the process of evaluation. Patton (1998) coined the term “process use” to describe this phenomenon and construct that addresses the use of “evaluation logic and processes to help people…learn how to learn….Learning how to think evaluatively is learning how to learn” (p. 226). While such ideas are helpful, it is still difficult to reconcile and extrapolate discussions about evaluative reasoning at the individual level versus the group level. To achieve deeper understanding, we need further clarification about how the construct is defined.

**Evaluative Thinking as Ideology**

The term evaluative thinking is often used in the literature in discussions of groups and organizations; specifically, it is addressed in terms of programmatic and organizational change, culture, or ideology. These discussions can be found in the literature on prescriptive evaluation theory (Alkin, 1991, 2013a; Cousins & Earl, 1992; Greene, 2005; Mark, Henry, & Julnes, 1998; Shadish, Cook, & Leviton, 1991), organizational learning (Preskill & Torres, 1999; Rogers & Williams, 2006), and capacity building (Arnold, 2006; Bourgeois, Chouinard, & Cousins, 2008; Preskill & Boyle, 2008).

Similar to professionals in other disciplines, evaluation scholars and practitioners are deeply invested in their work, and this is evidenced by the abundance of literature on prescriptive evaluation theory. As Alkin and Christie (2004) and Christie and Alkin
(2013) have indicated, all evaluators are concerned about use of evaluation findings, the methods employed to conduct evaluations, and the representation of values in the process. The motivations for pursuing these interests vary, however, as do the anticipated results of such courses of action. For example, some theorists have posited that ensuring use of evaluation will lead to improved decision-making (Cousins, 2013), while others have insisted that use will result in learning and investment in evaluation (Alkin, 2013b; Patton, 2013). Other theorists contend that a focus on rigorous evaluation methods will lead to better decisions (Weiss, 2013) or an improved social condition (Mark & Henry, 2013). There is also a class of theorists who believe ensuring that representing the views and values of broad stakeholder groups will lead to sustained program improvements (Greene, 2013). Thus, there are divergent views even from within the field of evaluation concerning what evaluation is, what purposes it serves, and how those purposes should be realized.

The issue then becomes one of how evaluation’s identity as a field and discipline is constructed and in what ways defining evaluative reasoning contributes to the development of this identity. A small number of divergent but related views on this issue can be found in the evaluation literature. A particularly distinct position that only a few evaluation writers have explicitly espoused pertains to the argument that evaluation is more than a technical activity. Dahler-Larsen (2012), Patton (1998, 2002b) and Schwandt (2002, 2008a, 2008b) are perhaps the strongest supporters of this view.
Patton (1998), in an article describing how he came to think about process use, indicated that evaluation is a culture with specific values not limited to logic, clarity, and specificity. It is a way of thinking that is unnatural and alien to many people, even though evaluators have come to take it for granted. Schwandt (2002) shares this view and indicated that perceiving evaluation as an activity that occurs within a vacuum or as an enterprise characterized by pre-determined and rigid procedures is too limited. Rather, the doing of evaluation is indicative of an evaluator’s “ethical and moral stance...[particular] responsibility...and a commitment to a democratic means” (Schwandt, 2002, p. 157). That is, evaluation is not only a technical and political activity, but also one with moral and ethical underpinnings. Patton further elaborated on this point in a 2002 publication where he indicated that evaluation contributes to the strengthening of democracy over the long term because engaging in the evaluative process results in “1. a more informed electorate through use of findings; and 2. a more thoughtful and deliberative citizenry” (2002b, p. 128). For Patton and Schwandt, doing evaluation, engaging in the process of evaluation, and thinking evaluatively are all ways of being in the world.

Additionally, while Schwandt (2008a) did not set out to define evaluative thinking, his description of an idea he referred to as “intelligent belief in evaluation” offers a philosophical lens through which it might be contemplated. Specifically, he described “a particular attitude and outlook on self and society [that is] demonstrated in a thorough understanding of what is involved in evaluative reasoning as well as a
robustly held, warranted conviction that such reasoning is vital to our well-being” (Schwandt, 2008a, p. 139). Though his writing suggests that he may have been approaching this idea from a sociological perspective, Schwandt’s conceptualization provides a sense of the kinds of cognitive processes and attitudinal indicators that may point to the presence of evaluative thinking at the group level. He has noted elsewhere that the doing of evaluation is “invariably marked by ambiguity, contingency, circumstantiality” (Schwandt, 2002, pp. 18–19). His observation speaks directly to the idea of evaluative reasoning because it deals with the underlying interpretive processes in which leaders within organizations must engage in order to arrive at informed decisions that have broad impact.

Dahler-Larsen (2012) contributed to this ongoing conversation through a sociological analysis of evaluation’s role and place in organizations and society at large. He opened by declaring that “We live in the age of evaluation” (p. 1) and that evaluation has become so enmeshed with our lives and minds that it is “a part of our culture...helping [us] to structure what it is we talk about, what we think are important problems, and what we think we will do about these problems” (p. 4). Further, building on Patton’s and Schwandt’s views that evaluation is a particular way of existing in the world, Dahler-Larsen (2012) indicated that evaluation can be considered a social practice—in particular, one that requires evaluators “to bring to their work many ideals, values, norms, and understandings that are taken for granted as they are embedded in the larger organizational and societal frameworks around evaluation” (p. 29). As such,
participating in such an enterprise calls for a mentality that is oriented towards critical reflection on and an examination of such norms, values, and expectations.

Taken together, this literature base and these authors’ writings suggest that when evaluative thinking is examined through a socio-philosophical lens, it is the input required for the kind of good organizational decision-making that is responsible for a just and fair society. In contrast, when considered through a functional lens, evaluative thinking might be comprised of a set of skills that stem from having undergone a transformational process marked by education and professional socialization. The literature does not suggest how this construct could be studied, however, because no scholars have undertaken such a project. The absence of a clear operational definition is likely a contributing factor.

**Research on Reasoning in Evaluation**

While most ideas underlying evaluative thinking are not entirely new, the construct itself is emergent and has only been recently introduced into the evaluation lexicon. As such, the empirical and theoretical literature bases on the topic are slowly growing while conceptual writings are abundant. In short, not much attention has been explicitly devoted to the empirical exploration and understanding of reasoning in evaluation. Nevertheless, there exist a few constructs that have been systematically examined in evaluation—some in greater depth than others. Constructs that have been more frequently studied and written about in the field include evaluation use, process use, and prescriptive evaluation theory.
Research methods and designs used to study each of these constructs range in their levels of sophistication and complexity. For example, of the various methods that have been employed to study use, simulation studies with factorial designs were most notable in the 1960s and 1970s for separating useful features of evaluation reports from less useful ones (Braskamp, Brown, & Newman, 1978; Newman, Brown, & Littman, 1979; Thompson, Brown, & Furgason, 1981). More recently, systematic reviews and survey studies have been used to gain a broad understanding of process use (Amo & Cousins, 2007; Fleischer & Christie, 2009; Harnar & Preskill, 2007; Preskill & Caracelli, 1997). In contrast, studies of prescriptive evaluation theory have used categorization schemas, content analysis, and model building as investigative tools (Alkin, 2013a; Alkin, Vo, & Hansen, 2013; Harnar, 2012; Shadish, Cook & Leviton, 1991).

With the exception of early experimental studies designed to better understand evaluation use (e.g., Brown & Newman, 1982; Brown, Newman, & Rivers, 1980; Newman, Brown, & Rivers, 1983), the majority of methods and designs used to examine new and developing constructs have been descriptive and exploratory in nature. The scope and purpose of research done on these constructs can inform method and design choice, establishing precedence for the use of similar tools in this investigation.

**The Current Study**

In sum, while the paucity of literature that explicitly describes evaluative thinking poses interesting challenges for fully understanding its meaning, conceptual discussions about evaluative reasoning can be found—implicitly and explicitly—in several areas
within the evaluation knowledge base. The present investigation sought to derive a fuller understanding of evaluative thinking by determining what is at the core of this construct. This was accomplished by engaging with a panel of evaluation experts in an iterative survey study guided by a forecasting method developed by the RAND Corporation—the Delphi technique. Data gathered from the Delphi were analyzed to empirically derive a working definition that can be used to inform future research. This adds to the existing knowledge base by establishing directions for subsequent studies and by making explicit the connections between the various bodies of knowledge previously summarized.
CHAPTER 3
RESEARCH METHODS

This chapter begins with a discussion of the sampling frame that was used to identify participants for the study as well as a brief description of the respondents who completed the Delphi. Descriptions of the study’s design, procedures, tools and instruments, and analytic frameworks are included. Key terms and ideas that guided this investigation are also defined.

Sampling Frame

One of the main purposes of this study was to better understand the ways in which evaluative thinking is conceptualized in the field. As such, it was critical to draw on a purposive sample of evaluation experts. To be considered for inclusion in this study, participants had to be knowledgeable about the topic of evaluative thinking, have effective communication skills, be willing and able to partake in the Delphi process, and have sufficient time to participate. To identify an appropriate sample, I used Alkin and Christie’s (2004) evaluation theory tree to guide the sampling procedures (see also Christie & Alkin, 2008). Prescriptive evaluation theorists who appear on the theory tree have been organized onto three branches based on the relative emphases that their approaches place on use, methods, or valuing. Alkin and Christie (2004) selected these scholars based on their definitive association “with a particular evaluation position...[that is, they appear] to have proposed a particular evaluation orientation,”
and they therefore represent expert thinking on their respective bodies of theory (p. 6). Several experts from each branch of the tree were invited to participate in the current research.

While the evaluation theory tree provided an appropriate starting point for the identification of study participants, a review of the literature suggested that there were others not represented on the tree who had written explicitly and/or made implied comments about evaluative thinking. A sample of these experts was also included in the research. This approach to sampling—drawing from established, recognized experts as well as from the broader field—was employed in order to generate data that are representative of the present thinking about this topic.

**Participants**

Thirty-four experts were identified, individually contacted, and invited to participate in the study. To be considered an expert, participants had to have at least 15 years of research, teaching, and practical experience in the field. Also, experts had to have written explicitly or made implied comments about evaluative thinking in their publications. Appendices A and B contain, respectively, a draft e-mail message and an information sheet about the study that was sent to experts for recruitment purposes and to inform them about the nature and scope of the investigation. The content of the information sheet was developed based on guidelines set forth by the Institutional Review Board at the University of California, Los Angeles.
Upon confirming experts’ willingness to partake in the study, a personalized letter and questionnaire were sent elaborating on the study’s rationale, purposes, and methods (see Appendices C and D). The letter included a brief explanation of the estimated time needed for participation, an assurance of confidentiality, and the investigator’s contact information.

Of the 34 individuals who were contacted, 28 accepted invitations to participate in the investigation and constituted the study’s purposive sample. Further, as presented in Table 3.1, 13 of these 28 evaluation respondents (6 females and 7 males) appear on the evaluation theory tree. The remaining 15 individuals (8 females and 7 males) who participated in the study were identified based on their published work in the broader literature. At the start of the study, 14 female and 14 male participants (n = 28) constituted the sample in Rounds 1 and 2 of survey administration. However, 27 respondents made up the study sample in Round 3 due to Carol Weiss’s unexpected passing. Overall, there was a 100% response rate to all surveys administered.
Table 3.1
Study Participants by Theoretical Orientation

<table>
<thead>
<tr>
<th></th>
<th>Use</th>
<th>Methods</th>
<th>Valuing</th>
<th>Unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chelimsky</td>
<td></td>
<td></td>
<td>Greene</td>
<td>Datta</td>
</tr>
<tr>
<td>Cousins</td>
<td></td>
<td></td>
<td>House</td>
<td>Donaldson</td>
</tr>
<tr>
<td>King</td>
<td></td>
<td></td>
<td>Mertens</td>
<td>Fitzpatrick</td>
</tr>
<tr>
<td>Patton</td>
<td></td>
<td></td>
<td></td>
<td>Fournier</td>
</tr>
<tr>
<td>Preskill</td>
<td></td>
<td></td>
<td></td>
<td>Hood</td>
</tr>
<tr>
<td>Wholey</td>
<td></td>
<td></td>
<td></td>
<td>Hopson</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Julnes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mabry</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Miller</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Morell</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rallis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rog</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Schwandt</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Stevahn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Shadish</td>
</tr>
</tbody>
</table>

The minimum number of respondents needed for a Delphi study ranges from 10 to 15, if their backgrounds are homogenous (Delbecq, Van de Ven, & Gustafson, 1975). In contrast, if participants’ backgrounds are diverse, the expert pool must be increased to ensure representativeness. The number of participants needed for a Delphi is also dependent on the field of study and the questions being addressed (Keeney, Hasson, & McKenna, 2011). Evaluation is a new but developing field with an approximately 40-year history and, as such, the pool of qualified panelists is small. Given the characteristics of the evaluation field and the sampling procedures employed, the pool of 28 study participants was sufficient to obtain an adequate amount of expert feedback.
Study Design

Delphi Technique

To address this study’s research questions, the Delphi technique—an iterative survey method developed by the RAND Corporation to systematically solicit informed opinions—was used (Helmer, 1967a; Hsu & Sandford, 2007). The study involved three rounds of survey administration where questionnaires administered during the second and third rounds were informed by responses from the previous round(s). Informal interviews with a subsample of respondents were conducted between each round to obtain feedback data for subsequent administrations.

Additionally, the Delphi technique—named for the Delphic Oracle at the Temple of Apollo in Greek antiquity—was originally created as a means to systematically gather expert opinions for planning purposes in the defense sector. The first Delphi study, named “Project Delphi” and commissioned by the U.S. Air Force, was completed in the 1950s by the RAND Corporation when its primary purpose was to conduct research on weapons strategy for the U.S. armed forces. The goal of the study was twofold: to determine the best missile target system based on expert opinion data and to estimate how many atomic bombs should be produced to offset the decreased production of other types of artillery. Interestingly, because the majority of this work took place in the middle of the Cold War, the study was conducted from a Soviet strategic planner’s point of view (Linstone & Turoff, 2002; Skulmoski, Hartman, & Krahn, 2007). The method was an attractive approach for addressing “Project Delphi’s” study purposes because the
cost of data collection using alternative approaches would have been prohibitive and the computers available at the time were unable to model the necessary complex simulations.

The Delphi method also witnessed significant early use in the armed services and aerospace research, due at least in part to its cost effectiveness. This approach was frequently used at the height of the Space Race in the 1960s after the Soviet Union successfully launched the first artificial satellite, the Sputnik I. In this era, the Delphi method’s application expanded to include long-range technological forecasting. Results of such studies were used to guide decision-making around which new directions to take in technology research, which technology initiatives to invest in, and which kinds of already-developed technology to withdraw research funding from (Linstone & Turoff, 2002).

While the Delphi method initially fulfilled military planning purposes and saw much early use in related fields, this investigative technique has since been employed to address decision- and policy-making challenges in a variety of other fields. Since Dalkey and Helmer first published the methodological findings stemming from “Project Delphi” in 1963, researchers in business, education, medicine, nursing, technology, and evaluation have all used the approach as a consensus-building and priority-setting tool (Adelson, Alkin, Carey, & Helmer, 1969; Bender, Strack, Ebright, & Von Haunalter, 1969; Christie & Barela, 2005; Day, 2002; Helmer, 1967b; Keeney, Hasson, & McKenna, 2011).
Key characteristics must be in place in order for an investigation to be considered a Delphi study. Such features include the use of a questionnaire that is adapted and altered based on participants’ responses, “anonymous debate-by-questionnaire” (Helmer, 1967a, p. 9), and “iteration with controlled feedback” (Dalkey, Rourke, Lewis, & Snyder, 1972, p. 20). Additionally, while a Delphi study may commence with either an open-ended or structured survey, it must follow a specified sequence of activities. In general, such activities include identifying the question or issue to be addressed, designing a survey that is intended to address that question, selecting a panel of experts, administering the survey to the panelists, evaluating participants’ responses, obtaining and distributing participants’ anonymous feedback on those responses, and redistributing the survey with summary statistics and feedback. These steps are repeated for two or three rounds of survey administration, and the results are interpreted in terms of areas of consensus and dissensus. The final step involves disseminating the study’s findings (Helmer, 1967a; Hsu & Sandford, 2007).

A closed-ended survey was implemented for this particular study; the Delphi procedures for such a design are summarized in Table 3.2. The primary procedural difference between studies that use a closed- versus open-ended survey in the first iteration of the Delphi occurs when feedback is collected. Specifically, when the Delphi starts with a structured, closed-ended questionnaire as it did here, feedback is obtained between the first and second survey administrations. When an open-ended survey is
used in the first iteration, feedback is instead collected as part of the second survey administration (Christie & Barela, 2005; Dalbecq, Van de Ven, & Gustafson, 1975).

The Delphi technique was selected to address this study’s research questions for several reasons, including anonymous consideration of group-level data, controlled feedback, and statistical group response (Dalkey et al., 1972, pp. 20–21). Anonymity was a key consideration in this study because it allowed participating individuals to be more forthcoming about their views on evaluative thinking. Participants accessed each other’s present thinking without knowing each other’s identities, and thus subject bias was minimized (Keeney et al., 2011). Concerns related to “group think,” persuasion by dominant individuals, politics, and other conflicts were circumvented using this approach. Such concerns may not be as easily addressed when respondents are physically assembled in the same space or when they are aware of who has shared which ideas.

This approach also enabled broad participation because respondents were not geographically bound. Additionally, because surveys were administered through e-mail, the response window was flexible and the pressure to produce on-the-spot responses was decreased. As such, participants were able to complete the questionnaires at their convenience and be as thoughtful as their availability allowed (Hurworth, 2004).
Table 3.2
Summary of Procedures for a Delphi Commencing with a Closed-Ended Survey

<table>
<thead>
<tr>
<th>Step #</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Define the question or issue of interest.</td>
</tr>
<tr>
<td>2</td>
<td>Determine methods of analysis.</td>
</tr>
<tr>
<td>3</td>
<td>Determine criteria for establishing consensus.</td>
</tr>
</tbody>
</table>
| 4      | Select study sample.  
|        | a. Determine the number of participants needed for the study.  
|        | b. Identify potential participants.  
|        | c. Recruit participants. |
| 5      | Implement Round 1.  
|        | a. Develop Questionnaire #1 based on review of the literature.  
|        | b. Pilot test the survey.  
|        | c. Determine the method of administration.  
|        | d. Administer the survey. |
| 6      | Analyze data from Questionnaire #1.  
|        | a. Determine summary statistics.  
|        | b. Collect comments from participants regarding outlier items on survey.  
|        | c. Prepare statistical and qualitative data for reporting. |
| 7      | Implement Round 1.  
|        | a. Develop Questionnaire #2 based on results of Questionnaire #1.  
|        | b. Pilot test the survey.  
|        | c. Administer the Questionnaire #2. |
| 8      | Analyze data from Questionnaire #2.  
|        | a. Determine summary statistics.  
|        | b. Collect comments from participants regarding outlier items on survey.  
|        | c. Prepare statistical and qualitative data for reporting. |
| 9      | Implement Round 1.  
|        | a. Develop Questionnaire #3 based on results of Questionnaire #2.  
|        | b. Pilot test the survey.  
|        | c. Administer the Questionnaire #3. |
| 10     | Analyze data from Questionnaire #3.  
|        | a. Determine summary statistics.  
|        | b. Collect comments from participants regarding outlier items on survey.  
|        | c. Prepare statistical and qualitative data for reporting. |
| 11     | Compile and prepare results across all iterations for final reporting. |

1Note: Adapted from Dalbecq, Van de Ven, and Gustafson’s (1975) Table 4.1 (p. 87).  
2Questionnaire #2 includes statistical and qualitative feedback data from Questionnaire #1.  
3Questionnaire #3 includes statistical and qualitative feedback data from Questionnaire #2.
Procedures & Instruments

In keeping with the design of a Delphi study, data in this investigation were collected over the course of three survey iterations and entailed follow-up communication with a select number of participants between rounds of surveys. All communication with respondents related to the study was done electronically, via e-mail.

Round 1 Questionnaire

The first Delphi questionnaire sent to participants contained an introductory letter describing the study, instructions for how to complete the instrument, and a list of 20 statements that described evaluative thinking. Statements were developed from a thorough review of the literature. The primary objective in Round 1 was for participants to rate the relative importance of these 20 statements on a scale of 1 (least important) to 6 (highly important).

The questionnaire had space for experts to record an importance rating for each statement and to add up to five alternative statements that, from their perspective, best described evaluative thinking. Panelists were asked to suggest items that were distinct from the 20 statements that they had already rated. A copy of this instrument can be found in Appendix D.

Round 1 Follow-Up

Following Round 1 survey administration (but before administration of Round 2), additional feedback was sought regarding any of the 20 original statements on which
there was strong dissensus regarding importance level (see Appendix F). Specifically, for each of these items, two panelists—one who had given the rating of “1=least important” and one who had given the rating of “6=highly important”—were asked to provide rationales for their responses. This information was used in the Round 2 questionnaire.

**Round 2 Questionnaire**

The questionnaire used in Round 2 of the Delphi included a description of the criteria used to determine agreement status for the items rated in Round 1 and a list of statements for which consensus had been reached. It also contained 12 statements for which dissensus remained (including select panelists’ reasoning for their answers in Round 1), and a list of eight new items based on panelists’ suggestions in Round 1. A copy of this instrument can be found in Appendix G.

The objective in Round 2 was for panelists to again indicate the importance level of 20 statements that described evaluative thinking. For the 12 statements for which dissensus was not reached in Round 1, participants were asked to consider the rationales that fellow panelists had provided for their earlier responses and to use the same scale to rate these items again. They were also asked to rate for the first time the eight new items that were developed based on panelists’ suggestions in Round 1.

**Round 2 Follow-Up**

As in the follow-up to Round 1, rationales were sought from a pair of panelists—one person who provided a rating of “1=least important,” and one person who rated it as “6=highly important”—for each Round 2 item on which there was strong dissensus
regarding the importance level (see Appendix F). Results of analyses and participants’ commentaries were organized and presented to all respondents for their consideration in the final survey.

**Round 3 Questionnaire**

The final Delphi questionnaire was similar in structure to the Round 2 questionnaire. During the third survey administration, respondents were presented with 14 statements with comments from fellow panelists and asked to rate them again. Panelists had rated nine of these 14 items twice already—in Rounds 1 and 2—while the remaining five items were based on suggestions from Round 1 and therefore had been rated only once previously, in Round 2.

Unlike the previous two surveys, the third survey did not include any new statements because no suggestions were requested in the previous round. The third questionnaire did offer space for panelists to provide optional comments, while the first and second surveys did not. A copy of the Round 3 instrument can be found in Appendix H.

**Post-Delphi Follow-Up**

Results from the final survey administration were shared with panelists after all data had been collected and analyzed. The post-Delphi follow-up message also summarized cumulative study findings (see Appendix I).
Research Tools

Two databases were constructed for the purposes of managing and analyzing the data that were derived from each survey iteration and the feedback process. These databases were created using Microsoft Excel 2008 and the Statistical Package for the Social Sciences 17 (SPSS, version 17).

Definition of Terms

Relative Importance

Respondents rated all statements on a 6-point scale (1=least important; 6=highly important). These measures of importance level were instrumental in determining the panel’s collective opinion about the centrality of the descriptive statements to the concept of evaluative thinking.

While on its own each item’s mean importance rating had no bearing on determining consensus, the mean for all items rated (represented with the horizontal dotted line in Figure 3.1) served as a guidepost for describing the statements’ relationships with each other in latter phases of the study. That is, when a single item’s mean rating was greater than the averaged mean for all items rated (i.e., the item was above the horizontal dotted line), it was considered relatively more important than items whose average ratings were less than the averaged mean (i.e., items that fell below the horizontal dotted line).
Consensus & Dissensus

As previously stated, the Delphi technique was originally developed to solicit opinion data from experts in a systematic fashion for planning purposes; thus, consensus is the primary outcome of interest. In this study, consensus was defined as the extent to which agreement had been reached about an item’s importance level on an individual survey. Specifically, consensus was reached if an item’s variance was less than the averaged variance for all items rated in that round. In Figure 3.1, an item for which consensus had been reached would fall in either Quadrant I or II, where the variance is not only low, but also less than the mean variance for all items rated (represented with the vertical dotted line at the center of the figure). In contrast, if the item’s mean
variance exceeded the averaged variance for all items rated (e.g., the item fell in either Quadrant III or IV, where the variance is high), dissensus remained and the item was included for re-rating in the next survey administration.

Traditionally, consensus criteria in Delphi studies are determined a priori and can be defined along the lines of any number of metrics, including: the percentage of votes in favor of a given item if the scale of measurement is dichotomous (Miller, 2006); the percentage of votes that fall into two adjacent categories if the outcome is measured on an ordinal or continuous scale (Ulschak, 1983); or the stability of panelists’ ratings as determined by measures of central tendency (Hasson, Keeney, & McKenna, 2000; Scheibe, Skutsch, & Schofer, 2002). Murray and Jarman (1987), for instance, argued for using mean ratings, while others have expressed a strong preference for using median values as a measure of consensus (Eckman, 1983; Hill & Fowles, 1975; Jacobs, 1996). Others, such as Ludwig (1994), have indicated that use of the mode is most appropriate for determining convergence because “the mean or median could be misleading,” particularly when “there was the possibility of polarization or clustering of the results around two or more points” (p. 57).

The suggestions outlined above are likely appropriate for analysis of individual rounds of survey data independently; however, measures of central tendency and distributions of percentages across categories provided only a limited understanding of the nature of consensus and dissensus in this study. The median, for example, was helpful for the purposes of determining the level of agreement reached on a single
survey item. Particularly, the median provided initial understandings about consensus when the distribution of ratings was skewed because it is, by definition, robust against outliers. However, normal and close-to-normal distributions were also observed and in these cases, use of the mean provided a more precise estimate of the level of consensus reached among participants. As such, the mean was examined consistently throughout this study.

Similarly, information summarized by the mean, median, and mode only provided insights about the relative importance panelists placed on each rated statement. These summary statistics were insufficient to provide deeper understandings about the nature of disagreement within the group. Examining proportions of ratings across categories did not offer a clear sense of the extent or nature of dissensus (e.g., the degree of variability in responses) either. Importantly, both were masked when using these metrics, while agreement was in the foreground. As such, variance—a measure of dispersion—was also used in this study to provide measures of difference in opinion.

**Analyses**

**Round 1: Quantitative Analyses**

To determine whether consensus had been reached on a given statement after all Round 1 survey data were collected, and to exclude it from the pool of items to be rated in the second round, mean ratings and variances were calculated for all 20 statements and these were plotted against the averaged mean and variance on a scattergram. Items whose variances were less than the averaged variance were excluded from the pool of
statements to be rated in Round 2; that is, agreement about their importance levels had been reached and thus they appeared in either Quadrants I or II in Figure 3.1. Conversely, items whose variances were greater than the averaged variance (i.e., they appeared in either Quadrants III or IV in Figure 3.1) were included for rating in the subsequent round. This process led to the identification of 12 statements that experts were asked to rate in Round 2. Additionally, items whose variance and mean were high (i.e., they were in Quadrant I) were considered highly important to the notion of evaluative thinking; items with low means and high variance (in Quadrant II) were considered relatively less important.

**Round 1: Qualitative Analyses**

Twenty-four of 28 participants responded to the request for suggested additional statements about evaluative thinking, yielding a total of 78 statements that could have potentially been included in the Round 2 questionnaire. Each statement was examined and compared to the others to determine the extent of overlap in the ideas that were expressed. Codes and categories were then inductively developed to describe each statement’s focus. This iterative process led to a number of statements being collapsed, and it reduced the overall pool of suggested statements from 78 to 33 items.

Next, these 33 statements were individually examined to determine the general ideas that were represented in each. They were then placed into like groupings. Subsequently, statements were randomly selected for inclusion in Round 2. The number of statements chosen from each category was proportional to the total number of
statements within that particular category. In an effort to respect panelists’ time and to ensure that the survey remained at a reasonable length, the maximum number of suggested statements that could have been selected was eight. Thus, eight items were randomly selected from the collection of 33 suggested statements, and these were combined with the 12 statements that remained from Round 1. This formed the new list of 20 statements that panelists rated in Round 2 of the study.

**Round 2: Quantitative Analyses**

Twenty items were rated in Round 2. Similar to the quantitative analyses described in Round 1, the mean rating and variance for each item was calculated and plotted against the averaged mean and averaged variance on a scattergram. This helped to determine if consensus had been reached and whether the item should be excluded from the pool of items to be rated in Round 3. Again, items whose variances were less than the averaged variance were excluded from the pool of statements to be rated in the final round. Conversely, items whose variances exceeded the averaged variance were included for rating in Round 3. This process led to the identification of 14 statements for which experts were asked to provide new ratings in Round 3.

**Round 3: Quantitative Analyses**

As was the case in Rounds 1 and 2, quantitative analyses in Round 3 involved determining the mean ratings and variances for 14 statements and plotting them against the averaged mean and variance on a scattergram. It was determined that consensus
was reached for items whose variances were less than the averaged variance. In contrast, dissensus remained for items whose variances exceeded the averaged variance.

**Cumulative Quantitative Analyses**

In addition to understanding areas of consensus and dissensus among panelists, arriving at a working definition of “evaluative thinking” was also a priority in this study. As such, it was important to determine the importance level that panelists placed on each of the statements they rated throughout the investigation. To accomplish this task, mean ratings and variances of all items rated were compared within each survey administration as well as cumulatively.

**Determining overall relative importance.** It was hoped that throughout the study, the variance of panelists’ ratings would gradually decrease as they reached consensus about the important concepts in evaluative thinking. It was also hoped that mean ratings for individual statements would be distributed along the 6-point scale, indicating different levels of importance for different statements. That is, rated statements would shift from Quadrants III and IV (high variance) to Quadrants I and II (low variance), as illustrated in Figure 3.2.
Determining statements’ relative importance under this analytic framework entailed tabulating their final mean ratings and respective standard errors, placing them in descending order, coding their contents by the domain emphasized in each, and examining how items clustered by domain and mean rating. Items that clustered in Quadrant I had consistently high mean ratings, and thus were considered of great importance to the notion of evaluative thinking. Items that clustered in Quadrant II, on the other hand, had consistently low mean ratings and were therefore considered of less importance to the notion of evaluative thinking.
In terms of domain, items were examined to determine whether they addressed reasoning, practice, valuing, or multiple issues (see Table 3.3). Those that fell into the reasoning domain required one to cognitively navigate through a number of decision points to reach a conclusion that would lead to action (e.g., considering the credibility of different kinds of evidence in context). Statements that fell into the practice domain involved behaviors tied to the conduct of evaluation (e.g., offering evidence for claims that one makes). Those items that dealt with personal conceptions of the nature and purpose of evaluation (e.g., endorsing the idea that evaluations are conducted for the purposes of assuaging social inequities) were placed in the valuing domain. And finally, items that touched on more than one domain were placed into the “multiple” category.

After the frequency and distribution of domain codes were examined, items were grouped by importance level to shed light on how panelists prioritized the 28 statements rated over three Delphi rounds. This was accomplished by sorting statements in descending order according to their mean ratings and then comparing the extent of overlap between the upper and lower limits of statements’ confidence intervals. More precisely, if an item’s upper limit overlapped with the lower limit of the preceding item, those statements’ mean importance ratings were considered similar to each other and, thus, they were grouped together. If a statement’s upper limit did not overlap with the preceding statement’s lower limit, then the items were rated differently on importance and a natural boundary, or cut point, was established between the two potential groups
of statements. Finally, each group of statements’ importance level was determined by their averaged mean rating.

**Understanding consensus and dissensus.** Unpacking the nature of agreement and disagreement among panelists required more in-depth analysis of the particular ideas conveyed in each of the rated statements. The statements’ themes were determined by examining the relative emphasis placed on each of the domain codes—reasoning, practice, and valuing. This offered insight into what types of statements (i.e., which domains) lent themselves most easily to consensus among these expert panelists.

Table 3.3
*Inventory of 28 Descriptive Statements by Domain*

<table>
<thead>
<tr>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statement</strong></td>
</tr>
<tr>
<td>S2</td>
</tr>
<tr>
<td>S3</td>
</tr>
<tr>
<td>S4</td>
</tr>
<tr>
<td>S5</td>
</tr>
<tr>
<td>S6</td>
</tr>
<tr>
<td>A6</td>
</tr>
<tr>
<td>A8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statement</strong></td>
</tr>
<tr>
<td>S15</td>
</tr>
<tr>
<td>S16</td>
</tr>
<tr>
<td>S17</td>
</tr>
<tr>
<td>S18</td>
</tr>
<tr>
<td>S19</td>
</tr>
<tr>
<td>S20</td>
</tr>
<tr>
<td>A3</td>
</tr>
<tr>
<td>A5</td>
</tr>
<tr>
<td>Statement</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td><strong>Valuing</strong></td>
</tr>
<tr>
<td>S7</td>
</tr>
<tr>
<td>S8</td>
</tr>
<tr>
<td>S9</td>
</tr>
<tr>
<td>S10</td>
</tr>
<tr>
<td>S11</td>
</tr>
<tr>
<td>S12</td>
</tr>
<tr>
<td>A1</td>
</tr>
<tr>
<td>A2</td>
</tr>
<tr>
<td>A4</td>
</tr>
<tr>
<td><strong>Multiple</strong></td>
</tr>
<tr>
<td>S1</td>
</tr>
<tr>
<td>S13</td>
</tr>
<tr>
<td>S14</td>
</tr>
<tr>
<td>A7</td>
</tr>
</tbody>
</table>
CHAPTER 4
RESULTS

This chapter presents results from all rounds of data collection. Findings from individual survey rounds are discussed first to demonstrate how final ratings of importance emerged across the study administration. These findings are then examined more holistically to address the broader questions that guided the investigation. Specifically, the findings shed light on what experts believe is important to evaluative thinking and how readily consensus emerged for specific aspects of the construct.

Round 1 Results

Quantitative Findings

In Round 1, the averaged mean rating and averaged variance of all 20 statements were 4.18 and 1.73, respectively. Mean ratings of individual statements ranged from 2.46 (Statement S9) to 5.57 (Statement S18), while their variances ranged from 0.74 (Statement S3) to 3.12 (Statement S10). Table 4.1 summarizes this information. It also indicates the statement numbers assigned to the items rated in this round, the statements themselves, and the summary statistics tabulated for them.

Examination of 20 statements’ mean ratings and variances relative to the averaged mean ($\bar{x}_g = 4.18$) and averaged variance ($v_g = 1.73$) led to the identification of eight statements for which panelists reached consensus concerning importance level. These items included Statements S2 through S7, as well as S18 and S19 (denoted with
asterisks in Table 4.1). All of these statements except S7 had mean ratings higher than the averaged mean of 4.18, and were therefore considered important to evaluative thinking. Statement S7 (\( \bar{x} = 2.50 \)) was deemed relatively less important.

Table 4.1
Summary Statistics for 20 Descriptive Statements Rated in Round 1

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>( \bar{x} )</th>
<th>( s^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
<td>5.14</td>
<td>2.57</td>
</tr>
<tr>
<td>S2*</td>
<td>I consider the availability of resources when setting out to conduct an evaluation.</td>
<td>4.68</td>
<td>1.26</td>
</tr>
<tr>
<td>S3*</td>
<td>I consider the importance of various kinds of data sources when designing an evaluation.</td>
<td>4.93</td>
<td>0.74</td>
</tr>
<tr>
<td>S4*</td>
<td>I consider alternative explanations for claims.</td>
<td>5.25</td>
<td>0.79</td>
</tr>
<tr>
<td>S5*</td>
<td>I consider inconsistencies and contradictions in explanations.</td>
<td>4.79</td>
<td>1.21</td>
</tr>
<tr>
<td>S6*</td>
<td>I consider the credibility of different kinds of evidence in context.</td>
<td>4.96</td>
<td>1.29</td>
</tr>
<tr>
<td>S7*</td>
<td>I conduct evaluation with an eye towards challenging personal beliefs and opinions.</td>
<td>2.50</td>
<td>1.37</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>2.61</td>
<td>2.40</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
<td>2.46</td>
<td>1.89</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
<td>3.68</td>
<td>3.12</td>
</tr>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
<td>4.18</td>
<td>1.71</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
<td>3.43</td>
<td>2.99</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
<td>3.39</td>
<td>2.25</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
<td>4.14</td>
<td>2.20</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
<td>3.57</td>
<td>2.03</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
<td>4.39</td>
<td>1.65</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
<td>4.36</td>
<td>1.87</td>
</tr>
<tr>
<td>S18*</td>
<td>I offer evidence for claims that I make.</td>
<td>5.57</td>
<td>0.85</td>
</tr>
<tr>
<td>S19*</td>
<td>I make decisions after carefully examining systematically collected data.</td>
<td>5.36</td>
<td>0.83</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
<td>4.14</td>
<td>1.61</td>
</tr>
</tbody>
</table>

\[ \text{Averaged Mean and Variance} \ 4.18 \ 1.73 \]

Note: Panelists reached consensus in Round 1 on items marked with asterisks (*).
Further, applying the consensus criteria described in Chapter 3 led to the conclusion that disagreement remained for nine of the original 20 statements. As shown in Figure 4.1, the variances of these items (Statements S1, S8–S10, S12–S15, and S17) exceeded the averaged variance. While Statements S11, S16, and S20 technically met consensus criteria, the extent of consensus was ambiguous and arguable because their means and variances were so close to the averaged values for all statements; thus, they were included in the list of items to be rated in Round 2. Therefore, in total, 12 statements from the original list of 20 were identified for re-rating in the second round.

Figure 4.1
Scattergram of Mean Ratings and Variances of 20 Statements Rated with Respect to Averaged Mean and Variance (Round 1)
Qualitative Findings

As described in Chapter 3, in Round 1 respondents contributed 78 additional statements that they considered important to evaluative thinking. Qualitative analyses reduced the overall pool of potential items to 33 (see Table 4.2). These 33 statements were then individually examined to determine the general ideas that were represented in each, so that they could be coded and placed into like groupings. In an effort to respect participants’ time and ensure that the length of the second survey did not exceed that of the first survey, the maximum number of suggested statements that could have been selected was eight.

Thus, eight items were selected from the collection of 33 suggested statements. The number randomly selected from each grouping was proportional to the number of items within the grouping. These eight items included Statements A5, A13, A15, A17–19, A30, and A31 (marked with asterisks in Table 4.2). The items were renumbered from A1 to A8 to prevent confusion on the second Delphi questionnaire, and they were then combined with the 12 statements mentioned above to form the list of 20 statements that panelists rated in Round 2 of the study.
<table>
<thead>
<tr>
<th>Statement #</th>
<th>Suggested Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>I allow existing evaluative principles and evaluation experience to guide the iterative process by which I conduct the evaluation (including how I choose to respond to shifting priorities and changing contexts).</td>
</tr>
<tr>
<td>A2</td>
<td>I allow flexibility in practice to accommodate different evaluation contexts.</td>
</tr>
<tr>
<td>A3</td>
<td>I conduct evaluation with an eye towards different audiences' intended use of findings by grounding data in evidence that they consider credible.</td>
</tr>
<tr>
<td>A4</td>
<td>I allow the evaluation question to guide all decisions during the study so that credible, useful evidence is produced.</td>
</tr>
<tr>
<td>A5*</td>
<td>Not everything can or should be professionally evaluated.</td>
</tr>
<tr>
<td>A6</td>
<td>I conduct evaluations with an eye towards questioning my own assumptions and preconceptions.</td>
</tr>
<tr>
<td>A7</td>
<td>I do evaluation if I have the appropriate expertise.</td>
</tr>
<tr>
<td>A8</td>
<td>I engage stakeholders in the process of interpreting and using evaluation data to make decisions.</td>
</tr>
<tr>
<td>A9</td>
<td>I engage multiple program experts in the process of establishing criteria by which the evaluand will be judged.</td>
</tr>
<tr>
<td>A10</td>
<td>I engage stakeholders in the process of making underlying program values explicit.</td>
</tr>
<tr>
<td>A11</td>
<td>I ask various stakeholder groups questions to help me understand how the evaluation will satisfy their information needs.</td>
</tr>
<tr>
<td>A12</td>
<td>Before starting a new initiative or revising an old one, I engage in conversations with others about how data will be collected and shared.</td>
</tr>
<tr>
<td>A13*</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
</tr>
<tr>
<td>A14</td>
<td>I attend to equity issues by ensuring that voices of the &quot;less powerful&quot; are legitimately and accurately represented.</td>
</tr>
<tr>
<td>A15*</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
</tr>
<tr>
<td>A16</td>
<td>I think about the multiple goals that evaluation can achieve.</td>
</tr>
<tr>
<td>A17*</td>
<td>I consider stakeholders' explicit and implicit reasons for commissioning the evaluation.</td>
</tr>
<tr>
<td>A18*</td>
<td>I think about the criteria that would qualify an evaluation as “good” or “bad.”</td>
</tr>
<tr>
<td>A19*</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
</tr>
<tr>
<td>A20</td>
<td>I consider the quality of evidence that is used to build evaluative claims.</td>
</tr>
<tr>
<td>A21</td>
<td>I consider the social dimensions of evaluation practice.</td>
</tr>
<tr>
<td>A22</td>
<td>I think about the intended and unintended impact that evaluation findings could have on program participants.</td>
</tr>
<tr>
<td>A23</td>
<td>I consider the unintended consequences of conducting the evaluation.</td>
</tr>
</tbody>
</table>
Table 4.2
Reduced List of 33 Suggested Statements Respondents Provided During Round 1, cont.

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Suggested Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A24</td>
<td>I think about the extent to which values of various stakeholder groups are represented.</td>
</tr>
<tr>
<td>A25</td>
<td>I think about the ways in which stakeholders' priorities impact the evaluation process.</td>
</tr>
<tr>
<td>A26</td>
<td>I think about potential use of evaluation results by potential users.</td>
</tr>
<tr>
<td>A27</td>
<td>I recognize tradeoffs between cost, quality, and time when trying to creatively obtain credible data to address evaluation questions.</td>
</tr>
<tr>
<td>A28</td>
<td>I work with community members to ensure use of the evaluation to support social justice and human rights.</td>
</tr>
<tr>
<td>A29</td>
<td>I do evaluations with an eye towards generating knowledge that will be used to support decision making.</td>
</tr>
<tr>
<td>A30*</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
</tr>
<tr>
<td>A31*</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
</tr>
<tr>
<td>A32</td>
<td>I consider the chain of reasoning linking the evaluand to desired change.</td>
</tr>
<tr>
<td>A33</td>
<td>I compare the best evidence available from multiple sources to confirm and disconfirm evaluative claims.</td>
</tr>
</tbody>
</table>

Note: Items marked with asterisks (*) were included for rating in the Round 2 questionnaire. They were first renumbered as A1—A8 (in the order that they appear here).

Round 2 Results

Quantitative Findings

In Round 2, the averaged mean rating and averaged variance of all 20 survey items were 4.08 and 2.26, respectively. Mean ratings of individual statements ranged from 2.57 (Statement S8) to 5.04 (Statement S17), while their variances ranged from 0.81 (Statement S11) to 3.51 (Statement A7). Table 4.3 summarizes this information for each statement that was rated during Round 2. It also indicates the statement numbers assigned to the items rated in this round, the statements themselves, and the summary statistics tabulated for them.
Table 4.3
Summary Statistics for 20 Descriptive Statements Rated in Round 2

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>$\bar{x}$</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
<td>4.50</td>
<td>2.70</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>2.57</td>
<td>2.62</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
<td>2.68</td>
<td>2.74</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
<td>4.32</td>
<td>2.15</td>
</tr>
<tr>
<td>S11*</td>
<td>I conduct evaluation with an eye towards transparency.</td>
<td>4.93</td>
<td>0.81</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
<td>3.61</td>
<td>2.84</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
<td>3.93</td>
<td>2.96</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
<td>3.75</td>
<td>2.42</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
<td>3.18</td>
<td>2.74</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
<td>4.86</td>
<td>2.13</td>
</tr>
<tr>
<td>S17*</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
<td>5.04</td>
<td>1.81</td>
</tr>
<tr>
<td>S20*</td>
<td>I set aside time to reflect on the way I do my work.</td>
<td>4.11</td>
<td>1.43</td>
</tr>
<tr>
<td>A1</td>
<td>Not everything can or should be professionally evaluated.</td>
<td>4.11</td>
<td>2.62</td>
</tr>
<tr>
<td>A2</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
<td>4.50</td>
<td>2.11</td>
</tr>
<tr>
<td>A3*</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
<td>4.68</td>
<td>1.86</td>
</tr>
<tr>
<td>A4</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
<td>3.61</td>
<td>2.25</td>
</tr>
<tr>
<td>A5*</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
<td>4.64</td>
<td>1.79</td>
</tr>
<tr>
<td>A6*</td>
<td>I consider stakeholders’ explicit and implicit reasons for commissioning the evaluation.</td>
<td>4.57</td>
<td>1.66</td>
</tr>
<tr>
<td>A7</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
<td>3.89</td>
<td>3.51</td>
</tr>
<tr>
<td>A8</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
<td>4.18</td>
<td>2.08</td>
</tr>
</tbody>
</table>

**Averaged Mean and Variance**

4.08  2.26

Note: Panelists reached consensus in Round 2 on items marked with asterisks (*).
Examination of 20 statements’ mean ratings and variances relative to the averaged mean ($\bar{x}_g = 4.08$) and averaged variance ($v_g = 2.26$) led to the identification of six statements for which panelists had reached consensus concerning importance level: Statements S11, S17, S20, A3, A5, and A6 (denoted with asterisks in Table 4.3). All of these items had relatively high mean importance ratings ($\bar{x} > 4.08$) and were therefore considered important to evaluative thinking.

Applying the consensus criteria previously described led to the conclusion that disagreement remained for nine statements. As illustrated in Figure 4.2, these items (S1, S8, S9, S12–S15, A1, and A7) had variances that exceeded the averaged variance. In addition, while Statements S10 and S16 met consensus criteria, the extent of consensus was arguable because averaged variance for all statements had increased between rounds (from 1.73 in Round 1 to 2.26 in Round 2) and thus the stability of these ratings was questionable. As a result, they were also included on the list of items to be rated in Round 3. Finally, Statements A2, A4, and A8 also met consensus criteria, but baseline ratings for these items were established in the current round. Due to the relatively high variance in Round 2 and the absence of earlier bases for comparison, the extent of agreement for these suggested items was also questionable. As such, they were prudently included for re-rating in Round 3. A total of 14 statements were identified for re-rating in Round 3.
**Round 3 Results**

**Quantitative Findings**

In the final survey round, the averaged mean rating and averaged variance of all items were 3.74 and 2.53, respectively. Mean ratings of individual statements ranged from 2.26 (Statement S9) to 4.74 (Statement S16), while their variances ranged from 1.58 (Statements S9 and S16) to 3.23 (Statement S14). Table 4.4 summarizes this information for the 14 statements that were rated during Round 3. It also indicates the statement numbers assigned to the items rated in Round 3, the statements themselves, and the summary statistics tabulated for them.
Table 4.4
Summary Statistics for 14 Descriptive Statements Rated in Round 3

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>$\overline{x}$</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
<td>4.52</td>
<td>2.57</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>3.33</td>
<td>2.62</td>
</tr>
<tr>
<td>S9*</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
<td>2.26</td>
<td>1.58</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
<td>3.56</td>
<td>2.79</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
<td>3.04</td>
<td>2.65</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
<td>3.56</td>
<td>2.79</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
<td>4.33</td>
<td>3.23</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
<td>3.48</td>
<td>2.49</td>
</tr>
<tr>
<td>S16*</td>
<td>I question claims and assumptions that others make.</td>
<td>4.74</td>
<td>1.58</td>
</tr>
<tr>
<td>A1*</td>
<td>Not everything can or should be professionally evaluated.</td>
<td>4.30</td>
<td>1.99</td>
</tr>
<tr>
<td>A2</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
<td>3.89</td>
<td>2.87</td>
</tr>
<tr>
<td>A4*</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
<td>3.37</td>
<td>2.17</td>
</tr>
<tr>
<td>A7</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
<td>4.07</td>
<td>3.15</td>
</tr>
<tr>
<td>A8</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
<td>3.85</td>
<td>2.90</td>
</tr>
</tbody>
</table>

**Averaged Mean and Variance** 3.74 2.53

Note: Panelists reached consensus in Round 3 on items marked with asterisks (*).

Examination of 14 statements’ mean ratings and variances relative to the averaged mean ($\overline{x}_g = 3.74$) and averaged variance ($s^2_g = 2.53$) led to the identification of four statements for which panelists had reached consensus concerning importance level: Statements S9, S16, A1, and A4 (denoted with asterisks in Table 4.4). Based on the mean
importance ratings of these four items relative to the averaged mean, S16 (\(\bar{x} = 4.74\)) and A1 (\(\bar{x} = 4.30\)) were considered more important to evaluative thinking, and S9 (\(\bar{x} = 2.26\)) and A4 (\(\bar{x} = 3.37\)) were considered less important.

As was the case in previous analyses, consensus criteria were applied to Round 3 data. This led to the conclusion that disagreement remained for nine of the 14 statements. As shown in Figure 4.3, the variances of these items (S1, S8, S10, S12–S14, A2, A7, and A8) exceeded the averaged variance (\(v_g = 2.53\)). In addition, Statement S15 met consensus criteria, but it remained on the cusp due to the instability of mean ratings and variances throughout the investigation. In particular, Statement S15's variance increased between Rounds 1 and 2 (from 2.03 to 2.74), but decreased between Rounds 2 and 3 (from 2.74 to 2.49). As such, the extent of consensus remained uncertain and this item was not included in the list of items for which agreement was reached. Thus, in Round 3, it was determined that panelists reached consensus on four statements, while dissensus remained for 10 others.
Cumulative Results

The preceding discussion of results within each round of data collection provides an in-depth, process-oriented understanding of panelists’ thinking during the Delphi. While this sheds important light on how the findings were derived, its fractured nature means its utility is somewhat limited. In particular, it does not clearly highlight the broader study findings. Thus, the current section discusses these cumulative results in terms of the research questions that guided this investigation:

1. What do evaluation experts consider important to evaluative thinking?
2. What is the nature of consensus and dissensus among experts?
What is Important to Evaluative Thinking?

Panelists rated the importance level of a total of 28 statements throughout the Delphi. These statements addressed a range of issues (or domains), from reasoning and values to issues of practice. The current section describes how all 28 items were distributed across the 6-point importance scale, paying specific attention to where statements within each of the three domains tended to fall on the scale. This provides a better sense of the ideas participants considered most central to evaluative thinking.

In Figure 4.4, all of the 28 statements that participants rated during the Delphi are placed in descending order along the 6-point scale on which they were rated ($\bar{x}_g = 4.22, SE(\bar{x}_g) = 0.26$). Sixteen of the statements’ mean ratings were greater than the averaged mean rating; their values ranged from 4.30 (Statement A1) to 5.57 (Statement S18). Twelve of these 16 items were equally distributed between the domains of reasoning ($n = 6$; in descending mean order: S4, S6, S3, S5, S2, A6) and practice ($n = 6$; in descending mean order: S18, S19, S17, S16, A3, A5) (see also Table 4.5). The remaining four items either fell into the valuing domain ($n = 2$; S11, A1) or into multiple domains ($n = 2$; S1, S14). In this case, Statements S1 and S14 both touched on reasoning and practice. (Statements pertaining to multiple domains are further discussed in a later section of this chapter.)
The remaining 12 of the 28 items’ mean ratings were lower than the averaged mean rating ($\bar{x}_g = 4.22$). These values ranged from 2.26 (Statement S9) to 4.11 (Statement S20). As illustrated in Figure 4.4, seven of these items fell into the valuing domain (in descending mean order: A2, S10, A4, S8, S12, S7, S9). Two of these 12 items dealt with practice issues (in descending mean order: S20, S15), one (A8) dealt with reasoning, and the remaining two statements fell into multiple domains—Statement A7 addressed issues of valuing and reasoning while Statement S13 dealt with practice and valuing.
Table 4.5
Descriptive Statistics for 28 Statements Rated During the Delphi with Domain Labels, In Order of Importance to Evaluative Thinking

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>( \bar{x} )</th>
<th>( SE_{\bar{x}} )</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>S18</td>
<td>I offer evidence for claims that I make.</td>
<td>5.57</td>
<td>0.17</td>
<td>Practice</td>
</tr>
<tr>
<td>S19</td>
<td>I make decisions after carefully examining systematically collected data.</td>
<td>5.36</td>
<td>0.17</td>
<td>Practice</td>
</tr>
<tr>
<td>S4</td>
<td>I consider alternative explanations for claims.</td>
<td>5.25</td>
<td>0.17</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
<td>5.04</td>
<td>0.25</td>
<td>Practice</td>
</tr>
<tr>
<td>S6</td>
<td>I consider the credibility of different kinds of evidence in context.</td>
<td>4.96</td>
<td>0.22</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S3</td>
<td>I consider the importance of various kinds of data sources when designing an evaluation.</td>
<td>4.93</td>
<td>0.16</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
<td>4.93</td>
<td>0.17</td>
<td>Valuing</td>
</tr>
<tr>
<td>S5</td>
<td>I consider inconsistencies and contradictions in explanations.</td>
<td>4.79</td>
<td>0.21</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
<td>4.74</td>
<td>0.24</td>
<td>Practice</td>
</tr>
<tr>
<td>S2</td>
<td>I consider the availability of resources when setting out to conduct an evaluation.</td>
<td>4.68</td>
<td>0.21</td>
<td>Reasoning</td>
</tr>
<tr>
<td>A3</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
<td>4.68</td>
<td>0.26</td>
<td>Practice</td>
</tr>
<tr>
<td>A5</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
<td>4.64</td>
<td>0.25</td>
<td>Practice</td>
</tr>
<tr>
<td>A6</td>
<td>I consider stakeholders’ explicit and implicit reasons for commissioning the evaluation.</td>
<td>4.57</td>
<td>0.24</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
<td>4.52</td>
<td>0.31</td>
<td>Reasoning/Practice</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
<td>4.33</td>
<td>0.35</td>
<td>Practice/Reasoning</td>
</tr>
</tbody>
</table>
Table 4.5
Descriptive Statistics for 28 Statements Rated During the Delphi with Domain Labels, In Order of Importance to Evaluative Thinking, cont.

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>$\bar{x}$</th>
<th>$SE_{\bar{x}}$</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Not everything can or should be professionally evaluated.</td>
<td>4.30</td>
<td>0.27</td>
<td>Valuing</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
<td>4.11</td>
<td>0.23</td>
<td>Practice</td>
</tr>
<tr>
<td>A7</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
<td>4.07</td>
<td>0.34</td>
<td>Valuing</td>
</tr>
<tr>
<td>A2</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
<td>3.89</td>
<td>0.33</td>
<td>Valuing</td>
</tr>
<tr>
<td>A8</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
<td>3.85</td>
<td>0.33</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
<td>3.56</td>
<td>0.32</td>
<td>Valuing</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
<td>3.56</td>
<td>0.32</td>
<td>Practice Valuing</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
<td>3.48</td>
<td>0.30</td>
<td>Practice</td>
</tr>
<tr>
<td>A4</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
<td>3.37</td>
<td>0.28</td>
<td>Valuing</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>3.33</td>
<td>0.31</td>
<td>Valuing</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
<td>3.04</td>
<td>0.31</td>
<td>Valuing</td>
</tr>
<tr>
<td>S7</td>
<td>I conduct evaluation with an eye towards challenging personal beliefs and opinions.</td>
<td>2.50</td>
<td>0.22</td>
<td>Valuing</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
<td>2.26</td>
<td>0.24</td>
<td>Valuing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Averaged Mean and Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.22</td>
</tr>
</tbody>
</table>

These results suggest that evaluative thinking has more to do with reasoning and practice than it does with valuing (i.e., one’s personal conceptions about evaluation and
its purpose), because statements in the reasoning and practice domains were rated as more important. Importantly, because participants appear to have placed equal emphasis on reasoning and practice, this finding needs to be unpacked further to determine if the relative importance of the two domains can be clarified.

This can be accomplished by grouping items according to their mean importance ratings and using the upper and lower limits of the statements’ confidence intervals to determine a cut point for each group. That is, if an item’s upper limit overlaps with the lower limit of the preceding item, those statements’ mean importance ratings are considered similar to each other and, thus, they are grouped together. If a statement’s upper limit does not overlap with the preceding statement’s lower limit, then the items are rated differently on importance and a natural boundary, or cut point, can be established between the two potential groups of statements.

For example, by tracing the first horizontal dotted line in Figure 4.5 from right to left (starting at Statement S5’s upper limit and continuing to Statement S18’s lower limit), the absence of overlap between these intervals is apparent. Thus, Statement S5’s mean rating is qualitatively different from the mean ratings of the seven items that precede it (S18, S19, S4, S17, S6, S3, S11). These seven statements, then, constitute Group 1. Statement S5 is then the first item in Group 2, which consists of 13 statements (S5, S16, S2, A3, A5, A6, S1, S14, A1, S20, A7, A2, A8), because the last item whose upper limit overlaps with Statement S5’s lower limit is Statement A8. By this logic, Group 3 is
comprised of seven statements—S10, S13, S15, A4, S8, S12, S7—and Statement S9 alone makes up Group 4.

Figure 4.5
*Scattergram of 28 Statements’ Mean Ratings and 95% Confidence Intervals, by Domain and Grouping*

While items can simply be grouped as illustrated in Figure 4.5, discussing these groups in terms of importance level is more meaningful in this investigation’s context. Specifically, based on individual grouping’s averaged mean rating, we might consider each group as corresponding to one of the following four importance categories: very important, important, moderately important, or minimally important. Note that this spectrum of importance represents the middle four of six levels of importance on the scale that panelists used for rating during the study. The anchors on the extreme ends of
the scale are not included because none of the mean ratings tabulated were exactly equal to 6 or 1 (i.e., highly important or least important, respectively).

With this schema in mind, the seven items in Group 1 (S3, S4, S6, S11, S17–19) were considered very important to describing evaluative thinking. As shown in Table 4.6, the averaged mean rating for these statements was 5.15, with values ranging from 4.93 to 5.57. The standard error of mean ratings ranged from 0.16 to 0.25. Careful examination of these statements’ contents indicates that their emphases are balanced between the reasoning process ($n = 3$; S3, S4, S6) and practice-related behaviors ($n = 3$; S17, S18, S19), while issues related to personal conceptions of evaluation’s purpose ($n = 1$; S11) were less prominent.

Table 4.6

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>$\bar{x}$</th>
<th>$SE_{\bar{x}}$</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>I consider the importance of various kinds of data sources when designing an evaluation.</td>
<td>4.93</td>
<td>0.16</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S4</td>
<td>I consider alternative explanations for claims.</td>
<td>5.25</td>
<td>0.17</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S6</td>
<td>I consider the credibility of different kinds of evidence in context.</td>
<td>4.96</td>
<td>0.22</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S18</td>
<td>I offer evidence for claims that I make.</td>
<td>5.57</td>
<td>0.17</td>
<td>Practice</td>
</tr>
<tr>
<td>S19</td>
<td>I make decisions after carefully examining systematically collected data.</td>
<td>5.36</td>
<td>0.17</td>
<td>Practice</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
<td>5.04</td>
<td>0.25</td>
<td>Practice</td>
</tr>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
<td>4.93</td>
<td>0.17</td>
<td>Valuing</td>
</tr>
</tbody>
</table>

Averaged Mean and Standard Error 5.15 0.19
Group 2 consists of 13 statements that were considered important to evaluative thinking. The averaged mean rating for Group 2 was 4.40, with a range from 3.85 to 4.79. The standard error of mean ratings in this group ranged from 0.21 to 0.35. Additionally, as summarized in Table 4.7, careful examination of these statements’ contents indicates that they equally emphasized practice-related behaviors \( (n = 4; S16, S20, A3, A5) \) and reasoning processes \( (n = 4; S2, S5, A6, A8) \). Personal conceptions about the nature of valuing and evaluation were de-emphasized \( (n = 2; A1, A2) \). Further, three items \( (S1, S14, A7) \) addressed multiple domains, with reasoning most frequently implied in all items, practice in two items \( (S1, S14) \), and valuing in only one \( (A7) \).

The “moderately important” category, Group 3, consisted of seven statements: \( S7, S8, S10, S12, S13, S15, \) and \( A4 \). The averaged mean rating for these statements was 3.26, with a range from 2.50 to 3.56. The standard error of mean ratings for this group ranged from 0.22 to 0.32. As indicated in Table 4.8, examination of these statements’ contents suggests that personal conceptions about evaluation’s purpose was the primary theme in this category \( (n = 5; S7, S8, S10, S12, A4) \), whereas practice \( (n = 1; S15) \) and reasoning \( (n = 0) \) were emphasized less. Also, one item \( (S13) \) addressed both the practice and valuing domains.
Table 4.7
Summary Statistics for 13 Statements Classified as “Important” for Evaluative Thinking, by Domain

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>$\bar{x}$</th>
<th>$SE_{\bar{x}}$</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>I consider the availability of resources when setting out to conduct an evaluation.</td>
<td>4.68</td>
<td>0.21</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S5</td>
<td>I consider inconsistencies and contradictions in explanations.</td>
<td>4.79</td>
<td>0.21</td>
<td>Reasoning</td>
</tr>
<tr>
<td>A6</td>
<td>I consider stakeholders’ explicit and implicit reasons for commissioning the evaluation.</td>
<td>4.57</td>
<td>0.24</td>
<td>Reasoning</td>
</tr>
<tr>
<td>A8</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
<td>3.85</td>
<td>0.33</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
<td>4.74</td>
<td>0.24</td>
<td>Practice</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
<td>4.11</td>
<td>0.23</td>
<td>Practice</td>
</tr>
<tr>
<td>A3</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
<td>4.68</td>
<td>0.26</td>
<td>Practice</td>
</tr>
<tr>
<td>A5</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
<td>4.64</td>
<td>0.25</td>
<td>Practice</td>
</tr>
<tr>
<td>A1</td>
<td>Not everything can or should be professionally evaluated.</td>
<td>4.30</td>
<td>0.27</td>
<td>Valuing</td>
</tr>
<tr>
<td>A2</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
<td>3.89</td>
<td>0.33</td>
<td>Valuing</td>
</tr>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
<td>4.52</td>
<td>0.31</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
<td>4.33</td>
<td>0.35</td>
<td>Practice</td>
</tr>
<tr>
<td>A7</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
<td>4.07</td>
<td>0.34</td>
<td>Valuing</td>
</tr>
</tbody>
</table>

Averaged Mean and Standard Error 4.40 0.27
Finally, Group 4, the “minimally important” group, consisted of one statement, S9 (\(\bar{x} = 2.26, SE_{\bar{x}} = 0.24\)), which emphasized personal conceptions of evaluation’s purpose.

Table 4.8  
Summary Statistics for Seven Statements Classified as “Moderately Important” for Evaluative Thinking, by Domain

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>(\bar{x})</th>
<th>SE_{\bar{x}}</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
<td>3.48</td>
<td>0.30</td>
<td>Practice</td>
</tr>
<tr>
<td>S7</td>
<td>I conduct evaluation with an eye towards challenging personal beliefs and opinions.</td>
<td>2.50</td>
<td>0.22</td>
<td>Valuing</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>3.33</td>
<td>0.31</td>
<td>Valuing</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
<td>3.56</td>
<td>0.32</td>
<td>Valuing</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
<td>3.04</td>
<td>0.31</td>
<td>Valuing</td>
</tr>
<tr>
<td>A4</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
<td>3.37</td>
<td>0.28</td>
<td>Valuing</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
<td>3.56</td>
<td>0.32</td>
<td>Practice Valuing</td>
</tr>
</tbody>
</table>

Averaged Mean and Standard Error 3.26 0.29

In general, Groups 1 and 2 tended to be more thematically diverse than Groups 3 and 4. Specifically, Groups 1 and 2 emphasized issues related to cognitive processes and practice relatively equally. Items that addressed multiple domains appeared in Group 2, but did not appear in Group 1. Groups 3 and 4 focused primarily on issues of valuing.

As a whole, these results suggest that participants viewed evaluative thinking as primarily having to do with reasoning and cognitive processes. Behaviors traditionally
recognized and defined as evaluation practice, on the other hand, were seen as secondary to evaluative thinking. Additionally, while personal conceptions about the nature and purpose of evaluation may guide one’s thinking and practice, they were seen as peripheral to the essence of evaluative thinking itself.

**Nature of Consensus and Dissensus**

The discussion of study findings presented earlier in this chapter indicates that, cumulatively, panelists clearly reached agreement about the importance level of 28 statements. However, the nature of consensus and dissensus concerning these items—that is, the relative ease with which participants reached agreement about the importance of particular types of statements—has not yet been examined. This issue is explored in further detail here.

First, it should be noted that panelists reached consensus about the statements that they considered to be of higher importance early in the Delphi study. This finding is best illustrated in Figure 4.6, which shows that in Rounds 1 and 2, 13 of the statements for which consensus was reached were identified as important, while one item was identified as comparatively less important. Further, agreement about the importance level of these 14 statements was reached for eight items in Round 1 ($v_g < 1.73$) and for the remaining six statements in Round 2 ($v_g < 2.26$).
In particular, statements that dealt with reasoning (S2–S6) were quickly identified as important to evaluative thinking in Round 1. Round 1 results additionally suggest that practice-related statements (S18 and S19, respectively) were also of great importance. In contrast, statements dealing with one’s personal views about the purposes of evaluation (S7) were identified as relatively less important than items that addressed reasoning and practice. These observations are summarized in Table 4.9.
Table 4.9
Classification of Eight Descriptive Statements for Which Consensus was Reached in Round 1, by Domain

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>I consider the availability of resources when setting out to conduct an evaluation.</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S3</td>
<td>I consider the importance of various kinds of data sources when designing an evaluation.</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S4</td>
<td>I consider alternative explanations for claims.</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S5</td>
<td>I consider inconsistencies and contradictions in explanations.</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S6</td>
<td>I consider the credibility of different kinds of evidence in context.</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S18</td>
<td>I offer evidence for claims that I make.</td>
<td>Practice</td>
</tr>
<tr>
<td>S19</td>
<td>I make decisions after carefully examining systematically collected data.</td>
<td>Practice</td>
</tr>
<tr>
<td>S7</td>
<td>I conduct evaluation with an eye towards challenging personal beliefs and opinions.</td>
<td>Valuing</td>
</tr>
</tbody>
</table>

While agreement about items that addressed reasoning dominated Round 1, consensus regarding practice-related statements was the primary focus of Round 2. Specifically, as summarized in Table 4.10, items that dealt with the technical aspects of practice (S17, S20, A3, A5) were considered rather important. Consideration for others’ motivations to engage in evaluation (A6) and attention to one’s own conception of evaluation’s purpose (S11) also ranked high in Round 2. Interestingly, respondents did not identify any statements as unimportant to the notion of evaluative thinking in Round 2.
Table 4.10
Classification of Six Descriptive Statements for Which Consensus was Reached in Round 2, by Domain

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6</td>
<td>I consider stakeholders’ explicit and implicit reasons for commissioning the evaluation.</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
<td>Practice</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
<td>Practice</td>
</tr>
<tr>
<td>A3</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
<td>Practice</td>
</tr>
<tr>
<td>A5</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
<td>Practice</td>
</tr>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
<td>Valuing</td>
</tr>
</tbody>
</table>

As the Delphi progressed, the rate of consensus regarding importance level plateaued (see Figure 4.6). This was evidenced by the small number of statements ($n = 4$) for which agreement was reached in Round 3 (see also Table 4.11). Unlike in the previous two rounds of survey administration, most of these statements emphasized valuing ($n = 3$; S9, A1, A4), while the practice domain was emphasized less ($n = 1$; S16) and the reasoning domain was not addressed at all ($n = 0$). Of the statements for which consensus was reached in Round 3, Statement S16 was considered relatively more important than the remaining valuing items for the purposes of describing evaluative thinking, and Statement S9 was considered less important.

Also noteworthy was the increase in variability of responses as the study progressed. This is most clearly summarized and depicted by the increase in widths of confidence intervals that were tabulated towards the end of the study (see Figure 4.6).
Note, in particular, that the standard error used to calculate confidence intervals increased from 0.16 in Round 1 to 0.35 in Round 3.

Table 4.11

Classification of Four Descriptive Statements for Which Consensus was Reached in Round 3, by Domain

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
<td>Practice</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
<td>Valuing</td>
</tr>
<tr>
<td>A1</td>
<td>Not everything can or should be professionally evaluated.</td>
<td>Valuing</td>
</tr>
<tr>
<td>A4</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
<td>Valuing</td>
</tr>
</tbody>
</table>

More careful examination of the items for which consensus had not been reached suggests that most of the disagreement could be attributed either to ambiguous phrasing in some statements or an emphasis on values and valuing in other statements. As summarized in Table 4.12, six of the 10 statements for which disagreement remained at the end of the Delphi dealt with either personal conceptions of the nature of evaluation \((n = 2; S15, A8)\) or evaluation’s purpose \((n = 4; S8, S10, S12, A2)\). In contrast, ambiguous wording in the remaining four statements \((S1, S13, S14, A7)\) could have resulted in multiple interpretations and, thus, divergence in opinion. Put another way, these arguably ambiguous statements could be captured in multiple domains. Consider Statement S14, for example, where the broad notion of “concepts and goals” could refer to the program, staff, or both. The process of teasing these issues apart might be considered part of one’s reasoning or of one’s practice. The ways in which individual panelists interpreted this item would have likely been influenced by their unique
practice and theoretical contexts. Similar interpretive issues are likely common to the remainder of items that appear in Table 4.12.

Table 4.12

Classification of 10 Descriptive Statements Rated for Which Dissensus Remained in Round 3, by Domain

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>A8</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
<td>Practice</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>Valuing</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
<td>Valuing</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
<td>Valuing</td>
</tr>
<tr>
<td>A2</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
<td>Valuing</td>
</tr>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
<td>Reasoning Practice</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
<td>Practice Valuing</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
<td>Practice Reasoning</td>
</tr>
<tr>
<td>A7</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
<td>Valuing Reasoning</td>
</tr>
</tbody>
</table>

Taken together, these results suggest that the process of identifying statements important to the notion of evaluative thinking was a relatively easy task for respondents. Important statements, for the most part, pertained to reasoning and practice. In contrast, identifying statements that were not central to evaluative thinking as a construct seemed to have posed an unexpectedly challenging task. Most of the disagreement appeared to have been rooted in respondents’ conceptions of evaluation’s
purpose (in other words, issues related to values and valuing). Statements organized by domain, irrespective of importance level and the round in which consensus was reached can be found in Table 3.3 (pp. 70–71).

**Summary of Results**

The following guiding research questions are revisited in this section of the chapter to provide a summarized discussion of the findings already presented.

1. What do evaluation experts consider important to evaluative thinking?
2. What is the nature of consensus and dissensus among experts?

**What do evaluation experts consider important to evaluative thinking?**

Study participants rated 28 descriptive statements throughout the Delphi. Of these 28 items, seven were considered “very important” and 13 “important.” On the whole, these 20 statements emphasized reasoning and practice over values/valuing. More precisely, equal numbers of items in the reasoning and practice domains appeared in the “very important” and “important” categories while items pertaining to values/valuing were underrepresented. This is an interesting pattern on its own, but a more careful examination reveals important nuances in participants’ thinking about the nature of evaluative thinking.

Let us first consider items that fell into the “very important” category (see Table 4.6). Thematically, these statements highlight the role that *data and evidence* play in the construction and consideration of evaluative claims. They focus not only on reasoning, but also on how one uses data to produce evidence and build arguments. As
such, given the degree of importance assigned to this category of statements, evaluative thinking seems primarily linked to one’s use of data and evidence in argumentation.

A slightly different pattern is found in statements that fell into the “important” category (see Table 4.7). Rather than emphasizing data and evidence, items in this category highlight the importance of context in solving evaluative problems. Participants acknowledged the importance of, among other things, stakeholders’ motivation to engage in evaluation and the need to adapt to ever-changing situations during the course of an evaluation. Thus, evaluative thinking seems to be secondarily about reasoning and practice in the face of contextual constraints.

The importance of using data when constructing evidence and crafting arguments together with the influence of context on these endeavors speaks to the intimate connection between reasoning and practice. Perhaps more importantly, this finding underlines the fine balance that must be struck between these two aspects of evaluation work to yield a fruitful experience and an effective product—whether a policy brief, a report for internal use, an organization that has the capacity to engage in evaluation, or otherwise.

What is the nature of consensus and dissensus among experts?

On the whole, there was more consensus than dissensus throughout the study. Early on, participants agreed on the importance of items pertaining to reasoning processes and the production of and engagement with evidence. Consensus also occurred for items dealing with issues that uphold democratic ideals—in other words,
that require some level of objectivity as well as some balance between professional judgment and personal conviction. For example, items pertaining to transparency and the need to be context-sensitive were considered of high importance to evaluative thinking.

Additionally, with the exception of a few items, the experts were unable to agree on the importance of statements that address personal values (what individuals deem important) and valuing issues (how importance is determined). There was less consensus regarding statements that addressed personal beliefs and special interests, in particular, and these were also considered less central to the notion of evaluative thinking. For example, there were differences in opinion regarding the importance of how standards for judging value are determined and who should set these standards; the methods by which value judgments could be reached; whether judgments should be rendered using delayed, precise findings or timely, slightly less accurate findings (i.e., use versus accuracy); and the purposes of evaluation (e.g., to inform public debate, to address social injustice, etc.).

Importantly, the participants’ disagreement was difficult to reconcile, suggesting that the nature, demands, and constraints of the evaluation (i.e., the context of the evaluation) dictate how panelists think about and respond to these issues in practice. Likewise, the panelists’ individual contexts (e.g., the settings in which their practice is carried out, their theoretical orientations, etc.) have an effect as well. Thus, it seems
logical that the group’s inherent heterogeneity would lend itself to the sort of differences highlighted here.
CHAPTER 5
CONCLUSION

This study was designed to shed light on experts’ thinking about ideas and elements that are central to the notion of evaluative thinking. It sought an improved understanding of the areas of agreement and disagreement among evaluation experts on this often used but vaguely defined term. The following research questions guided the investigation.

1. What do evaluation experts consider important to evaluative thinking?
2. What is the nature of consensus and dissensus among experts?

This concluding chapter returns to these two questions to offer a broader examination of the research findings. In the course of this examination, I place the findings in the context of the extant literature on the topic, and suggest an empirically-derived working definition of evaluative thinking. I also discuss the significance and implications of the study’s findings, the limitations of the study, and possible directions for future research.

**Evaluative Thinking: A Working Definition**

Fournier (1995b) noted that “[evaluation] logic serves to distinguish evaluation from nonevaluation” (p. 30); so, too, does evaluative thinking. The divergent and sometimes fractured discussions about evaluation and its aims and purposes have resulted in a splintered understanding—and often misunderstandings—as to what
exactly evaluation is and what evaluators do. This has earned evaluators and the evaluation community a negative reputation among those outside of the field (Donaldson, 2001) and led to issues such as evaluation anxiety among stakeholders, which is a deeply-rooted fear of evaluation and the unpredictable consequences that can come from it (Donaldson, Gooler, & Scriven, 2002). Therefore, it is imperative to the growth and development of the field that we arrive at a working definition of evaluative thinking.

It is inarguable that most evaluators are expected to make value claims. That is not the full scope of work that they engage in, however, and the process of reaching such claims does not take place in a vacuum. A strict and narrowed focus on the rendering of judgment is a reductive view of this co-constructed social practice. Rather, it is more productive to recognize that evaluators create knowledge in the process of determining an entity’s merit or significance through the ways in which they address context. As such, the evaluative act and the thinking that accompanies it can—and should—be extended to include considerations for other dimensions that provide a more nuanced understanding of the evaluand and enable one to make evaluative claims about it. In this sense, evaluative thinking does not simply happen in the mind; rather, this cognitive process is manifested as a problem-solving practice—the doing of evaluation.

The ways in which evaluation problems are viewed, thought about, and subsequently resolved occur in a social context and have real consequences. By showing their ability to deal with and account for the contextual constraints and determinants
that inform the value claim being made, evaluators explicitly demonstrate this aspect of evaluative thinking; this is how evaluators *translate* their thinking into practice. Logic and reasoning jointly provide a unifying lens through which the evaluation enterprise can be considered; they anchor the field’s sense of professional identity in the goal of solving social problems, on the one hand, and in an educative enterprise, on the other hand.

Taken together, these ideas suggest a working definition of evaluative thinking as follows:

*Evaluative thinking is a particular kind of critical thinking and problem-solving approach that is germane to the evaluation field. It is the process by which one marshals evaluative data and evidence to construct arguments that allow one to arrive at contextualized value judgments in a transparent fashion.*

Understood this way, the notion of evaluative thinking allows us to summarize and capture what is at the heart of an evaluator’s practice and at the core of the field’s theories. It takes the field’s conceptualization of “evaluation” beyond determinations of merit and worth, beyond the idea that “Bad is bad and good is good and it is the job of evaluators to decide which is which” (Scriven, 1986, p. 19). Framing the concept around problem-solving and educational identities provides an accessible, useful frame of reference.

Dahler-Larsen (2012) noted that, “Not all terms will be defined the same way at all times because they are embedded in a changing social context, informed by different
philosophies, paradigms, and points of views” (p. 5). To be sure, evaluative thinking is a dynamic construct. Other studies will surely confirm that it is multi-layered and multidimensional (e.g., Buckley and Archibald, 2013). This observation begs the question, then, of why we should bother going through the trouble of defining a concept as abstract and elusive as evaluative thinking when we can expect its meaning to eventually change. The answer is quite simple: If we do not define it, there will be nothing to revise when the time comes to do so. And we will not be able to trace its evolution, effectively omitting an important piece of the field’s development. This study reflects an effort to provide the evaluation community with a starting point to refine what we mean when we say “evaluative thinking.” It is therefore reasonable to anticipate that the working definition of evaluative thinking that has been offered here should be revised and refined as the field continues to develop a richer and more precise understanding of this construct.

Implications

The findings of this study, in conjunction with the existing literature, have a number of implications for evaluator training, evaluation practice, and research on evaluation. Each of these categories is addressed in turn below.

Evaluator Training

The recognition that reasoning underlies practice sheds light on how the field can continue to grow. The literatures on the teaching of evaluation and evaluator competencies have, up to this point, focused primarily on ensuring that evaluators gain
technical, “hard” skills. While there is some effort to broaden the field’s focus (Thomas & Madison, 2010), explicit attention needs to be placed on development and attainment of the even “harder” skills that are at the field’s core. The ability to reason is among these skills because it involves thinking in a dynamic, fluid, and integrative fashion. Teaching thinking skills is no easy task and remains a call to be answered.

An intentional focus on thinking and reasoning during the course of evaluators’ professional development entails bringing the social and technical aspects of evaluation more in line with each other, thus moving away from the idea that evaluation is strictly a mechanical activity and that evaluators are simply technicians (Schwandt, 2008b). Endeavoring to teach evaluators how to reason evaluatively means teaching them how to think critically as they conduct evaluations. Framing the relationship between evaluation and evaluative reasoning in this way also means that these concepts should be taught in a manner that enables the next generation of evaluators to draw on their portfolio of experiences and apply their skills across a range of evaluative domains. That is, training programs should equip future evaluators with the ability to transfer knowledge and skills from one evaluation context to another, but existing programs may not be designed to meet this goal.

The landscape of evaluator training at the graduate level includes formal knowledge acquisition (e.g., as by lecture). Further, the teaching of evaluation literature, currently emphasizes five other modes by which evaluators gain theoretical, technical, and practical knowledge: simulations, role-play, discussion groups, single course
projects, and practicum experience (Trevisan, 2004). Depending on the course content, focus of course projects, and scope and duration of the practicum, these different learning modalities are likely designed with the intention of helping students build and hone technical skills. At present, argumentation and related topics such as reasoning and logic are not explicitly addressed, at least to a degree that warrants their mention in the literature. It seems, however, that activities emphasizing these skill sets would dovetail seamlessly into existing course activities. For example, the use of a debate format for discussion would allow students to practice constructing and delivering arguments. The explicit use and integration of the Socratic method would engage students in discussion. These and other similar activities could easily be adapted to suit different audiences, particularly program staff and other adult learners, and then be presented through in-person workshops or webinars.

**Evaluation Practice**

The Canadian Evaluation Society (CES) has led the charge to professionalize evaluation by implementing a Professional Designations Program that allows Canadian evaluators to apply for an evaluation credential that proves their ability to practice evaluation by the order’s highest standard (Canadian Evaluation Society, 2010). Currently, the CES requires interested applicants to submit proof of educational and experiential qualifications for review. Applicants must also demonstrate competence in interpersonal, management, reflective, situational, and technical practices. This is
indicative of the CES’s efforts to go beyond technical skills and of its concern for the “harder” skills that are difficult to attain, including evaluative thinking.

As the evaluation field continues moving towards professionalization at a broader level, the current findings will be of use to domestic organizations and other international evaluation associations (e.g., African Evaluation Association, American Evaluation Association, etc.) that need guidance in determining content and requirements for licensure beyond those outlined by the CES. A working definition of evaluative thinking can simultaneously underscore the complexity of measuring evaluative expertise while giving evaluators in a national and global context a common understanding of what matters in evaluation work.

**Research on Evaluation**

The implications of this study’s findings for the research on evaluation enterprise are many. Perhaps the most important is the need to shift towards both intra- and interdisciplinary work. To date, a fair amount of evaluation scholarship has emerged from within the field of evaluation; this is a natural part of a field’s development and should be expected. But if we take, for example, the body of literature on prescriptive evaluation models (i.e., theories of what practice should be), we find a leading example of the kind of inbred work currently in question. That is, evaluation theorists have developed evaluation theories based on the characteristics that evaluation shares with research and these theories have been refined, challenged, and studied primarily by those within the discipline. However, the current study suggests that we have perhaps been too narrowly
focused and should consider how a range of disciplinary points of view—both within and outside of evaluation—might inform our theories in particular and our understandings of evaluation more generally. Up to this point, only a paucity of such efforts exist (Mark, Donaldson, & Campbell, 2011). Moving evaluation in the direction proposed here will certainly enrich and enhance the integrity of our scholarship.

Increased collaboration between evaluators and scholars from different substantive areas within evaluation (e.g., policy and program evaluation) could be a first step towards progress in this area. This would allow for the fruitful exchange of ideas concerning how to approach evaluation and how to address important contextual issues. Likewise, collaboration among evaluators who evaluate similar entities but with varying disciplinary training would also improve understandings of different areas within the field that can be empirically studied. Eventually, collaboration between evaluators and members of other fields where evaluation is influential and also influenced (e.g., cognitive science, political science, sociology) will determine the future landscape of research on evaluation, contributing to an enriched sense of what evaluation is from multiple perspectives.

**Study Limitations**

As with any research project, there are certain limitations that must be acknowledged and addressed. The methodological, analytic, and logistical obstacles encountered during the study are described in this section.
Methodological Limitations

The Delphi method has important advantages but it also carries certain limitations. Perhaps the most significant critique of the approach relates to the inability of participants to converse in real-time. Some of the depth and richness of experts’ thinking may be sacrificed through iterative and sequential administration of multiple surveys because they cannot actively exchange ideas—something that would be feasible through focus groups or round table discussions. Other approaches, such as the nominal group technique (NGT) described by Dalbecq and colleagues (1975), resemble the Delphi technique but have important distinctions. For example, the NGT is nearly identical to the Delphi technique, except it brings participants together physically. As a result, however, this approach loses one advantage of the Delphi in that it cannot guarantee anonymity and avoid “bandwagon thinking.” That is, this method may limit the degree to which experts feel comfortable being candid in their responses, because they know who has offered which ideas, and they may be swayed by their colleagues’ reputations or persuasiveness. And, from a logistical standpoint, experts tend to be geographically dispersed, making multiple in-person meetings challenging if not impossible. As such, a method such as the NGT was not suitable for this study.

The methodological advantages and disadvantages of the Delphi technique are certainly worthy of further discussion and empirical study. Nonetheless, despite the known limitations of this method of inquiry, given the research questions and study goals, it was still preferred over other methods of obtaining expert opinion.
**Analytic Limitations**

Analytic issues are also of concern in this investigation. Specifically, challenges arose in the process of disentangling an item’s importance level from the extent to which consensus had been reached. The analytic tension in the consideration of importance level and consensus is represented in Figures 4.5 (p. 90) and 4.6 (p. 96), respectively. As described in chapter 4, each of these two figures provides a unique point upon which one can focus: the data in Figure 4.5 (p. 90) emphasize the notion of importance, leading to intermingling of “high variance” items (i.e., items with very wide confidence intervals and, thus, divergent views regarding their importance) with “low variance” items (i.e., items with narrower confidence intervals and higher levels of consensus regarding their importance). In contrast, the data presented in Figure 4.6 (p. 96) privilege consensus and dissensus to highlight areas of disagreement among participants. In short, the processes of determining consensus and importance level seem to be in competition with each other, making it challenging to discuss the relationship between the two. This study should serve as a starting point for further investigation of dissensus within a group, as well as for methodological inquiry into how this problem of competing foci can be best addressed.

**Logistical Limitations**

Finally, while the study did not suffer from a poor response rate, use of multiple surveys and multiple efforts to follow up with participants presented a challenging and time consuming responsibility for the investigator as well as participants. For example,
the data collection period for this study took place over the span of 10 months, most of which overlapped with spring, summer, and winter holidays as well as professional conference seasons. As such, the initial plan of providing a two-week window for participants to respond to and provide feedback on each survey was often extended to four (sometimes six) weeks to accommodate various schedules. For some participants, flexibility in the schedule was appreciated, but for others, it contributed to difficulties recalling why they had provided their initial responses. This was an unanticipated source of frustration for at least some of the participants as they were effectively asked to reconstruct their reasoning after a significant period of time had elapsed. Further, while some of the difficulties typically associated with data collection were mitigated by the use of e-mail to administer surveys and obtain participants’ feedback, the cognitively demanding nature of the task surely added to the perception that participation was time consuming and wearisome; these sentiments were always understandable but were also occasionally challenging to manage.

**Directions for Future Research**

The present investigation led to interesting findings concerning experts’ views regarding evaluative thinking. Perhaps more importantly, it also highlights a few descriptive, empirical, and methodological areas in which additional research could further expand the knowledge base. Future potential studies are outlined below.
Descriptive & Empirical Studies

More research is needed to understand evaluative thinking in action. This can be accomplished in a number of ways. One possibility involves linking participants’ responses in this study to a number of variables that might explain differences in the item ratings. Results of such a study would contribute to an understanding of the various facets of participants’ mental models that drive—either covertly or overtly—the responses that they provided. Another possibility involves examining the ways in which participants’ responses differ by theoretical orientation as described by Alkin and Christie (2004) and Christie and Alkin (2013). Such analyses will shed light on how conceptualizations of evaluative thinking vary depending on theoretical emphasis.

Alternatively, a representative evaluation task can be assigned to expert and novice evaluators whose practices differ by the sector in which they occur (e.g., education, public health, human services, etc.) and a protocol analysis (Ericcson & Lehmann, 1996; Ericcson & Simon, 1993) could be used to record and analyze their cognitive process. Doing so would yield deeper insight into how decisions are made across different contexts and how evaluation problems are solved by various sub-groups.

Collectively, studies such as these would not only further the field’s understanding of the nature of evaluative thinking and how it occurs, but would also help us to begin to understand how we can use that information to improve practice and impact social change.
Methodological Studies

The Delphi technique was originally developed “to make effective use of informed intuitive judgment” to solve particular policy-related problems by “forecasting the consequences of alternative policies” (Helmer, 1967a, pp. 3–4). Thus, it was not necessarily created to address the types of social science research questions that guided the current investigation. It is perhaps not surprising then that issues pertaining to the reliability and validity of findings generated using the Delphi technique have been raised in the literature (Ament, 1970; Rowe, Wright, & Bolger, 1991; Sackman, 1975). These concerns are typically epistemologically driven, especially during a period when there has been so much interest in classifying the Delphi technique as either a post-positivist or constructivist research tool. These issues are certainly worthy of further exploration, especially as the technique is applied in broader, more diverse areas of inquiry.

For example, the political process is inherently driven by relationships and values. Applying the Delphi technique—with its focus on consensus building—in the policy-making context could potentially lead to more objective and efficient decision-making. Policy-makers might use it to explore the advantages and disadvantages of various political courses of action in an anonymous fashion. In this way, perspectives of a number of different decision-makers can be cogitated without the pressure of having to consider the source of such opinions. The process would also be data-driven rather than values-driven.
Within the educational context, test-makers might use the Delphi approach to determine the pool of potential questions and the difficulty levels of questions that could be included on a national exam in a given year. Because the perception among many educational professionals is that high-stakes testing is far-removed from the actual teaching that occurs within classroom walls, obtaining and considering the opinions of teachers and master teachers, for example, would be a step towards closing the gap between these two activities. Alternatively, school administrators could potentially use the Delphi method to determine the focus and direction of different strategic plans and initiatives. Due to the approach’s flexibility, this process could be used among a small group of stakeholders (e.g., among school administrators only) or expanded to include a more diverse group of vested individuals (e.g., teachers, parents, students).

Studies that address the Delphi technique itself might examine how consensus criteria should be determined between rounds of survey administration. Moreover, this type of inquiry might address the question of how the stability of ratings ought to be defined, under which circumstances, and for what purposes. Such efforts will contribute to the current literature base dedicated to the Delphi method’s refinement.

**Replication Studies**

Another possibility with respect to future research includes efforts to replicate or refine the present study’s results. For example, the present investigation defines experts as evaluation scholars; thus, participants have been chosen based on their contributions to and visibility in the field, but not necessarily based on how representative they are of
the evaluation field as a whole. Future studies might draw from a broader sample of evaluation scholars and practitioners from various disciplines and backgrounds. Doing so would not only address concerns about potential biases towards inclusion of evaluation scholars with particular perspectives stemming from their training and/or backgrounds, but could also serve as a means of validating and confirming this study’s findings. Further, because the present investigation is the first to explore this topic, the findings can be used as a benchmark for our understanding of evaluative thinking. Future studies will be useful in tracking the evolution of the term over time, as alluded to earlier in this chapter.

**Final Remarks**

The current investigation indicates that evaluative reasoning is a particular way of thinking and problematizing; it is what differentiates the evaluation discipline from other practice-based fields. Evaluative thinking is at the discipline’s center and is a large part of what makes evaluation both an art and a science (Lincoln, 1991). More importantly, its translatable and transferrable attributes are the means by which the fine balance between the social and technical aspects of evaluation is maintained.

This study suggests a number of ways that the evaluation field might continue to bridge the practice-theory gap, chiefly through a deepened understanding of the influence of one’s worldview on practice, how one’s thinking is translated into practice, and vice versa. Evaluative thinking is a developing area within the evaluation field that is worthy of further study. Insights gained from future research on evaluative thinking
will surely contribute in important ways to the field’s efforts to articulate clear
descriptive theories of practice and its identity as a rigorous area of study, enriching the
discipline and leading to improved practice and policy-making.
Appendix A

EXPERT RECRUITMENT LETTER

(INsert DATE HERE)

Dear Dr. (INSERT NAME HERE),

My name is Anne Vo and I am a doctoral candidate in the Graduate School of Education and Information Studies at the University of California, Los Angeles (UCLA). I have been studying program evaluation under Professor Marvin Alkin’s guidance and supervision. I am e-mailing to inquire about the possibility of your participation in my dissertation study, tentatively titled “Toward a Definition of Evaluative Thinking.”

Attached to this message is an information sheet that has been developed based on guidelines set forth by UCLA’s Institutional Review Board. It outlines the study’s scope, purpose, and methods. The information sheet also describes what is to be expected as part of your involvement in the study.

If you are willing to participate in this investigation upon reviewing the information sheet, please respond to this message and I will provide details concerning how we might proceed.

Thank you for your time and consideration. I look forward to hearing from you.

Sincerely,

Anne Vo
Doctoral Candidate
Principal Investigator

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779
Appendix B

UNIVERSITY OF CALIFORNIA LOS ANGELES
STUDY INFORMATION SHEET

Toward a Definition of Evaluative Thinking

Anne T. Vo, M.A. (Principal Investigator) and Marvin C. Alkin (Faculty Sponsor), from the Graduate School of Education & Information Studies at the University of California, Los Angeles (UCLA) are conducting a research study.

You were selected as a possible participant in this study because you are considered an expert on the topic of evaluative thinking based on a review of the literature. Your participation in this research study is voluntary.

**Why is this study being done?**

The evaluation literature frequently discusses an idea referred to as “evaluative thinking.” However, there is little agreement in the evaluation field about the meaning of this term. The purpose of this study is to understand how evaluation experts think about this idea so that a definition can be derived to inform future research and policy decisions.

**What will happen if I take part in this research study?**

If you volunteer to participate in this study, the researcher will ask you to:

- Complete a series of three web-based questionnaires. The questionnaires will ask you to:
  - Read a list of 20 statements describing various aspects of evaluative thinking.
  - Rank the relative importance of each statement.
  - Suggest additional descriptive statements.

- Potentially complete a brief, semi-structured follow-up interview through e-mail after each questionnaire. During the interviews, you will be asked to:
  - Describe the rationale for some of the answers provided on the surveys.

Questionnaires will be completed electronically via e-mail. Follow-up interviews will be completed via e-mail and/or telephone.
How long will I be in the research study?

Participation will take a total of about 15 to 30 minutes per survey and 5 to 10 minutes per interview.

Are there any potential risks or discomforts that I can expect from this study?

There are no anticipated risks or discomforts due to participating in this study.

Are there any potential benefits if I participate?

You will not directly benefit from your participation in the study.

The results of the research may be used to develop an instrument that measures evaluative thinking, to develop a conceptual framework for understanding this idea, and to inform research, program, and policy decisions.

Will information about me and my participation be kept confidential?

Any information that is obtained in connection with this study and that can identify you will remain confidential. It will be disclosed only with your permission or as required by law. Confidentiality will be maintained by means of removing identifying information that can be linked to responses provided through questionnaires and interviews; implementing password protection for storing of electronic files; and restricting file access to the Principal Investigator.

What are my rights if I take part in this study?

- You can choose whether or not you want to be in this study, and you may withdraw your consent and discontinue participation at any time.
- Whatever decision you make, there will be no penalty to you, and no loss of benefits to which you were otherwise entitled.
- You may refuse to answer any questions that you do not want to answer and still remain in the study.

Who can I contact if I have questions about this study?

- The research team: If you have any questions, comments or concerns about the research, you can talk to one of the researchers. Please contact:
Anne T. Vo, M.A.
Principal Investigator
Doctoral Candidate

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779

Marvin C. Alkin, Ed.D.
Faculty Sponsor
Professor Emeritus

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 3026
Los Angeles, CA 90095-1521

E: alkin@gseis.ucla.edu
P: 310.825.4800

- **UCLA Office of the Human Research Protection Program (OHRPP):**
  If you have questions about your rights while taking part in this study, or you have concerns or suggestions and you want to talk to someone other than the researchers about the study, please call the OHRPP at (310) 825-7122 or write to:

  UCLA Office of the Human Research Protection Program
  11000 Kinross Avenue, Suite 211, Box 951694
  Los Angeles, CA 90095-1694
Appendix C

EXPERT INTRODUCTORY LETTER

Dear Dr. (INSERT NAME HERE),

Thank you for agreeing to participate in my dissertation study, tentatively titled “Toward a Definition of Evaluative Thinking.”

My interest in this topic stems from the observation that while the term “evaluative thinking” has been frequently used in the evaluation literature (including Dahler-Larsen, 2012; Patton, 2002; Schwandt, 2008; and Scriven, 1995), there remains little clear indication of what “evaluative thinking” means or how it might be studied systematically.

With that in mind, the purpose of this study is to empirically articulate the various facets of evaluative thinking using the Delphi technique. This method will allow us to better understand areas of consensus and dissensus among a purposive sample of evaluation experts about appropriate indicators of evaluative thinking.

The questionnaire you are being asked to complete is rooted in the existing literature, and seeks to build on the existing evaluation knowledge base to inform a more nuanced understanding of just what is meant by “evaluative thinking.”

As indicated in the study information sheet that you previously received, you are being asked to complete a total of three questionnaires. The screens that appear via the survey link below will take you through the first of these questionnaires. It includes a brief, introductory note; sorting instructions; the 20 descriptive statements that you will sort into categories of importance; the actual sorting activity; and one follow-up question. The descriptive statements are also attached to this message for your review.

Please note that because this is Round 1 of the Delphi, you are being asked to sort the 20 statements into 6 categories of importance; each category may contain up to 5 descriptive statements only. Second, you are asked to suggest additional descriptive statements that would be helpful in capturing the essence of evaluative thinking.

If possible, I would sincerely appreciate receiving your completed survey by Friday, June 8th.

Please feel free to contact me at annevo@ucla.edu if you have questions or concerns.

Thank you for your time, attention, and participation.
Sincerely,

Anne Vo
Doctoral Candidate
Principal Investigator

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779

Works Cited:
Appendix D

DELPHI QUESTIONNAIRE 1

Dear Dr. (INSERT NAME HERE),

Thank you for agreeing to participate in my dissertation study, tentatively titled “Toward a Definition of Evaluative Thinking.”

As previously indicated, you will be asked to sort 20 statements describing various aspects of evaluative thinking into 6 categories of importance on this questionnaire; each category may contain up to 5 descriptive statements only. Second, you will be asked to suggest additional descriptive statements that would be helpful in capturing the essence of evaluative thinking.

As you complete this questionnaire, please note that each of the descriptive statements is intended to capture a different facet of evaluative thinking. Collectively, they are designed to reflect the behaviors, actions, and attitudes that we recognize as the “doing of evaluation.” Thus, these statements should represent the essence of evaluative thinking, which for the purposes of this study is different from evaluation practice itself. Your participation in this research will help to clarify and refine our understanding of this important emerging concept.

Lastly, please note that individual responses from each questionnaire will be kept confidential and will be reported back to the group of participating experts only in aggregate. If possible, I would sincerely appreciate receiving your completed survey by Friday, June 8th.

Please contact me at annevo@ucla.edu if you have questions or concerns.

Sincerely,

Anne Vo
Doctoral Candidate
Principal Investigator

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779
Section 1. Statement Descriptions.

The following 20 statements have been excerpted or adapted from various evaluation scholars’ writings about evaluative thinking; that is, the cognitive processes that subsequently result in behavior that we recognize as the physical doing of evaluation.

Table 1
Statements to be Rated

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
</tr>
<tr>
<td>S2</td>
<td>I consider the availability of resources when setting out to conduct an evaluation.</td>
</tr>
<tr>
<td>S3</td>
<td>I consider the importance of various kinds of data sources when designing an evaluation.</td>
</tr>
<tr>
<td>S4</td>
<td>I consider alternative explanations for claims.</td>
</tr>
<tr>
<td>S5</td>
<td>I consider inconsistencies and contradictions in explanations.</td>
</tr>
<tr>
<td>S6</td>
<td>I consider the credibility of different kinds of evidence in context.</td>
</tr>
<tr>
<td>S7</td>
<td>I conduct evaluation with an eye towards challenging personal beliefs and opinions.</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
</tr>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
</tr>
<tr>
<td>S18</td>
<td>I offer evidence for claims that I make.</td>
</tr>
<tr>
<td>S19</td>
<td>I make decisions after carefully examining systematically collected data.</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
</tr>
</tbody>
</table>
Section 2. Round 1 Rating Form.

Given the information provided above, please place each of the 20 statements that appear on the previous page into one of the following six categories of importance by writing the statement number on a line in the appropriate category.

Please note that each category should have no more than 5 statement numbers. That is, each statement number should appear only once in the table below.

<table>
<thead>
<tr>
<th>Highly Important</th>
<th>Very Important</th>
<th>Important</th>
<th>Moderately Important</th>
<th>Minimally Important</th>
<th>Least Important</th>
</tr>
</thead>
<tbody>
<tr>
<td># 1</td>
<td># 2</td>
<td># 3</td>
<td># 4</td>
<td># 5</td>
<td># 6</td>
</tr>
<tr>
<td># 7</td>
<td># 8</td>
<td># 9</td>
<td># 10</td>
<td># 11</td>
<td># 12</td>
</tr>
<tr>
<td># 13</td>
<td># 14</td>
<td># 15</td>
<td># 16</td>
<td># 17</td>
<td># 18</td>
</tr>
<tr>
<td># 19</td>
<td># 20</td>
<td># 21</td>
<td># 22</td>
<td># 23</td>
<td># 24</td>
</tr>
</tbody>
</table>

Section 3. Suggestions for Additional Statements.

What additional statements (or items) might you include in an effort to describe “evaluative thinking,” if any?

1. 
2. 
3. 
4. 
5. 

Thank you for your time and participation.
Appendix E

EXPERT REMINDER LETTER

Dear Dr. (INSERT NAME HERE),

Thank you for agreeing to participate in my dissertation study.

If you have not already submitted your responses, I would greatly appreciate receiving your questionnaire by **Friday, June 8th**.

Please note that because this is Round 1 of the Delphi, you are being asked to sort the 20 statements into 6 categories of importance; **each category may contain up to 5 descriptive statements only**. Second, you are asked to suggest additional descriptive statements that would be helpful in capturing the essence of evaluative thinking.

Please feel free to contact me at annevo@ucla.edu if you have questions or concerns.

Thank you for your time, attention, and participation.

Sincerely,

Anne Vo
Doctoral Candidate
Principal Investigator

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779
Appendix F

FOLLOW-UP INTERVIEW PROTOCOL

Dear Dr. (INSERT NAME HERE),

Thank you for taking the time to complete the (INSERT SURVEY NUMBER HERE) Delphi questionnaire for my dissertation study on evaluative thinking. I have analyzed the survey data and will share results from Round (INSERT SURVEY NUMBER HERE) with the expert panel shortly.

I am currently preparing to launch the next questionnaire. To do so, I need to obtain feedback from participants about the statements for which consensus has not been reached. Specifically, I am contacting experts who provided ratings at the ends of the importance scale for statements that will be re-rated in Round (INSERT SURVEY NUMBER HERE).

Upon examining results from Round (INSERT SURVEY NUMBER HERE), it appears that:

- You rated Statement #(INSERT STATEMENT NUMBER HERE) (INSERT STATEMENT DESCRIPTION HERE) as "1=least important."
- You rated Statement #(INSERT STATEMENT NUMBER HERE) (INSERT STATEMENT DESCRIPTION HERE) as "6=highly important."

Would you please provide a rationale for your ratings so that fellow panelists may use your feedback to revise their responses if they so choose? Further, in keeping with the design and purpose of the Delphi, I will share your response with the panel, but your identity will remain confidential.

Thank you for your time and I look forward to hearing from you.

Sincerely,

Anne Vo
Doctoral Candidate
Principal Investigator

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779
Appendix G

DELPHI QUESTIONNAIRE 2

Dear Dr. (INSERT NAME HERE),

Thank you for your participation in my dissertation study about evaluative thinking. You have completed Round 1 of the investigation and the information contained in this packet will be used for the next phase of the study. Round 2 requires panelists to review results from the previous phase and to provide feedback for the last round of the study. To facilitate this task, I have organized all of the necessary information in this packet into the following sections:

- Section 1: List of Participating Panelists
- Section 2: Summary of Round 1 Results
- Section 3: Panelists’ Feedback Based on Round 1 Data
  - A: Overview of Statements to be Rated in Round 2
  - B: Panelists’ Feedback for Statements to be Rated in Round 2
- Section 4: Round 2 Rating Form

Each section contains an explanation of the information contained therein. For panelists, the goal of this phase is to rate the statements that appear in Section 3 after considering fellow experts’ feedback. Please return the item ratings by **Monday, October 8th**.

If any questions or concerns arise as you review materials provided in this packet, please feel free to contact me at annevo@ucla.edu.

Thank you for your time and participation.

Sincerely,

Anne T. Vo, M.A.
Principal Investigator
Doctoral Candidate

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779
Section 1. List of Participating Panelists.

This investigation is not possible without the valuable time and effort of the following 28 evaluation experts who have agreed to partake in the study. All panelists have completed Round 1 of the Delphi study and are included in Round 2 of the investigation.

Robert Boruch             Rodney Hopson             Hallie Preskill
Eleanor Chelimsky        Ernest House            Sharon Rallis
J. Bradley Cousins       George Julnes            Debra Rog
Lois-ellin Datta         Jean King               Thomas Schwandt
Stewart Donaldson        Linda Mabry              William Shadish
Jody Fitzpatrick         Melvin Mark              Laurie Stevahn
Deborah Fournier         Donna Mertens            Carol Weiss
Jennifer Greene          Robin Miller             Joseph Wholey
Gary Henry               Jonathan Morell
Stafford Hood            Michael Patton

Section 2. Summary of Round 1 Results.

This section of the informational packet provides a summary of study results based on analysis of qualitative and quantitative data from Round 1.

Survey Results - Quantitative. Experts reviewed and rated 20 statements on a 6-point scale in terms of their relative importance (1=least important; 6=highly important) when considering how to characterize evaluative thinking. Results indicate the averaged mean rating and averaged variance of all items in Round 1 were 4.18 and 1.73, respectively. Mean ratings of individual statements ranged from 2.46 to 5.57, while their variances ranged from 0.74 to 3.12. Examination of 20 statements’ mean ratings and variances relative to the averaged mean ($\mu = 4.18$) and averaged variance ($\sigma^2 = 1.73$) led to the identification of eight statements (40%) for which panelists have reached consensus concerning importance level. Table 1, below, specifies the statements on which consensus has been reached from Round 1 as well as each item’s summary statistics. Additionally, Round 1 results suggest that experts’ opinions converged quickly about what was most important when thinking about how to describe evaluative thinking. However, agreeing about what was least important was a challenge for the group.
Table 1
Summary Statistics of Eight Statements on Which Consensus Has Been Reached in Round 1

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>( \bar{x} )</th>
<th>( s^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>I consider the availability of resources when setting out to conduct an evaluation.</td>
<td>4.68</td>
<td>1.26</td>
</tr>
<tr>
<td>S3</td>
<td>I consider the importance of various kinds of data sources when designing an evaluation.</td>
<td>4.93</td>
<td>0.74</td>
</tr>
<tr>
<td>S4</td>
<td>I consider alternative explanations for claims.</td>
<td>5.25</td>
<td>0.79</td>
</tr>
<tr>
<td>S5</td>
<td>I consider inconsistencies and contradictions in explanations.</td>
<td>4.79</td>
<td>1.21</td>
</tr>
<tr>
<td>S6</td>
<td>I consider the credibility of different kinds of evidence in context.</td>
<td>4.96</td>
<td>1.29</td>
</tr>
<tr>
<td>S7</td>
<td>I conduct evaluation with an eye towards challenging personal beliefs and opinions.</td>
<td>2.50</td>
<td>1.37</td>
</tr>
<tr>
<td>S18</td>
<td>I offer evidence for claims that I make.</td>
<td>5.57</td>
<td>0.85</td>
</tr>
<tr>
<td>S19</td>
<td>I make decisions after carefully examining systematically collected data.</td>
<td>5.36</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Survey Results – Qualitative. In addition to rating statements on the questionnaire, experts were asked to provide up to five suggested statements that were meant to address areas and ideas not covered on the original list. Twenty-four of 28 experts responded to this prompt, yielding 78 new statements. Each statement was examined and compared to the others to determine the extent of overlap in the ideas that were expressed. This process led to a number of statements being collapsed and reduced the overall pool of suggested statements to 33. These statements were then individually examined to determine the general ideas that were represented in each and placed into like groupings. Subsequently, statements were randomly selected based on the proportion of items within each category. In an effort to respect experts’ time and ensure that the survey remains at a reasonable length, the maximum number of suggested statements that could have been selected was eight. Thus, eight items were randomly selected from the collection of 33 suggested statements (i.e., Statements A1–A8; see Part 3A below) and combined with the 12 statements that remained from Round 1 (i.e., Statements S7–S17 and S20; see Part 3A below) to form a new list of 20 statements that panelists will rate in Round 2 of the study.
Section 3A. Panelists’ Feedback Based on Round 1 Data – Overview of Statements to be Rated in Round 2.

As described in Section 2, above, a list of statements to be rated in this phase of the investigation was created based on data that were collected and analyzed in Round 1. The newly generated list consists of:

a) 12 statements on which consensus has not been reached from the previous round and
b) eight statements that were randomly selected from a pool of suggestions that panelists provided.

These statements are provided for initial review in Table 2, below.

Table 2
Overview of Statements to be Rated in Round 2

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement Description Remaining Statements From Round 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
</tr>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
</tr>
</tbody>
</table>
Table 2
Overview of Statements to be Rated in Round 2, cont.

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Not everything can or should be professionally evaluated.</td>
</tr>
<tr>
<td>A2</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
</tr>
<tr>
<td>A3</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
</tr>
<tr>
<td>A4</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
</tr>
<tr>
<td>A5</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
</tr>
<tr>
<td>A6</td>
<td>I consider stakeholders’ explicit and implicit reasons for commissioning the evaluation.</td>
</tr>
<tr>
<td>A7</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
</tr>
<tr>
<td>A8</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
</tr>
</tbody>
</table>

Section 3B. Panelists’ Feedback Based on Round 1 Data – Panelists’ Feedback for Statements to be Rated in Round 2.

Table 3, below, expands on information contained in the preceding section such that it identifies the statements to be rated in this round of the investigation. The table also contains summary statistics that indicate the ways in which the 12 statements from Round 1 do not meet consensus criteria. Additionally, it highlights comments from panelists who rated these statements on either of the extreme ends of the importance scale (e.g., 1=least important; 6=highly important). Panelists are asked to review and use the summary statistics along with fellow colleagues’ feedback to inform how they assign statement ratings in Round 2. Ratings will be recorded on the form that appears in Section 4.
Table 3
Panelists’ Feedback for Statements to be Rated in Round 2

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
<td><strong>Mean</strong> 4.50</td>
</tr>
<tr>
<td></td>
<td><strong>Variance</strong> 2.70</td>
</tr>
<tr>
<td></td>
<td><strong>Median</strong> 5.00</td>
</tr>
<tr>
<td></td>
<td><strong>Skewness</strong> -0.78</td>
</tr>
</tbody>
</table>

**Least Important Rationale**
...the evidence that different stakeholders see as being important may vary even though they might agree on the importance of the question. For example, parents, teachers, principals, may agree on the importance of students’ academic achievement but require different types of evidence as being credible for them to conclude whether it has happened or not.

**Highly Important Rationale**
If an evaluation question will not likely be able to be answered...then conducting the evaluation does not seem prudent. Put differently, evaluation questions are at the heart of evaluation work—such questions drive the entire process thereafter—so they must be framed well and in answerable ways or it most probably is pointless to continue....Therefore, I place a high priority on evaluation questions right from the start, and on how they are framed, which can make or break a successful study.
<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>Mean 2.61</td>
</tr>
<tr>
<td></td>
<td>Variance 2.40</td>
</tr>
<tr>
<td></td>
<td>Median 2.50</td>
</tr>
<tr>
<td></td>
<td>Skewness 0.53</td>
</tr>
</tbody>
</table>

**Least Important Rationale**

I do not believe that evaluation can necessarily change the ideology of someone who is entrenched in their particular ideology. For example someone who believes that African Americans, Hispanics, and Native Americans are intellectually inferior to whites and therefore should be expected to [under-achieve academically] and are consequently destined for lower economic status will maintain that belief in the face of evaluative evidence that...suggests otherwise. Therefore...I conduct evaluation with an eye towards fully and accurately understanding the phenomena/evaluand with culture and cultural context being central in this effort when undertaken in culturally diverse settings....It is more likely that the resulting evidence and interpretations will have greater validity [this way].

**Highly Important Rationale**

On rating number 8 as highly important, I don’t mean that we necessarily have to find that the unquestioned ideological views are untrue. They may in fact be true. But, I do think that it is very important for evaluators to make others aware of the need to be more open, unquestioning, and seeking evidence in regard to their beliefs.
Table 3
Panelists’ Feedback for Statements to be Rated in Round 2, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>S9</td>
<td>2.46</td>
</tr>
<tr>
<td></td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
</tr>
</tbody>
</table>

**Least Important Rationale**
The issue for me is I try to conduct evaluation without a hidden agenda. I try to conduct it as openly as I can to look at the issues and to take in all perspectives. It’s not that the evaluation won’t challenge special interests; I just don’t set my designs or methods quite in that way.

**Highly Important Rationale**
I take seriously the charge for evaluation in the public interest and to rigorously test the statements about public policies and programs that are made by folks with an interest in the continuation, expansion or dissolution of the program, whether they are program administrators, evaluation or program sponsors, vendors, or politicians. All have a direct interest in the program’s status and all positions and assumptions should be checked...All evaluations of public policies or programs and those provided by foundations who receive special tax exemptions (public tax expenditures) should be transparent and we should attempt to get them in front of the public. This creates a rationale for evaluators to include unanticipated side effects or negative consequences....Many special interests don’t focus on the anticipated outcomes but the negative side-effects. The side-effects deserve to be examined with the same rigor as the intended outcomes.
Table 3
Panelists’ Feedback for Statements to be Rated in Round 2, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S10</em> I conduct evaluation with an eye towards informing public debate.</td>
<td>Mean 3.68</td>
</tr>
<tr>
<td></td>
<td>Variance 3.12</td>
</tr>
<tr>
<td></td>
<td>Median 3.00</td>
</tr>
<tr>
<td></td>
<td>Skewness -0.12</td>
</tr>
</tbody>
</table>

*Least Important Rationale*

My evaluation practice is [such that] the framing of issues/concerns for the evaluation comes from my interactions with my clients. The outcomes of any of my evaluations are similarly focused on the specific context from which they came and the ways in which people there can use the results. I do not conduct studies with the intention of “informing public debate”; to my mind that would be research rather than evaluation, unless you consider large-scale policy studies to be “evaluation” (but in that case the users would frame the study “with an eye towards informing public debate”).

*Highly Important Rationale*

My reported ratings of the importance of the twenty statements are contextual—they represent my views of importance for a particular type of evaluation work that I do but would be quite different for other types of evaluation work. Choosing my work with federal agencies as the exemplar to guide my ratings (I could have chosen others), I rated “informing public debate” as highly important because I see that as a critical role for federal evaluations. Leaving aside for now debates over the relative value of what have been labeled instrumental use and enlightenment, informing public debates is important for both the desired process of democratic deliberation and the desired impacts on public policy.
### Table 3  
*Panelists’ Feedback for Statements to be Rated in Round 2, cont.*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td>S11</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td><strong>Median</strong></td>
</tr>
<tr>
<td></td>
<td>4.00</td>
</tr>
</tbody>
</table>

**Least Important Rationale**

Maybe it’s a question of my understanding of what “transparency” means.... “Transparency” can mean many things. For instance does it refer to methodology?...Does it mean something about my motivations for doing the evaluation?...Does it mean telling the world what I have been doing? ...Sometimes I might dissemble a little bit because I am playing the role of organizational change consultant rolled in with that of evaluator. If that is the case, telling everything to everyone at all times can be counterproductive.

**Highly Important Rationale**

My thinking is that clients and other stakeholders (i.e., program funders, participants or data providers) should be aware of the data collection in progress, who’s collecting the data and why (the evaluation questions and design) and, as preliminary analysis permits, aware of interim results that might serve either a formative purpose or as a head’s up so that final results are not a shock.
Table 3
Panelists’ Feedback for Statements to be Rated in Round 2, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>S12 I conduct evaluation with an eye towards addressing social inequities.</td>
<td>Mean 3.43 Variance 2.99</td>
</tr>
<tr>
<td></td>
<td>Median 3.50 Skewness 0.01</td>
</tr>
</tbody>
</table>

**Least Important Rationale**

“Least important” does not mean “unimportant.”...There are many areas important to me....As individuals, we can choose areas of greatest significance to us in which to practice evaluation.

...However, I am not comfortable with requiring my field—evaluation—to embrace any or all of the causes which seem to me of great significance as the core value of our field or as essential to our practice. I see such a requirement as...

“my cause” advocacy, which can be counter to our special skills...[which help] us shine the light on what is happening and why across a wide range of concerns in a fair trustworthy manner....In so doing, we are likely to contribute to good government, broadly conceived, and the areas about which many of us may care.

**Highly Important Rationale**

Many of the programs we are asked to evaluate are focused on addressing disparities in health care, education, social services, environmental safety, and employment. Such disparities are commonly found to be associated with characteristics such as gender, disability, poverty, deafness, and race/ethnicity. By focusing on social inequities in our evaluations, we are in a better position to understand the cultural complexities that lead to the disparities and hence to contribute to effective solutions that support social justice.
Table 3
Panelists’ Feedback for Statements to be Rated in Round 2, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>S13</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
</tr>
</tbody>
</table>

Least Important Rationale
[This statement is] not consistent with my core value of providing accurate, trustworthy, and timely information to my clients (“getting it right”). “Getting it now” suggests that I might provide information I'm not confident is accurate (“right”) because my client needs something (anything) now to justify or make a decision. I see this as a form of malpractice and unfortunately too common in our field.

Highly Important Rationale
Use requires both accuracy and timeliness. Greater accuracy (“getting it right”) that is late (after decisions have been made) is useless. Timely data that lack basic credibility are also relatively useless. Thus the high importance of balance.
<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
<th>Least Important Rationale</th>
<th>Highly Important Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S14</strong> I operationalize concepts and goals before examining them systematically.</td>
<td><strong>Mean</strong> 4.14</td>
<td>Operationalizing concepts and goals should not be standard operating procedure in my view (nor) should (it)...be treated as a standard of excellence. Qualitative inquiry treats concepts as “sensitizing concepts” for exploration, inquiry and dialogue. For many concepts different stakeholders define them and use them differently. Premature operationalization often ignores this diversity. Operationalization is not appropriate in highly complex, dynamic and turbulent environments and situations...[It] can over-generalize a concept and reduce sensitivity to context. Finally, philosophy of science has found much operationalization flawed by positivist assumptions: <a href="http://en.wikipedia.org/wiki/Operationalization">http://en.wikipedia.org/wiki/Operationalization</a></td>
<td><strong>Mean</strong> 3.57</td>
</tr>
<tr>
<td></td>
<td><strong>Median</strong> 4.00</td>
<td></td>
<td><strong>Median</strong> 3.50</td>
</tr>
<tr>
<td></td>
<td><strong>Variance</strong> 2.20</td>
<td></td>
<td><strong>Variance</strong> 2.03</td>
</tr>
<tr>
<td></td>
<td><strong>Skewness</strong> -0.70</td>
<td></td>
<td><strong>Skewness</strong> -0.07</td>
</tr>
</tbody>
</table>

**S15** I devise action plans that guide how I subsequently examine concepts and goals.

**Least Important Rationale**

I thought we were planning an evaluation study, and I didn't see where action plans would come in.

**Highly Important Rationale**

Such action plans are needed to get agreement on concepts and goals among intended users of the evaluation.
Table 3
Panelists’ Feedback for Statements to be Rated in Round 2, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
<th>Least Important Rationale</th>
<th>Highly Important Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S16</strong> I question claims and assumptions that others make.</td>
<td>Mean: 4.39</td>
<td>#16 was low in importance for me because it was already addressed within #6: “I consider the credibility of different kinds of evidence in context.”</td>
<td>I question claims and assumptions others make because that IS our job—to question everyone’s assumptions and seek evidence for all claims...I see this point as fundamental to evaluation.</td>
</tr>
<tr>
<td></td>
<td>Median: 5.00</td>
<td></td>
<td>Mean: 4.36</td>
</tr>
<tr>
<td></td>
<td>Variance: 1.65</td>
<td></td>
<td>Median: 5.00</td>
</tr>
<tr>
<td></td>
<td>Skewness: -1.26</td>
<td></td>
<td>Skewness: -0.71</td>
</tr>
<tr>
<td><strong>S17</strong> I seek evidence for claims and hypotheses that others make.</td>
<td>Mean: 4.36</td>
<td>I interpreted this on the first round as “claims and hypotheses” external to the evaluation in question, i.e., out there in the literature. I do not highly value externalities like this, as I believe firmly that context matters in all claims and hypotheses.</td>
<td>It’s difficult to see how evaluators can rely on the findings of a study without being able to triangulate and validate what participants say. What they say provides both data for findings, more so in some studies than others, and also underlies the framework for interpreting results.</td>
</tr>
<tr>
<td></td>
<td>Median: 5.00</td>
<td></td>
<td>Mean: 4.36</td>
</tr>
<tr>
<td></td>
<td>Variance: 1.87</td>
<td></td>
<td>Median: 5.00</td>
</tr>
<tr>
<td></td>
<td>Skewness: -0.71</td>
<td></td>
<td>Skewness: -0.71</td>
</tr>
</tbody>
</table>
Table 3  
*Panelists’ Feedback for Statements to be Rated in Round 2, cont.*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
<th>Least Important Rationale</th>
<th>Highly Important Rationale</th>
</tr>
</thead>
</table>
| S20 | **Mean** 4.14  
**Variance** 1.61  
**Median** 4.00  
**Skewness** -0.05 | I don’t think it’s critical to “set aside time,” but I do think it’s important to reflect on your choices and your work. | If I did not set aside time to think about my work, I’d have a tough time improving it and a tougher time uncovering new opportunities for making distinctive contributions. |
| A1 | Not everything can or should be professionally evaluated. | N/A | |
| A2 | I design the evaluation so that it is responsive to the cultural diversity in the community. | N/A | |
| A3 | I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time. | N/A | |
| A4 | I do evaluations to develop capacity in program community members’ evaluation knowledge and practice. | N/A | |
| A5 | I work with stakeholders to articulate a shared theory of action and logic for the program. | N/A | |
| A6 | I consider stakeholders’ explicit and implicit reasons for commissioning the evaluation. | N/A | |
| A7 | I think about the criteria that would qualify an evaluand as “good” or “bad.” | N/A | |
| A8 | I consider the chain of reasoning that links composite claims to evaluative claims. | N/A | |
**Section 4. Round 2 Rating Form.**

Given the information provided above, please place each of the 20 statements that appear on page 5 into one of the following six categories of importance by writing the statement number on a line in the appropriate category.

Please note that each category should have no more than 5 statement numbers. That is, each statement number should appear **only once** in the table below.

<table>
<thead>
<tr>
<th>Highly Important</th>
<th>Very Important</th>
<th>Important</th>
<th>Moderately Important</th>
<th>Minimally Important</th>
<th>Least Important</th>
</tr>
</thead>
<tbody>
<tr>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
</tr>
<tr>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
</tr>
<tr>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
</tr>
<tr>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
<td># _______</td>
</tr>
</tbody>
</table>

*Thank you for your time and participation.*
Appendix H

DELPHI QUESTIONNAIRE 3

Dear Dr. (INSERT NAME HERE),

Thank you for your participation in my dissertation study about evaluative thinking. You have completed 2 questionnaires for the investigation and the information contained in this packet will be used for the next—and last—phase of the study. Round 3 requires panelists to review results from the previous survey and to re-rate the remaining items for which consensus has not been reached. To facilitate this task, I have organized all of the necessary information in this packet into the following sections:

- Section 1: List of Participating Panelists
- Section 2: Summary of Round 2 Results
- Section 3: Panelists’ Feedback Based on Round 2 Data
  - A: Overview of Statements to be Rated in Round 3
  - B: Panelists’ Feedback for Statements to be Rated in Round 3
- Section 4: Round 3 Rating Form

Each section contains an explanation of the information contained therein. For panelists, the goal of this phase is to rate the statements that appear in Section 3 after considering fellow experts’ feedback. Please return the item ratings by **Friday, February 1st, 2013**.

If any questions or concerns arise as you review materials provided in this packet, please feel free to contact me at annevo@ucla.edu.

Thank you for your time and participation.

Sincerely,

Anne T. Vo, M.A.
Principal Investigator
Doctoral Candidate

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779
**Section 1. List of Participating Panelists.**

This investigation is not possible without the valuable time and effort of the following 28 evaluation experts who have agreed to partake in the study. All panelists have completed Rounds 1 and 2 of the Delphi study. Additionally, all panelists are included in Round 3 of the investigation with the exception of Carol Weiss who passed in January 2013.

<table>
<thead>
<tr>
<th>Robert Boruch</th>
<th>Rodney Hopson</th>
<th>Hallie Preskill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eleanor Chelimsky</td>
<td>Ernest House</td>
<td>Sharon Rallis</td>
</tr>
<tr>
<td>J. Bradley Cousins</td>
<td>George Julnes</td>
<td>Debra Rog</td>
</tr>
<tr>
<td>Lois-ellin Datta</td>
<td>Jean King</td>
<td>Thomas Schwandt</td>
</tr>
<tr>
<td>Stewart Donaldson</td>
<td>Linda Mabry</td>
<td>William Shadish</td>
</tr>
<tr>
<td>Jody Fitzpatrick</td>
<td>Melvin Mark</td>
<td>Laurie Stevahn</td>
</tr>
<tr>
<td>Deborah Fournier</td>
<td>Donna Mertens</td>
<td>Carol Weiss</td>
</tr>
<tr>
<td>Jennifer Greene</td>
<td>Robin Miller</td>
<td>Joseph Wholey</td>
</tr>
<tr>
<td>Gary Henry</td>
<td>Jonathan Morell</td>
<td></td>
</tr>
<tr>
<td>Stafford Hood</td>
<td>Michael Patton</td>
<td></td>
</tr>
</tbody>
</table>
Section 2. Summary of Round 2 Results.

This section of the informational packet provides a summary of study results based on analysis of survey data from Round 2.

Survey Results. Experts reviewed and rated 20 statements on a 6-point scale in terms of their relative importance (1=least important; 6=highly important) when considering how to characterize evaluative thinking. Results indicate that the averaged mean rating and averaged variance of all items were 4.08 and 2.26, respectively. Mean ratings of individual statements ranged from 2.57 to 5.04, while their variances ranged from 0.81 to 3.51. Examination of 20 statements' mean ratings and variances relative to the averaged mean ($\mu = 4.08$) and averaged variance ($\sigma^2 = 2.26$) led to the identification of six statements (30%) for which panelists had reached consensus concerning importance level. Of these six statements, three were drawn from a list of items that panelists suggested for inclusion during Round 1 and, thus, were rated for the first time in Round 2. Table 1, below, specifies the statements on which consensus has been reached from Round 2 as well as each item's summary statistics. Additionally, Round 2 results suggest that experts' opinions about what was most important when thinking about how to describe evaluative thinking did not converge as quickly as in Round 1. Agreeing to what was least important in Round 2 remained a challenge for the group.

Table 1
Summary Statistics of Six Statements on Which Consensus Has Been Reached in Round 2

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>$\bar{x}$</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
<td>4.93</td>
<td>0.81</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
<td>5.04</td>
<td>1.81</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
<td>4.11</td>
<td>1.43</td>
</tr>
<tr>
<td>A3</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
<td>4.68</td>
<td>1.86</td>
</tr>
<tr>
<td>A5</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
<td>4.64</td>
<td>1.79</td>
</tr>
<tr>
<td>A6</td>
<td>I consider stakeholders' explicit and implicit reasons for commissioning the evaluation.</td>
<td>4.57</td>
<td>1.66</td>
</tr>
</tbody>
</table>
Section 3A. Panelists’ Feedback Based on Round 2 Data – Overview of Statements to be Rated in Round 3.

Table 2, below, contains a list of 14 statements to be re-rated in the current, final round. They are provided here for initial review.

Table 2
Overview of Statements to be Rated in Round 3

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
</tr>
<tr>
<td>S8</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
</tr>
<tr>
<td>S10</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
</tr>
<tr>
<td>S12</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
</tr>
<tr>
<td>S13</td>
<td>I balance “getting it right” and “getting it now.”</td>
</tr>
<tr>
<td>S14</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
</tr>
<tr>
<td>S15</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
</tr>
<tr>
<td>A1</td>
<td>Not everything can or should be professionally evaluated.</td>
</tr>
<tr>
<td>A2</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
</tr>
<tr>
<td>A4</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
</tr>
<tr>
<td>A7</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
</tr>
<tr>
<td>A8</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
</tr>
</tbody>
</table>
Section 3B. Panelists’ Feedback Based on Round 2 Data — Panelists’ Feedback for Statements to be Rated in Round 3.

Table 3, below, expands on information contained in the preceding section such that it contains summary statistics that indicate the ways in which the 14 statements from the previous round do not meet consensus criteria. Additionally, it highlights comments from panelists who rated these statements on either of the extreme ends of the importance scale (e.g., 1=least important; 6=highly important). Panelists are asked to review and use the summary statistics along with fellow colleagues’ feedback to inform how they assign statement ratings in this final round. Ratings will be recorded on the form that appears in Section 4.
## Table 3

*Panelists’ Feedback for Statements to be Rated in Round 3*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>I consider the answerability of an evaluation question before trying to</td>
<td>Mean  5.14  Variance  2.57</td>
</tr>
<tr>
<td>address it.</td>
<td>Median 6.00  Skewness  -1.87</td>
</tr>
</tbody>
</table>

### Least Important Rationale

I learned long ago that it is better to answer the right evaluation question however challenging that is from a measurement perspective than to answer the wrong question really well. Evaluations often raise important questions that are incredibly difficult to answer, but I would not let the “answerability” of any questions limit my asking them in the first place. If, instead, this statement means that once a question is chosen, an evaluator has to think about how to answer it, then I would rate that as important, but a tautology. Of course you have to think about how to answer any evaluation question once it is decided upon.

### Highly Important Rationale

In the governmental setting in which I planned my evaluations, it turned out to be extraordinarily important that an evaluation question be as precise and answerable as possible, so as to be sure of (a) being perceived as spending money wisely; (b) advancing the idea of evaluation as a routine way for policymakers to examine theories, programs, past history and ambient “knowledge” (i.e., conventional wisdom) about public problems; (c) removing at least one of the obvious impediments to use constituted by the failure of an evaluation to deliver a cogent set of findings; (d) making it harder for partisan interests to find fault with the work; and finally, (e) helping evaluation achieve its role of participant in the public debate that assures transparency and accountability in government.
Table 3  
Panelists’ Feedback for Statements to be Rated in Round 3, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
<td>2.57</td>
</tr>
<tr>
<td>Least Important Rationale</td>
<td>Median</td>
</tr>
<tr>
<td>[This statement is related to the next one; both] focused on what I am willing to challenge during the course of an evaluation. As usual, my thinking is contingent on circumstances, and this aspect serves to differentiate the importance I accorded each statement. Ideologies are always among the contextual variables in conducting an evaluation, [but] their frequency and magnitude have varied enormously in my practice....Ideological challenges figure infrequently in my practice, possibly because my ideological commitments are apparent in some of my [writings].</td>
<td>2.00</td>
</tr>
<tr>
<td>Highly Important Rationale</td>
<td></td>
</tr>
<tr>
<td>I believe that evaluators should challenge ideology, which are often untested assumptions, especially about program performance. For example, school voucher advocates maintain that vouchers will improve educational outcomes due to competition for students. If I were evaluating a voucher program the ideologically based assumptions would be a top priority for rigorous evaluation. This is true for any policy or program. Untested beliefs should be identified and tested whether the beliefs are from the right, left or centrist ideologies.</td>
<td></td>
</tr>
</tbody>
</table>
Table 3
Panelists’ Feedback for Statements to be Rated in Round 3, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
<th>Least Important Rationale</th>
<th>Highly Important Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
<td><strong>Mean</strong> 2.68  <strong>Variance</strong> 2.74</td>
<td>Part of my response centers on the term “special interests.” It can mean as seen from a special viewpoint, e.g., teachers, or as vested interests, e.g., teacher unions, which has a pejorative overtone. I don't want to enter any evaluations where I have judged people preemptively. All those involved are entitled to having their opinions heard honestly and fairly. After collecting evidence, I might decide some have taken overly selfish positions or behaved too opportunistically.</td>
<td>Challenging special interests means that power relations are explicitly identified and the basis for the power is examined. In this way, power derived from unearned privilege can be interrogated and the need to address power inequities can be made visible. This is undertaken with a goal of insuring representation of voices that are not in powerful positions or moderating the effect of those who may represent a minority voice that is not in the best interests of those most in need.</td>
</tr>
<tr>
<td>S10 I conduct evaluation with an eye towards informing public debate.</td>
<td><strong>Mean</strong> 4.32  <strong>Variance</strong> 2.15</td>
<td>My job as an evaluator is to work with my clients (my primary intended users) to help them learn what matters to them in moving forward. If these clients were policy makers, then I suppose the evaluation might help to inform public debate, but that is not the world in which I move and I do not keep an eye on that particular prize. That is why I rated it “least important.”</td>
<td>The high rating I gave to both of these statements reflects four factors that are exogenous to the evaluation thinking process, as well as one that is endogenous. The exogenous factors are: setting; goal(s); a priority given to use; and the effort to constrain predictable political obstructions related to setting, goal and use. The endogenous one is the particular conception I have of the role of evaluation in government.</td>
</tr>
</tbody>
</table>
Table 3
*Panelists’ Feedback for Statements to be Rated in Round 3, cont.*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>S12 I conduct evaluation with an eye towards addressing social inequities.</td>
<td>Mean 3.61, Variance 2.84</td>
</tr>
<tr>
<td></td>
<td>Median 4.00, Skewness -0.07</td>
</tr>
</tbody>
</table>

**Least Important Rationale**

I conduct evaluation to give as fair and honest a “test” of an intervention (if it is an outcome evaluation). If that program is designed to address social inequities, then I am assessing whether IT does indeed address them. I also strive to conduct evaluation in a manner that includes all key voices, especially voices of the beneficiaries of the programs. This is important to me, to have a pluralistic view and assessment of the intervention. I conduct evaluations because I think they can make a difference, and yes, I think they can level the playing field. But I don’t think I approach each individual evaluation with that stance. My stance is one based on what is an appropriate, complete, thorough, fair test of the intervention from all perspectives.

**Highly Important Rationale**

The particular lenses I bring into evaluation [are] based on my own lived experience....I find my leanings in evaluation foster questions that address power, racial, class, sexual differentials, deliberately asking questions that tease these larger issues. The last few evaluations that I have been contracted to do very much address issues of inequities and I find that these commissioners appreciate a discussion about evaluation but especially with those who understand how real these issues [are] in the work we do.
Table 3  
Panelists’ Feedback for Statements to be Rated in Round 3, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
</table>
| **S13** I balance “getting it right” and “getting it now.” | **Mean** 3.93  
**Variance** 2.96 |
|                                                | **Median** 4.00  
**Skewness** -0.12 |

**Least Important Rationale**
Accuracy or validity is the most important criterion for judging an evaluation. The reason for conducting “systematic inquiry” is to produce findings that are valid and reliable. When the quick answer is unlikely to be accurate or sufficiently precise, it is better to get it right or leave the evaluation undone than to provide quick and incorrect findings.

**Highly Important Rationale**
[This] balance...is highly important [due to] the (i) type & characteristics of the evaluand & the (ii) realities of the context in which it resides....[T]he type of evaluand with which I typically work include large-scale, adaptive research enterprises and multi-site programs. [The] evaluand [is usually] nested within evaluands...& is best characterized...as both complicated (lots of moving parts across multiple sites & stakeholders that are firing at different times and rates) & complex (lots of interacting parts that independently respond & adapt to one another & in doing so generate unexpected novel behavior for the system as a whole)....Decision-making that guides such strategy & programming development & resource allocation entail power distributions & organizational politics that are among the most influential of features with this type of evaluand. [Thus, my evaluations are typically designed for] a moving target...[and] must take into account both fixed, preordinate features...and flexible, adaptive ones...in order to provide the most accurate & relevant data that can be used by administrators leading the research enterprise. These two key features ensure an evaluation plan that continues to be robust because fixed designs do not deal with the realities of rapid change, nor recognize the major practical significance of the decision-making context.
Table 3  
*Panelists’ Feedback for Statements to be Rated in Round 3, cont.*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>I operationalize concepts and goals before examining them systematically.</td>
<td>Mean 3.75</td>
</tr>
<tr>
<td></td>
<td>Variance 2.42</td>
</tr>
<tr>
<td></td>
<td>Median 4.00</td>
</tr>
<tr>
<td></td>
<td>Skewness -0.31</td>
</tr>
</tbody>
</table>

*Least Important Rationale*

My rating is based on the fact that I don't (a) fully understand what it means and (b) don't see how it relates to matters of evaluative reasoning.

*Highly Important Rationale*

I cannot figure out how to examine concepts or goals systematically until they get defined (and definition(s) could take a lot of more or less systematic negotiation among stakeholders) and operationalized (following negotiation). After that, the systematic work for me involves systematic acquisition of data/information on the achievement of goals.
<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
<th>Least Important Rationale</th>
<th>Highly Important Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
<td><strong>Mean</strong> 3.18</td>
<td><strong>Mean</strong> 3.00</td>
<td>Without some plan of action regarding how key concepts and goals (e.g., those criteria of merit that are initially identified) are to be addressed, the evaluation is not likely to result in warranted judgments of merit and worth. This is not to say that additional concepts cannot emerge. Nor is it to say that plans of action won’t have to be modified. Rather, it’s to say that, without a sensible plan to guide upcoming actions and inquiry, problems ensue.</td>
</tr>
<tr>
<td><strong>Variance</strong> 2.74</td>
<td><strong>Skewness</strong> 0.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Although rating this item as “least important” does not mean I think it is unimportant, there are plenty of other statements that we rated in both rounds that better capture the notions that one must carefully consider how best to translate concepts into concrete operations, develop criterion which can inform evaluative judgments, and devise plans to obtain evaluative evidence that is acceptable given the purpose and nature of a particular evaluation. These other statements better reflect the unique critical reasoning processes I associate with evaluative thinking. Relative to other statements, this statement stood out as among the least helpful in distinguishing evaluative thinking from thinking associated with any other activity for which one might need a plan about how best to move forward.
<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S16</strong> I question claims and assumptions that others make.</td>
<td><strong>Mean</strong> 4.86</td>
</tr>
<tr>
<td></td>
<td><strong>Median</strong> 5.00</td>
</tr>
</tbody>
</table>

**Least Important Rationale**

I *do* conduct my evaluations with democratic values commitments....I perceive these commitments as structural and societal, not personal. At the personal level, I invoke commitments to respect and tolerance for others’ viewpoints, and commitments to dialogue and conversation. Thus the low ratings for [this and other similar] items.

**Highly Important Rationale**

Questioning claims and exploring assumptions through seeking evidence and logical analysis is at the heart of an evaluator’s work....I immediately associate “claims and assumptions” with the common assertions program staff and managers, consumers, funders, and other stakeholders make about a program’s merit, value or worth (or lack of) and with the assumptions or presumed logic/theory of a program. These are an important driver of my search for evidence to support, refute, clarify, etc. the basis of those claims....Attending closely to claims that may or may not be well substantiated can be a productive part of focusing an evaluation.
<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>A1</td>
<td>4.11</td>
</tr>
<tr>
<td></td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td>4.00</td>
</tr>
</tbody>
</table>

**Least Important Rationale**

I very much agree with this statement, but it has more to do with policy or management decisions than evaluation practice. So, in terms of importance for what an evaluator can do, we have little power over these decisions. We should try to have more.

**Highly Important Rationale**

Evaluation in one sense is like breathing: we do it all the time. Evaluation in the more systematic sense of our field is not something we do all the time....[S]ystematic evaluation has direct costs to everyone concerned in time and money. It can have indirect costs like poking a stick in a bee-hive. Our pieties can be that benefits outweigh these costs in all situations, but sometimes, the benefits seem primarily the full employment of evaluators. The saying “If it works, don't fix it” should be considered before we do an evaluation, particularly an evaluation required by funders, thinking more Shaker simplicity than Gaudi.
Table 3  
*Panelists’ Feedback for Statements to be Rated in Round 3, cont.*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
<td><strong>Mean</strong></td>
<td><strong>Median</strong></td>
</tr>
<tr>
<td>A2</td>
<td>4.50</td>
<td>4.50</td>
</tr>
</tbody>
</table>

**Least Important Rationale**  
I do evaluation for very specific purposes [which does not include a focus on cultural diversity]....I may have larger societal concerns as a person, but as an evaluator I’m...very practical [and have] narrow interests....Some of [the stakeholders’] need is to know what is happening and why in order to improve the program. Some is for political reasons. Both are legitimate....The questions I do address are the immediate needs of work that is consuming taxpayer money....Sometimes I am able to convince a funder that some of [the] broader questions should be included, and when I do I feel really, really good about it. But notice that none of these important larger questions has anything to do with “cultural diversity” in any but the most tenuous sense....Cultural diversity is just too far down the priority list of broader questions that I’m willing to try to get included in the evaluation scope.

**Highly Important Rationale**  
I think it is important to draw upon a wide range of theories and methods to design an evaluation that is optimally matched to the context and responsive to the cultural diversity in the community of interest.
Table 3  
*Panelists’ Feedback for Statements to be Rated in Round 3, cont.*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
<th>Least Important Rationale</th>
<th>Highly Important Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4 I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
<td>Mean: 3.61</td>
<td>Least Important Rationale:&lt;br&gt;Not every evaluation should or can focus on capacity building. Large-scale evaluations, which are the type I often conduct, afford limited opportunities to engage in capacity building with stakeholders. Developing their knowledge of the program outcomes is important but then is different that evaluation capacity building.</td>
<td>Highly Important Rationale:&lt;br&gt;We've learned that evaluation use is neither natural nor easy. Knowledgeable intended users are more likely to become knowledgeable actual users. Thus, every evaluation is also an opportunity to enhance capacity, teach and train intended users, not just for the specific evaluation underway, but to enhance future evaluations as well, including a commitment to support and engage in evaluative thinking and use in the future.</td>
</tr>
</tbody>
</table>
Table 3
Panelists’ Feedback for Statements to be Rated in Round 3, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
<td><strong>Mean</strong> 3.89  <strong>Variance</strong> 3.51</td>
</tr>
<tr>
<td><strong>Median</strong> 4.00  <strong>Skewness</strong> -0.34</td>
<td></td>
</tr>
</tbody>
</table>

**Least Important Rationale**
To me, an evaluand is what will be evaluated—i.e., the program, organization, policy, etc. So, when I think about any given evaluand, I think about describing it...and determining its evaluability (i.e., the extent to which [it] can and/or should be evaluated to meet a specified/desired evaluation purpose—i.e., the evaluand’s readiness for evaluation in light of what stakeholders/clients wish to know). I never think of an evaluand as “good” or “bad”—instead I think about its evaluability, and if determined evaluable, then how best to frame evaluation questions and designs, collect and analyze data to address those questions, interpret the findings, and make recommendations—all in light of the evaluation study’s strengths and limitations.

**Highly Important Rationale**
Fundamentally, this is what evaluators do. How can we evaluate anything without a solid understanding of quality criteria?
Table 3
Panelists’ Feedback for Statements to be Rated in Round 3, cont.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>A8 I consider the chain of reasoning that links composite claims to evaluative claims.</td>
<td><strong>Mean</strong> 4.18 <strong>Variance</strong> 2.08</td>
</tr>
<tr>
<td>Least Important Rationale</td>
<td><strong>Median</strong> 4.00 <strong>Skewness</strong> -0.74</td>
</tr>
</tbody>
</table>

I’ve no idea what the adjective “composite” means and in what context.

Highly Important Rationale
A good chain of reasoning between findings and conclusions is critical to good evaluation practice. It is one of the primary things that distinguishes evaluation from judgment or criticism.
Section 4. Round 3 Rating Form.

Given the information provided above, please place each of the 14 statements that appear on page 4 into one of the following six categories of importance by writing the statement number on a line in the appropriate category.

Please note that each category should have no more than 3 statement numbers. That is, each statement number should appear only once in the table below.

<table>
<thead>
<tr>
<th>Highly Important</th>
<th>Very Important</th>
<th>Important</th>
<th>Moderately Important</th>
<th>Minimally Important</th>
<th>Least Important</th>
</tr>
</thead>
<tbody>
<tr>
<td># ______</td>
<td># ______</td>
<td># ______</td>
<td># ______</td>
<td># ______</td>
<td># ______</td>
</tr>
<tr>
<td># ______</td>
<td># ______</td>
<td># ______</td>
<td># ______</td>
<td># ______</td>
<td># ______</td>
</tr>
</tbody>
</table>

Optional: Please use the space below to comment on the ratings provided above. Please note that the box will automatically expand to accommodate the length of your comments.

Thank you for your time and participation.
Appendix I

POST-DELPHI FOLLOW-UP MESSAGE

Dear Dr. (INSERT NAME HERE),

In addition to thanking you for participating in my dissertation study about evaluative thinking, I am writing to share results from the final survey as well as cumulative results of the investigation. Round 3 results are summarized below, followed by an overview of cumulative study findings.

Summary of Results.

Round 3 Results. Experts reviewed and rated 14 statements on a 6-point scale in terms of their relative importance (1=least important; 6=highly important) when considering how to characterize evaluative thinking. Results indicate that the averaged mean rating and averaged variance of all items were 3.74 and 2.53, respectively. Mean ratings of individual statements ranged from 2.26 to 4.74, while their variances ranged from 1.58 to 3.23. Examination of 14 statements’ mean ratings and variances relative to the averaged mean ($\mu = 3.74$) and averaged variance ($\sigma^2 = 2.53$) led to the identification of four statements (14%) for which panelists had reached consensus concerning importance level. Table 1, below, specifies the statements on which consensus has been reached from Round 3 as well as each item’s summary statistics. Consistent with results from Round 2, findings from the final survey administration suggest that experts’ opinions converged quicker earlier in the study and that agreeing to what was least important remained a challenge for the group. The highly contextual nature of experts’ practices was the reason most frequently provided for disagreement.

Table 1
Summary Statistics of Four Statements on Which Consensus Has Been Reached in Round 3

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement</th>
<th>$\bar{x}$</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
<td>2.26</td>
<td>1.58</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
<td>4.74</td>
<td>1.58</td>
</tr>
<tr>
<td>A1</td>
<td>Not everything can or should be professionally evaluated.</td>
<td>4.30</td>
<td>1.99</td>
</tr>
<tr>
<td>A4</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
<td>3.37</td>
<td>2.17</td>
</tr>
</tbody>
</table>
Cumulative Results. Throughout the course of the Delphi, experts reviewed and rated a total of 28 statements. Overall, results indicate that experts reached consensus on:

- Eight of 28 statements (29%) in Round 1, most of which dealt with the importance of evidence while reasoning;
- Six of 28 statements (21%) in Round 2, most of which dealt with behavioral/procedural processes; and
- Four of 28 statements (14%) in Round 3, which pertained not only to procedural issues, but also the purpose of evaluation.

Of the statements rated, those considered:

- “Very Important” included Statements S3, S4, S6, S11, S17–19;
- “Important” included Statements S1, S2, S5, S14, S16, S20, A1–3, A5–8;
- “Moderately Important” included Statements S7, S8, S10, S12, S13, S15, A4;
- “Minimally Important” included Statement S9.

Interestingly, six out of eight statements that fell into the “Moderately Important” category as well as six out of 13 statements that were considered “Important” to evaluative thinking were also those for which consensus had not been reached by the end of the third survey administration. With respect to the “Moderately Important” category, these items consisted of Statements S8, S10, S12, S13, S15, A4. In terms of the “Important” category, such items consisted of Statements S1, S14, A1, A2, A7, A8. These 12 items are denoted with an asterisk [*] in Table 2, below.

<table>
<thead>
<tr>
<th>Statement #</th>
<th>Statement Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1*</td>
<td>I consider the answerability of an evaluation question before trying to address it.</td>
</tr>
<tr>
<td>S2</td>
<td>I consider the availability of resources when setting out to conduct an evaluation.</td>
</tr>
<tr>
<td>S3</td>
<td>I consider the importance of various kinds of data sources when designing an evaluation.</td>
</tr>
<tr>
<td>S4</td>
<td>I consider alternative explanations for claims.</td>
</tr>
<tr>
<td>S5</td>
<td>I consider inconsistencies and contradictions in explanations.</td>
</tr>
<tr>
<td>S6</td>
<td>I consider the credibility of different kinds of evidence in context.</td>
</tr>
<tr>
<td>S7</td>
<td>I conduct evaluation with an eye towards challenging personal beliefs and opinions.</td>
</tr>
<tr>
<td>Statement #</td>
<td>Statement Description</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>S9</td>
<td>I conduct evaluation with an eye towards challenging special interests.</td>
</tr>
<tr>
<td>S8*</td>
<td>I conduct evaluation with an eye towards challenging unquestioned ideology.</td>
</tr>
<tr>
<td>S10*</td>
<td>I conduct evaluation with an eye towards informing public debate.</td>
</tr>
<tr>
<td>S11</td>
<td>I conduct evaluation with an eye towards transparency.</td>
</tr>
<tr>
<td>S12*</td>
<td>I conduct evaluation with an eye towards addressing social inequities.</td>
</tr>
<tr>
<td>S13*</td>
<td>I balance “getting it right” and “getting it now.”</td>
</tr>
<tr>
<td>S14*</td>
<td>I operationalize concepts and goals before examining them systematically.</td>
</tr>
<tr>
<td>S15*</td>
<td>I devise action plans that guide how I subsequently examine concepts and goals.</td>
</tr>
<tr>
<td>S16</td>
<td>I question claims and assumptions that others make.</td>
</tr>
<tr>
<td>S17</td>
<td>I seek evidence for claims and hypotheses that others make.</td>
</tr>
<tr>
<td>S18</td>
<td>I offer evidence for claims that I make.</td>
</tr>
<tr>
<td>S19</td>
<td>I make decisions after carefully examining systematically collected data.</td>
</tr>
<tr>
<td>S20</td>
<td>I set aside time to reflect on the way I do my work.</td>
</tr>
<tr>
<td>A1*</td>
<td>Not everything can or should be professionally evaluated.</td>
</tr>
<tr>
<td>A2*</td>
<td>I design the evaluation so that it is responsive to the cultural diversity in the community.</td>
</tr>
<tr>
<td>A3</td>
<td>I modify the evaluation (e.g., design, methods, and theory) when evaluating a complex, complicated evaluand as it unfolds over time.</td>
</tr>
<tr>
<td>A4*</td>
<td>I do evaluations to develop capacity in program community members’ evaluation knowledge and practice.</td>
</tr>
<tr>
<td>A5</td>
<td>I work with stakeholders to articulate a shared theory of action and logic for the program.</td>
</tr>
<tr>
<td>A6</td>
<td>I consider stakeholders’ explicit and implicit reasons for commissioning the evaluation.</td>
</tr>
<tr>
<td>A7*</td>
<td>I think about the criteria that would qualify an evaluand as “good” or “bad.”</td>
</tr>
<tr>
<td>A8*</td>
<td>I consider the chain of reasoning that links composite claims to evaluative claims.</td>
</tr>
</tbody>
</table>
Further examination of ratings that experts assigned to these eight statements per survey administered indicated that panelists’ opinions continued to diverge as the study progressed. This observation was substantiated by unstable ratings and increased variation in ratings with each new round. Upon further analyses, it seems that dissensus could be attributed to ambiguous phrasing for some items and differences in respondents’ contexts for other items.

If any questions or concerns arise as you review materials that I have provided, please feel free to contact me at annevo@ucla.edu.

Thank you for your time and continued support.

Sincerely,

Anne T. Vo, M.A.
Principal Investigator
Doctoral Candidate

UCLA Graduate School of Education & Information Studies
Social Research Methodology (SRM) Division
Box 951521, Moore Hall 2027
Los Angeles, CA 90095-1521

E: annevo@ucla.edu
P: 310.845.6779
References


Futures, 1, 35–44.

http://www.eval.org/doestatement.htm


http://www.eval.org/eptf.asp

http://archive.constantcontact.com/fs014/1100694346064/archive/1102975077013.html


Retrieved from
http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_fy2009/m09-20.pdf


Schauer, F. (2008). Why precedent in law (and elsewhere) is not totally (or even substantially) about analogy. *Perspectives on Psychological Sciences, 3*, 454–460.


Sage Publications.


