The practice of evaluative reasoning in the Aotearoa New Zealand public sector

HEATHER NUNNS

A thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy at Massey University, Wellington, New Zealand

ABSTRACT

This study argues that sound evaluative reasoning, defined as “the systematic means for arriving at evaluative conclusions . . . the principles that support inferences drawn by evaluators” (Fournier, 1995, p.1), is an essential element of evaluation quality. As such, evaluative reasoning is a lens through which to consider how to improve the quality of evaluations undertaken or commissioned by the Aotearoa New Zealand public sector. The argument is grounded in the theory of evaluation derived from western philosophy, specifically, informal logic. This theory underpins the conceptualisation and design of this study examining how evaluative reasoning is understood and practised by professionals who undertake public sector evaluation in Aotearoa New Zealand. A multiple method research design is used to generate diverse understandings of the topic and offer opportunities for abductive thinking. The methods used are Q methodology, meta-evaluation, and key informant interviews with local and international evaluation experts.

The findings from this study point to three ways in which evaluative reasoning has an impact on the quality of evaluation. It increases the robustness of the reasoning chain from value claim to evaluative conclusion/judgment; underpins the professional competencies required of evaluation practitioners; and reinforces the ethical dimensions of evaluation practice in a public sector context. Lastly, two abductively-derived conjectures point evaluators toward diverse ways of knowing in their reasoning from evaluative claim to evaluative conclusion/judgment. Amplifying the work of previous theorists, it is suggested that expert intuition and abductive inference provide further paths of evaluative knowing in addition to inductive logic and probative inference.

ACKNOWLEDGEMENTS

This thesis represents a six year journey, combining study with work and family life. It is fitting therefore that my first thanks is to my husband Ray who has provided emotional support as well as practical support in day-to-day household activities. I am truly in your debt. Thanks also to our daughter Megan who since the age of five has been accustomed to a mother with her head in a book as I have worked through post graduate studies. Thank you Megan for your understanding and patience.

I also thank my supervisors Associate Professor Robin Peace and Professor Karen Witten. Your helpful guidance, together with your enthusiasm about my study and confidence in me has been hugely encouraging and kept me going during the times when I questioned whether my thinking made any real sense.

My sincere thanks are also due to my valued colleagues - Mathea Roorda, Robyn Bailey, Rae Torrie, and Vicki Wilde. Since 2008 we have met quarterly for professional support and development purposes. Your ongoing interest in my study, willingness to listen to my latest mental block, and feedback on my work is greatly appreciated.

Lastly, I want to thank Massey University's Distance Library Service. Their behind-the-scenes support of distance students is of immense value.

This thesis is dedicated to my mother Barbara, who like many of her generation growing up in England in the war years did not have the opportunity for education beyond 14 years of age.

TABLE OF CONTENTS

PART A

PART D

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Aroha	Love, affection, sympathy, charity, compassion.
Evaluand	A generic term for whatever is being evaluated.
Fono	Councils or meetings. Applies to national assemblies and legislatures, as well as local village councils or any type of meeting between people.
Haka	To dance, perform the haka - vigorous dances with actions and rhythmically shouted words.
Hīkoi	Step, march, hike.
Iwi	Extended kinship group, tribe, nation, people, nationality, race.
Kanohi ki kanohi	Face to face, in person.
Karakia	Prayer, grace, blessing.
Kaumātua	Elders, man or woman, who are held in high esteem.
Kaupapa Māori	Customary practice, principles incorporating the knowledge, skills, attitudes and values of Māori society.
Kāwanatanga	Government, dominion, rule, authority, governorship.
Koha	Gift, offering, donation, contribution.
Mana	Prestige, authority, control, power, influence, status, spiritual power, charisma.
Manaakitanga	Hospitality, kindness, generosity, support. The showing of respect, generosity and care for others.
Māori	The indigenous people of Aotearoa New Zealand.
Marae	The open area in front of the wharenui (the main building of a marae) where formal greetings and discussions take place. Often also used to refer to the complex of buildings around the marae.
Mātauranga	Knowledge, wisdom, understanding, skill.
Pākehā	New Zealander of European descent.
Palagi	A person of European descent.

Pasifika	Pasifika does not refer to a single ethnicity, nationality, gender or culture. The term is one of convenience used to encompass a diverse range of peoples from the South Pacific region now living in Aotearoa New Zealand who have family and cultural connections to their South Pacific countries of origin.
Pōwhiri	Welcome ceremony on a marae.
Rangatira	Chief.
Tangata whenua	Local people, hosts, indigenous people - people born of the whenua (land).
Taonga	Treasure, anything prized. Applied to anything considered to be of value including socially or culturally valuable objects, resources, phenomenon, ideas and techniques.
Tapu	That which is sacred, prohibited, restricted, set apart.
Te Ao Māori	Māori world.
Te Puni Kōkiri	Ministry of Māori Development.
Te Reo	Māori language.
Te Tiriti o Waitangi	The Treaty of Waitangi.
Tikanga	The customary system of values and practices that have developed over time and are deeply embedded in the social context.
Tino Rangatiratanga	Self-determination.
Tūrangawaewae	Place where one has the right to stand. Place where one has rights of residence and belonging through kinship and whakapapa.
Va	Va is a Samoan concept relating to the space between, which is not empty or void but is relational and sacred.
Wairuatanga	Spirituality.
Whakapapa	Genealogy, lineage, descent.
Whānau	Extended family, family group.
Whānaungatanga	A relationship through shared experiences and working together which provides people with a sense of belonging.

ABBREVIATIONS

AEA	American Evaluation Association
AES	Australasian Evaluation Society
ANZEA	Aotearoa New Zealand Evaluation Association
APA	American Psychological Association
CBA	Cost benefit analysis
CEA	Cost effectiveness analysis
EES	European Evaluation Society
EVALTALK	The email discussion group of the American Evaluation Association
FBI	Federal Bureau of Investigation
HIA	Health impact assessment
ISO	International Organisation for Standardization
OECD	Organisation for Economic Co-operation and Development
Q	Q methodology
SSC	State Services Commission
SPEaR	Social Policy Evaluation and Research Committee
UK	United Kingdom
UNDP	United Nations Development Programme
UNICEF	United Nations International Children’s Fund
USA	United States
VFM	Value for money
WEG	Wellington Evaluation Group

INTRODUCING THE STUDY

This study explores how evaluative reasoning is understood and practised in the context of public sector evaluation in Aotearoa New Zealand. My argument is that evaluative reasoning is a lens through which to consider how to improve the quality of evaluations being conducted or commissioned by Aotearoa New Zealand public sector agencies. The objectives of the study are to (i) present a theoretical account of evaluative reasoning from western philosophy and evaluation literatures; (ii) examine the practice of evaluative reasoning in the Aotearoa New Zealand public sector (the public sector); (iii) identify contextual factors that influence how evaluative reasoning is being practised in the public sector; (iv) generate insights into how evaluative reasoning practice can be improved.

The research questions are: (i) How is evaluative reasoning understood and practised by professionals working in or commissioned by the public sector? (ii) How do contextual factors influence how evaluative reasoning is practised in the public sector? and (iii) How can evaluative reasoning practice be strengthened in the public sector context?

This study is presented in four parts: Part A (chapters 1, 2 and 3) situates the thesis and describes the research methodology, Part B (chapters 4 and 5) provides a theoretical foundation for the study, Part C (chapters 6, 7 and 8) offers three perspectives on evaluative reasoning in the Aotearoa New Zealand public sector, and Part D (chapters 9 and 10) integrate and interpret the findings from these three perspectives.

PART A

SITUATING THE RESEARCH

> Researchers are neither neutral nor objective, but rather are inevitably and intrinsically interested inquirers . . . interested implies a situated inquirer, one who inevitably brings to the process of social inquiry his or her own sociocultural history, beliefs about the social world and about what constitutes warranted knowledge of it, theoretical preferences, and moral and political values (Greene, 2011, p.81, 82).

Greene (2011) reminds us of the situated nature of research and the researcher in relation to the topic of the research. It is therefore appropriate that Part A situates the study by introducing the research topic and its rationale, and presenting the researcher’s epistemological stance (chapter 1); explaining the multiple methods design (chapter 2); and describing the context for the research, namely, Aotearoa New Zealand and its public sector (chapter 3).

CHAPTER 1

INTRODUCTION

Chapter 1 presents the topic of this study and argues the reasons for its relevance. An overview of the thesis structure is then provided. I then describe my epistemological stance and the theoretical framework underpinning the study.

1.1 Defining evaluation

The term evaluation describes a cognitive act that occurs in every aspect of everyday life - the shopper in the supermarket comparing one product with another, the parent with a sick child assessing whether their symptoms require a visit to the doctor. Similarly, evaluation underpins workplace activity, whether it is a motor mechanic doing a safety check on a car, or a farmer considering whether pasture levels are adequate for stock to feed on.

Moreover, evaluation underpins intellectual and professional endeavours such as science, medicine, engineering, planning and law. Such examples illustrate what Scriven (1991) describes as the “transdisciplinary” (p.363) nature of evaluation, that is, it is a discipline (like logic and statistics) that underpins other disciplines. In more recent years, Scriven (2013a) has referred to evaluation as the “alpha discipline” (p.29) because it offers the means by which other disciplines (such as those listed above) may be examined and assessed.

1.2 Evaluation as a western knowledge construct

The 1,000 plus pages of the Oxford Companion to Philosophy (2005) demonstrate that what is known in western societies as logical reasoning is the outcome of social and historical influences spanning over two thousand years, beginning in 600 BC with the first Greek philosophers. Thus logic and the knowledge produced by the application of such logic are socially constructed, arising from human intellectual endeavour. Consequently, it is important to clarify that the constructs of evaluation and evaluative reasoning used in this study are derived from western philosophy, specifically informal logic as described in chapter 4.

While this study does not draw on indigenous epistemologies to conceptualise evaluative reasoning, Māori epistemologies feature in this study (Māori are the indigenous people of Aotearoa New Zealand and comprise 15 percent of the population).[1] This is due to the involvement of Māori evaluators in public sector evaluation, some of whom were participants in the Q methodology study (chapter 6). Further, the meta-evaluation (chapter 7) includes three evaluations conducted according to Kaupapa Māori principles (Kaupapa Māori refers to the customary practice and principles incorporating the knowledge, skills, attitudes and values of Māori society). Given the study's focus on western epistemologies, it could be argued that these evaluators and the three evaluation reports should have been excluded from the study. However, the Māori participants in the Q study are working in or for the public sector. Similarly, the three evaluation reports have been commissioned by the public sector and, as such, are in a western domain of influence based on the Westminster system of democratic accountability. This raises a challenging question of whether an indigenous logic of evaluative reasoning is needed. Whatever the potential answer to such a question, it is beyond the scope of consideration here. However it does raise an important question for future research. In the meantime, I have chosen to reflect on the findings from the three reports from the meta-evaluation and Māori evaluators who participated in the Q study as they add richness to our understanding about evaluative reasoning practice in Aotearoa New Zealand. This has also enabled some consideration of the influence that Māori epistemologies are having on the practice of evaluative reasoning in Aotearoa New Zealand, as discussed in chapter 9. While there is an increasing number of Pasifika evaluators, the influence of Pasifika epistemologies on evaluation practice is not as obvious at the present time. (The Tongan term Pasifika is used in this thesis to refer to peoples from the Pacific Islands who have a significant presence in Aotearoa New Zealand, making up 7 percent of the population and, like Māori, are subject to a range of policy interventions).

[1] The Māori to English translations are taken from Moorfield (2005).

1.3 Defining the evaluation space of this study

This study focuses on a specific application of the discipline of evaluation, namely, evaluation for the purposes of assessing public sector policy, and the impacts of such policy. Two definitions of evaluation are provided for this context. Weiss' (1998) definition emphasises the comparative nature of evaluation, in addition to addressing the purpose and intended outcome of evaluation: "Evaluation is the systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy" (p.4). The second definition provided by Scriven (1991) identifies the assessment of value as the primary focus of evaluation: "Judgment of the merit (quality), worth (value), significance (importance) of an evaluand" (p.13). (The term evaluand is a generic term for whatever is being evaluated (Scriven, 1991, p.139)).

These definitions provide expansive boundaries, enabling evaluation practice to be conceptualised in different ways, for example as a technical or management or professional practice (Stern, 2006). The primary concern of this study is not with such conceptualisations of practice, but rather with the theoretical grounds on which the practice of evaluation in a western knowledge context is based. The study is premised on evaluation practice as being founded on a theory-based discipline derived from western philosophy. (This premise is elaborated in chapters 4 and 5). Weiss (1998) asserts that evaluation is a craft. My focus is on the theory that underpins this craft and the implications of this theory for the craft.

1.4 Defining evaluative reasoning

Two definitions of evaluative reasoning are provided. The first from Blair (1995), a logician, is chosen for its comprehensive description:

> . . . the reasoning involved in drawing inferences from the information base plus the evaluative principles (criteria, standards, and so on) to the evaluative conclusions, in the generation of criteria and of information-gathering methods (warrant-establishing reasoning for example), and in the formation of an overall evaluative conclusion based on the component evaluative judgments (p.78).

The second is from Fournier (1995), an evaluation theorist whose work has contributed to the evaluative reasoning discourse: " . . . the systematic means for arriving at evaluative conclusions . . . the principles that support inferences drawn by evaluators" (p.1).

The methodology literature offers a useful insight for this study. Maxwell (1992) identifies five types of validity in qualitative research, one of which is evaluative validity (p.295) defined as "evaluative judgments about a study are warranted and appropriate" (Bamberger, Rugh & Mabry, 2006, p.146). Evaluative reasoning, from the perspective of my research, concerns the validity of inferences, claims and evaluative conclusions/judgments as they relate to a value statement about an evaluand. This is elaborated further in chapter 6.

1.5 Evaluative reasoning: a neglected topic

This study demonstrates the centrality of evaluative reasoning to the practice of evaluation in the western tradition. House (2004a) describes evaluative reasoning as "the substance of evaluation" (p.219), a view reinforced by Fournier and N. L. Smith (1993) who state "In essence, building a justifiable argument is the crux of evaluation practice" (p.316). Despite its importance, the evaluative reasoning discourse has not been elaborated to the same extent as has occurred in other areas of professional evaluation, such as methods and evaluation use (Alkin, Vo & Christie, 2012). This is demonstrated in a recent study which examined articles published in the American Journal of Evaluation (AJE) and Evaluation, the Journal of the European Evaluation Society (Kallemeyn, Hall, Friche & McReynolds, 2015). The study examined the number of times the terms method, use and valuing were used in 171 papers in these journals in the period 2008-2011. The frequency of the term method (2711) was double that of the term use (1380), while terms relating to valuing were least used (719).

The neglected state of the evaluative reasoning discourse is described by Fournier (1995):

> . . . professionalised evaluation has spent much of its time and effort on developing methodological sophistication, and less so on logical sophistication. Understanding the reasoning process used to establish evaluative conclusions drawn in practice has to be the field's greatest unmet challenge (p.1).

Writing nearly twenty years later, Patton (2012) describes the situation as being unchanged: " . . . valuing is fundamentally about reasoning and critical thinking. Evaluation as a field has become methodologically manic-obsessive. Too many of us, and those who commission us, think that it's all about methods. It's not. It's about reasoning" (p.105).

1.6 Rationale for the research topic

I arrived at the research topic, namely, how evaluative reasoning is understood and practised in the Aotearoa New Zealand public sector, after considerable reading and reflection. As a consultant undertaking evaluation and policy work for government agencies (and a former public servant), my interest in the quality of evaluative work being undertaken in New Zealand led me to initially develop a topic focused on what evaluation quality means in a public sector context. I read literature on topics such as the role of evidence, evaluation standards and specifications, evaluation designs, approaches and methods. I gradually came to understand that one aspect of evaluation quality is sound evaluative reasoning. My argument in this thesis is that evaluative reasoning is a lens through which to consider how to improve the quality of evaluative work undertaken or commissioned by the Aotearoa New Zealand public sector.

Furthermore, I sought to understand that evaluative reasoning is Evaluation Theory. "Evaluation Theory" in this thesis is differentiated from "evaluation theory". The former refers to a prescriptive account of "the logic of evaluative discourse . . . how evaluation should be done in order to be valid" (Scriven, 1991, p.155, p.156). The latter refers to evaluation models, methods and approaches "that are often simply metaphors for, conceptualizations of, or procedural paradigms for evaluation" (Scriven, 1991, p.155, p.156). (The terms evaluation models, methods and approaches are used interchangeably in Aotearoa New Zealand). I aim to demonstrate that there are conceptual and practical reasons for professionals who undertake public sector evaluation to have an in-depth understanding of Evaluation Theory and its application. Consequently, my research topic was revised to consider the practice of evaluative reasoning in the context of public sector evaluation in Aotearoa New Zealand.

1.7 Public sector evaluation

The context for this study is public sector evaluation. In democratic nation states, evaluation is associated with "the interests and needs of governments" (Schwandt, 2009a, p.26). In the Aotearoa New Zealand context, such interests include "allocative and productive efficiency, the effectiveness of government programmes, and the accessibility, cultural sensitivity and responsiveness of public services for customers" (Boston, Martin, Pallot & Walsh, 1996, p.4).

Chelimsky (2006, p.39) has identified four ways in which evaluation serves the interests and needs of the United States government, namely to (i) support government oversight, (ii) build a stronger base for decision making, (iii) help agencies to improve capabilities, as well as greater openness, (iv) strengthen public information about government activities through dissemination of findings. The principles of oversight, accountability, decision-making, transparency, and public sector capability that underpin Chelimsky's evaluation purposes are equally relevant for the Westminster-based government of Aotearoa New Zealand. The origins and role of evaluation in the Aotearoa New Zealand public sector are discussed in section 2.3.

While public sector evaluation primarily concerns public management, the potential impact of such evaluative activities is far-reaching. The consequences of evaluation findings may be substantive for policy or programme recipients who may rely on such services. Similarly, communities may be adversely impacted by a reduction in or withdrawal of government resources. Given the consequential nature of evaluation, evaluative conclusions/judgments must be defensible (Greene, 2011). For evaluation to be viewed by governments and the public as a credible contributor to public policy, evaluative conclusions/ judgments need to be robust (Chelimsky, 1998). Such defensibility and credibility involve sound evaluative reasoning.

1.8 A bounded study

The bounded nature of this study must be emphasised. Firstly as noted above, the study draws on western philosophy to conceptualise evaluative reasoning. Reasoning in an evaluative manner may not hold the same meaning or have value in indigenous cultures as it does in western philosophic thought. However Māori epistemologies feature in the study as a result of the influence of Māori evaluators, some of whom participated in the study or were authors of reports examined in the meta-evaluation.

Secondly, this study focusses on the professional practice of evaluation for the purposes of public administration in the Aotearoa New Zealand public sector. Therefore the study excludes evaluative reasoning approaches that may exist in other practice-based professions such as psychiatry or policing, and in other contexts such as philanthropy and the non-government sector. Lastly, the study focuses only on the reasoning process in public sector evaluation. The role and contribution of evidence, and evaluation models/methods/approaches in public sector evaluation is beyond the scope of the thesis.

1.9 Overview of the thesis

This section details the presentation of my research which reflects its iterative nature and the order in which it was undertaken.

Part A (chapters 1, 2 and 3) situates the thesis. Chapter 1 presents the research questions and objectives, and justifies the relevance of the study. Chapter 2 describes aspects of the Aotearoa New Zealand context relevant to the study and its findings. Chapter 3 introduces the epistemological stance and theoretical framework underpinning the study, and presents the multiple method research design.

Part B (chapters 4 and 5) provides a theoretical foundation for the study by examining the main aspects of the evaluative reasoning discourse. Chapter 4 examines the place of values in the western philosophy of science. Chapter 5 describes the development of the evaluative reasoning discourse in the evaluation literature. The purpose of Part B is to provide a theoretically-based argument for the value of evaluative reasoning in professional evaluation practice undertaken in a western context. In doing so, Part B provides a theoretical scaffold to support the design of the Q methodology study and meta-evaluation (both of which are reported in Part C).

Part C offers three perspectives on evaluative reasoning practice in the Aotearoa New Zealand public sector. The purpose of Part C is to examine the research topic from different perspectives through use of three methods: a Q methodology study (chapter 6), a meta-evaluation of 30 evaluation reports produced or commissioned by 20 government agencies (chapter 7), and key informant interviews with international and New Zealand evaluation experts (chapter 8).

Part D (chapter 9) integrates and interprets the findings from the three perspectives presented in Part C. Insights and hypotheses about the conceptualisation and practice of evaluation reasoning in public sector evaluation and their implications are discussed.

The conclusion (chapter 10) that follows Part D presents contributions to new knowledge about evaluative reasoning practice in a public sector context.

1.10 Conclusion

Chapter 1 has defined the research topic and argued its relevance. The chapter has also established the bounded nature of the study.

CHAPTER 2

AOTEAROA
NEW ZEALAND CONTEXT

2.1 Introduction

Dahler-Larsen (2012), Dahler-Larsen & Schwandt (2012), and Stern (2006) describe how evaluation practice at the country-level is shaped by the societal, political and institutional contexts in which evaluation is being undertaken. It is therefore important to describe aspects of the Aotearoa New Zealand context relevant to this study to inform understanding of the findings. Such contextual information assumes increased significance given that one of the key findings discussed in chapter 9 is the role of contextual factors in shaping evaluative reasoning practice in Aotearoa New Zealand.

Four contextual topics relevant to this study are discussed. The first topic is the Treaty of Waitangi, the founding document of Aotearoa New Zealand, and its implications for public policy and evaluation. The New Zealand public sector is then described, including an account of the influences that have shaped public sector evaluation since its beginnings in the 1980s. A short account of the public sector evaluation community is then presented. Finally, the contribution of New Zealand evaluators to the discourse on indigenous evaluation is discussed. The chapter ends with a short introduction about the researcher.

2.2 Aotearoa New Zealand founding document

This section begins by providing a brief introduction to Aotearoa New Zealand, followed by an explanation about Te Tiriti o Waitangi (the Treaty of Waitangi, hereinafter referred to as the Treaty), its founding document.

Aotearoa New Zealand is a geographically isolated nation in the South Pacific. While it is a substantial country geographically - 270,500 square kilometres, similar to the size of Japan or the British Isles (Statistics New Zealand, n.d.) - Aotearoa New Zealand is a small country demographically with a population of 4.5 million. It is a multicultural society comprising some 200 ethnic groups made up of European (74 percent), Māori (15 percent), Asian (12 percent), Pasifika peoples (7 percent), and a very diverse group characterised as 'other' (1 percent) (Statistics New Zealand).[2] [3] Māori are the indigenous people of Aotearoa New Zealand and the Māori language is one of three official languages, alongside English and Sign language.

Local historian Professor Anne Salmond (2012) describes Aotearoa New Zealand as "a small, intimate society" (p.6). A recent report of the Royal Society of New Zealand (Hawke, Bedford, Kukutai, McKinnon, Olssen & Spoonley, 2014) about the 2013 National Census of population results describes " . . . (the) multiple cultural identities and values" (p.8) of Aotearoa New Zealand. The intimate and relational nature of Aotearoa New Zealand society and its cultural diversity feature in the findings discussed in chapter 9.

Turning now to the Treaty, this document is regarded as "the founding constitutional charter" of Aotearoa New Zealand (Mulgan, 1994 as cited in Boston, Martin, Pallot & Walsh, 1996). In the 175 years since the Treaty was signed it has become more than a historical document, being described by Justice Chilwell (1989) in a High Court judgment as " . . . part of the fabric of New Zealand" (as cited in Te Puni Kōkiri, 2001, p.24) and by M. King (2003) as representing "a living relationship between Māori and the Crown" (p.515). The Treaty was signed in 1840 between some (approximately 500) but not all Māori rangatira (chiefs) and representatives of Queen Victoria. There are two versions of the Treaty, an English version and a Māori version which contain different meanings. The explanation for the differences between the two versions has been the subject of much conjecture and legal challenge over the last 175 years. These differences have never been resolved and are the source of ongoing debate. According to the English version of the Treaty, Māori relinquished to the British Crown the power to govern in New Zealand (sovereignty) described as Kāwanatanga. However for Māori,

[2] The percentages do not add to 100 percent as individuals may identify with more than one ethnicity.

[3] The term Pasifika is used in New Zealand to refer to indigenous peoples from the Pacific Islands living in Aotearoa New Zealand.

Kāwanatanga means something less than absolute authority as is implied by the word sovereignty. In exchange, the Crown promised to protect the chiefly authority of Māori (Tino Rangatiratanga: self determination), including Māori rights to their lands and other possessions (taonga). The Crown also promised to extend to Māori the same rights and privileges as British citizens. Despite its undertakings in the Treaty, over the following 20 years the colonial government confiscated land from Māori and a Legislative Council was set up without Māori representation (Salmond, 2008). These and other breaches of the Treaty culminated in the outbreak of armed conflict in 1860 (referred to as the New Zealand Wars or the Land Wars) between government forces and settlers, and Māori (Salmond, 2008).

It was not until the 1960s-early 1970s that sufficient momentum grew to address the injustices that Māori had suffered and continued to suffer as a result of the loss of their land, language, and mana (authority, control, influence, prestige and power). In 1975 a hīkoi (march) travelled from the far north of the North Island to the capital city Wellington (a distance of about 1000 kilometres) to petition the government to address land grievances. In the same year the Waitangi Tribunal was established by the Treaty of Waitangi Act as an Independent Commission of Inquiry to make recommendations to government on claims relating to actions or omissions of the Crown that may breach the Treaty of Waitangi (Barrett & Connolly-Stone, 1998; Byrnes, 2010). Since the mid-1980s the “constitutional significance, legal status and policy implications of the Treaty of Waitangi have received increasing recognition” (Boston et al., 1996, p.142). Around 72 Treaty settlements have been reached to-date between claimants and the Crown, with many other claims in pre-negotiation, negotiation or draft Deed of Settlement stages (the draft deed contains the details of a settlement initialled by the Crown and the claimant negotiators which is then put to all members of the claimant group for ratification) (Office of Treaty Settlements, 2015, p.21).

Since the mid-1980s successive governments have given recognition to the principles of the Treaty in significant legislation (Boston et al., 1996, p.145). The term “Treaty principles” is now commonly used by government and the public sector rather than reference to specific Treaty clauses. All legislation is required to comply with the Principles of the Treaty (Parliamentary Counsel Office, n.d.). Examining the findings and recommendations of the Waitangi Tribunal, Byrnes (2010, p.6) notes there is “increasing recognition of the Treaty as a developing social contract, rather than a static historical document". Byrnes (2010, p.10) describes the Tribunal's findings as emphasising "a new and ongoing relationship (between Māori and the Crown), where some power was ceded to the Crown (by Māori), but conditional on the Crown fulfilling its Treaty obligations towards Māori". Despite this notion of the Treaty as an ongoing social contract with Māori, there continues to be a lack of clarity by successive governments about Treaty obligations in the social policy area (Barrett & Connolly-Stone, 1998). This lack of clarity, together with the failure of these governments to effectively address social and economic disparities for Māori have contributed to some parts of Māoridom calling for Tino Rangatiratanga (self-determination), particularly relating to the funding and provision of social services for Māori (for example, Wihongi, 2010).

The requirement for Government to meet its Treaty obligations has implications for public policy and its evaluation. Guidelines for research and evaluation with Māori have been developed by a number of government agencies, including Te Puni Kōkiri guidelines for evaluation with Māori (Ministry of Māori Development, 1999); Nga Ara Tohutohu Rangahau Māori: Guidelines for Research and Evaluation with Māori (Ministry of Social Development, 2004); SPEaR Good Practice Guidelines (Social Policy Evaluation and Research Committee, 2008); Te Ara Tika Guidelines for Māori Research Ethics (Health Research Council, 2010). In summary, these guidelines articulate principles and approaches that support culturally-responsive practices: the involvement of Māori stakeholders in evaluation planning and design, the use of culturally appropriate methodologies, analysis of evaluative information about Māori as a distinct group, the protection of Māori knowledge and valuing of Māori expertise, and timely reporting back to Māori stakeholders.

The SPEaR Guidelines (2008) and Te Puni Kōkiri Guidelines (1999) extend culturally responsive evaluation practices by illustrating how evaluations may incorporate Māori values, be controlled by Māori and produce Māori knowledge to a greater or lesser extent. Both guidelines present conceptual frameworks which are based on a continuum - at one end of the continuum are evaluations which are based on public sector values and control (referred to below as the "public sector" end of the continuum), and have no engagement with Māori stakeholders or involvement of Māori evaluators. At the opposite end of the continuum are evaluations which are based on an evaluation approach in which Māori values are at the fore, Māori stakeholders control the design and conduct of the evaluation, and which produces Māori knowledge. This approach is referred to as a Kaupapa Māori approach, discussed further in section 2.5. (Referred to below as the "Kaupapa Māori" end of the continuum). Both frameworks identify two types of evaluation between these opposite ends of the continuum - the first type of evaluation is located towards the public sector end of the continuum in which Māori have some involvement but the agency controls the evaluation and knowledge of interest to the public sector is produced. The second type of evaluation which is located towards the Kaupapa Māori end of the continuum incorporates Māori values and uses a co-production approach with Māori stakeholders. Knowledge of interest to Māori is more likely to be produced.

While the impact of these conceptual frameworks and culturally-responsive practices described above on public sector evaluation has not been systematically examined, the guidelines can be seen as contributing to the shaping of present-day evaluation practice in Aotearoa New Zealand, particularly through their emphasis on stakeholder participation in evaluation, and the valuing of indigenous knowledge and expertise. This will be demonstrated in the Q methodology findings presented in chapter 6.

2.3 Public sector evaluation in Aotearoa New Zealand

The Aotearoa New Zealand public sector is made up of two levels - central government and local government (comprising 78 local authorities). As this study focuses on central government, the following information about the public sector provides context for the findings in chapters 6, 7 and 8.

The Parliament which is located in Wellington, the capital city, consists of one House made up of 120 members. The central government public sector is made up of 28 core departments and 27 Crown Entities, the majority of which have head offices in Wellington where the majority of evaluations are undertaken or commissioned. The core public service had 35,623 full-time equivalent positions as at 30 June 2015 (State Services Commission, 2015). The public sector spends approximately $111M per annum on policy, research, evaluation and development contracts undertaken by external contractors (Ministry of Business, Innovation and Employment, 2016). As will be demonstrated in this study, the small size of the public sector (relative to other Commonwealth countries) and the consequential thin “bureaucratic layers” (Williams, 2003, p.199) affords some features that are not present in countries with larger governmental structures.

Evaluation in Aotearoa New Zealand has its origins in public sector management due to the demand for government agencies to have increased internal and external accountability as expressed in the State Sector Act 1988 and the Public Finance Act 1989 (Lunt & Trotman, 2005). Public sector evaluation has been variously conceptualised as a tool to assist decision making, an accountability mechanism, and the end component of the public policy cycle (Bahler, 2003). This managerial foundation is in contrast to the US where evaluation has strong philosophic traditions as evidenced in the writing of theorists such as Greene (1990, 2011), Schwandt (1997, 2002b), and Scriven (1972, 1976, 1980a, 1995, 2007a), and Europe with theorists such as Abma (2006), Dahler-Larsen (2012), and Leeuw (2003, 2008).

Reforms to public management in New Zealand aimed at improving the effectiveness and efficiency of the public service are ongoing some 30 years after they began in the 1980s (Morrison, 2014). Since the mid to late 1990s, such reforms have been driven by New Public Management (Boston et al., 1996), a term that describes an approach to public management with features such as “a preference for private ownership . . . the contracting out of most publicly funded services . . . accountability for quantifiable output or outcome measures . . . (and) an emphasis on cost-cutting and efficiency” (p.26). Successive programmes of work directed by the New Zealand State Services Commission (the central agency that coordinates the management of the state sector) (SSC) to support these public sector reforms have influenced the Government’s focus on and commitment to public sector evaluation (2002; 2003). This was illustrated in the mid-1990s when government agencies were required to focus on the achievement of the then Government’s Strategic Priority Areas. The SSC promoted the use of evaluation to measure such achievement (Lunt & Trotman, 2005). A review of the state sector in 1996 by Professor Alan Schick, a professor of public policy at the University of Maryland identified the importance of measuring the outcomes of government activities (Morrison, 2014; Ryan, 2003). This was reinforced by a report of the Advisory Group on the Review of the Centre (2001) which identified the need for increased emphasis on outcome specification and evaluation (Lunt, 2003).

The Government's response was the introduction of a work programme titled Managing for Outcomes (2002) which required agencies to adopt an outcome-focussed approach to planning, management, evaluation and reporting (Ryan, 2003). According to Ryan and Gill (2011) Managing for Outcomes was "one of the most important developments in public management in New Zealand" (p.311) given its focus on both efficiency and effectiveness. In 2003 a related government initiative, Improving the Knowledge Base for Social Policy (2003), was introduced aimed at enhancing the production and use of research for policy purposes (Lunt & Trotman, 2005). The Social Policy Evaluation and Research Committee (SPEaR) was established to facilitate best practice research and evaluation for public sector purposes (Lunt, 2003). The period from 2001 to around 2008 may be regarded as the heyday years for public sector evaluation, given the focus on evidence-based policy both in New Zealand and internationally (Nutley, Davies & Walter, 2003). The tightening of public expenditure in the late 2000s in the aftermath of the global financial crisis impacted on the funding available for government funded research and evaluation activities, and SPEaR was disestablished in 2010. In 2011 Government passed legislation to support Better Public Services (2011) which sets out the strategic direction and priorities for public sector agencies for the ten years to 2021 (State Services Commission, n.d.). The strategy aims to clarify the priority areas on which agencies are to focus their efforts, in addition to fostering coordination among agencies (State Services Commission, n.d.). Morrison (2014) notes that the language of outcomes which dominated public sector discourse for the previous ten years has now been replaced by results which "are bite-sized pieces of an outcome, similar to what were previously called intermediate outcomes" (p.47). The State Services Commision (2011) identifies five areas for improved public sector performance, each of which has two or more result targets: "reducing long-term welfare dependence, supporting vulnerable children, boosting skills and employment, reducing crime, and improving interaction with government" (n.p.). There are ten result targets in total, all of which are measured quantitatively via indicators, and reported publically. As there is no literature or other commentary about the impact of Better Public Services on public sector evaluation, the following are personal observations based on my work as a public sector contractor working in both evaluation and policy roles. The Better Public Services issues paper titled Results (State Services Commission, 2011) identifies five characteristics of effective results, one of which is that they are measurable:

“(Results) . . . are tightly specified in terms of scope so they can be captured by one or two indicators that can be used to define and measure performance” (p.9). Working with one of the largest public sector agencies on a number of policy and evaluation projects over the last two years, I have observed a significant reliance on quantitative indicators to measure performance and less interest in qualitative approaches. This agency has recently reduced its five research and evaluation teams to two teams, with a corresponding reduction in the number of staff. The agency’s monitoring and data analysis teams are not affected. The New Zealand situation described here echoes Stern’s (2006) description of public sector reforms in the United Kingdom (UK) and their impact on evaluation: “ . . . performance management has come to displace evaluation in government circles” (p.299). Penny Hawkins, a New Zealander who is now working in the UK made a similar observation in her keynote address to the 2015 conference of the Aotearoa New Zealand Evaluation Association. She describes routinely hearing comments from officials such as: “Monitoring is all we need, evaluation takes too long, costs too much and is not sufficiently responsive to the information needs of policy-makers and programme managers” (2015, p.2).

An unrelated development in Aotearoa New Zealand has brought public sector evaluation under increased scrutiny, namely, the appointment of Professor Sir Peter Gluckman as the Government’s first Chief Science Adviser in 2009. Professor Gluckman is tasked with advising Government on how the public sector might improve its use of evidence in both the formation and evaluation of policy (Office of the Prime Minister’s Science Advisory Committee, 2013). Professor Gluckman’s stance towards the production of evidence for public policy-making is predicated on hypothetico-deductive logic. The task of science (defined as including social science) is to produce “a high degree of objectivity” (2013, p.12). Writing about the role of evidence in public policy formation, Professor Gluckman promotes value-free inquiry: “ . . . where evidence is conflated with values, its power is diminished” (p.4). Given the centrality of values to evaluation as will be described in chapters 4 and 5, Professor Gluckman’s stance sits in sharp contrast to a body of social science in which values are present. The potential impacts of this stance are discussed in chapter 9.

2.4 *Public sector evaluation community*

There is little research-based literature describing the current state of the public sector evaluation community in Aotearoa New Zealand. The most recent account is an edited book published in 2003 (Lunt, Davidson & McKegg, 2003) examining a range of practice-related topics such as commissioning and managing evaluations, evaluation utilisation, and evaluation with Māori and Pacific peoples. Therefore the information in this section is based on discussions with public sector evaluation colleagues (including at workshops where I have presented the findings from this study), and my experiences of working in and for public sector agencies over a twenty-year period.

The majority of the evaluation occurring in Aotearoa New Zealand is government funded, with lesser funding available from other sources such as private and philanthropic organisations. The community of professional evaluators working within public sector agencies (internal evaluators) or evaluators commissioned by agencies (external evaluators) is relatively small, reflecting the size of the public sector (Wehipeihana, Bailey, Davidson, & McKegg, 2014, p.50). There are two evaluation organisations operating in New Zealand - the Aotearoa New Zealand Evaluation Association (ANZEA) established in 2006, and the Australasian Evaluation Society (AES) established in 1986 which covers the Australasian-Pacific region and operates out of Melbourne Australia. There are around 180 members of ANZEA and approximately 160 New Zealand members of AES (some people belong to both organisations). (These figures refer to individual memberships, not corporate memberships).

The formation of ANZEA grew out of a desire of New Zealand evaluators for an organisation to “represent the unique needs, values, obligations and working context of the Aotearoa New Zealand evaluation community” (ANZEA, n.d.). Central to this is ANZEA’s commitment to the Treaty of Waitangi as articulated in the first objective of its Constitution:

> To promote and facilitate the development of evaluation practices and standards which are relevant to Aotearoa New Zealand, with specific reference to the principles and obligations established by Te Tiriti o Waitangi and reflecting the unique bi-cultural context of Aotearoa New Zealand, while also providing a framework from which multi-culturalism can be embraced and responded to (n.p).

Two organisations operate within the ANZEA umbrella - Mā te Rae established in 2015 “by Māori, for Māori to advance the social, cultural and economic development of iwi Māori through participation in and contribution to quality evaluation” (ANZEA, 2015 , n.p.) and a Pasifika Fono (meeting) established in 2014 to support the development of Pasifika evaluation capacity and capability.

The Australasian Evaluation Society’s New Zealand branch operates out of Wellington through the Wellington Evaluation Group (WEG) which was established in 1991 (Trotman, 2003). People who are not AES members can participate in WEG activities. Approximately 300 people currently subscribe to the WEG email list and attend WEG events such as workshops and presentations by overseas and local evaluators.

My perception of the public sector evaluation community (from having been part of this community for over 20 years) is that it is diverse, with people coming to evaluation from areas such as policy analysis, social work, community development, management, market research, and social science research. The limited amount of public sector work available means evaluators have to be generalists, as specialist opportunities do not exist (E. J. Davidson, personal communication, 3 June, 2015). The first (and only) formal evaluation qualification (a post graduate diploma) began at Massey University in 2005. Up until this time formal evaluation education was limited to courses run by some universities as part of health, education or research qualifications. The Australasian Evaluation Society played a significant training role in the late 1990s-early 2000s providing one-week evaluation courses delivered by Australian and New Zealand academics which I attended in 2001. For my part, it was not until I began the post graduate diploma in 2005 that I was challenged to deepen my understanding of evaluation theory.

Given our physical isolation, we rely on overseas experts visiting us or we attend overseas conferences, principally the annual AES Conference in Australia. A group of around 10-15 people travel further afield to the annual conference of the American Evaluation Association (AEA), while a handful attend the bi-annual conference of the European Evaluation Society (EES). In my experience, our reliance on visiting experts makes us susceptible to adopting the practices espoused by the most recent visitor. This is demonstrated by the evaluation methods that have become flavour of the day in New Zealand following a theorist’s visit (predominantly theorists from the US), for example, Michael Quinn Patton's Utilisation-focused Evaluation, Tessie Catsambas' Appreciative Enquiry, Jess Dart's Most Significant Change, David Fetterman's Empowerment Evaluation, and Donna Mertens' Transformative Evaluation. A personal observation is that the small size of the public sector evaluation community, our relative isolation, and the lack of academic courses and challenging professional development opportunities may limit the range of discourses in which we participate.

2.5 Contribution to indigenous evaluation discourse

Chapter 9 describes how the study findings suggest that some aspects of Māori epistemologies are influencing evaluation practice by some non-Māori evaluation practitioners. As an introduction to this discussion, it is helpful to understand the contribution of Māori evaluators to the discourse of indigenous evaluation (Fitzpatrick, 2012), for example, Cram (1997, 2009), Kerr (2012), Moewaka Barnes (2003, 2009), and Wehipeihana (2008, 2013). This section provides an overview of the origin of these developments and their implications for public sector evaluation practice.

In past years most research about Māori was, according to Cram (2009), funded by government "to objectify and problematise Māori" (p.309). There was a lack of regard for Māori aspirations regarding the research, as well as "a lack of researcher accountability to Māori" (Moewaka Barnes, 2003, p.146). The way research was conducted served to reinforce the asymmetric power dynamics between Māori and Pākehā. This approach led to the production of "mainstream knowledge of Māori," rather than "Māori knowledge" (C. W. Cunningham and Durie, 1998, p.1). The former serves the needs of the government as the funder of the research/evaluation, while the latter serves the needs of Māori (Moewaka Barnes, 2003, p.149). In the late 1990s, Māori scholars including Professor Sir Mason Durie (1998), Professor Chris Cunningham (1998), and Professor Linda Tuhiwai Smith (1999) challenged this approach based on dominant western epistemology, and argued for the need for Māori epistemology and Mātauranga Māori (traditional knowledge) to be embraced. Moreover, Māori asserted their rights under the Treaty to conduct research "that is by Māori for Māori, using tools that we see as valid" (Jackson, 1987/1988, cited in Cram, 2001, p.39). Research theory based on Māori ontology and epistemology emerged, referred to as Kaupapa Māori research. Kaupapa Māori means “the Māori way or agenda, a term used to describe traditional Māori ways of doing, being and thinking, encapsulated in a Māori world view or cosmology” and as such “ . . . is both a set of philosophical beliefs and social practices (tikanga) (E. Henry & Pene, 2001, p. 235, 237). Expressed simply, Kaupapa Māori research and evaluation refers to that which is “by Māori, for Māori and with Māori” (Cram, 2009, p.312). L. T. Smith (1999, p.120; 2008, p.130) identified seven research practices based on cultural values to guide the behaviour of Māori researchers/evaluators. These practices have been elaborated further by Cram (2001, p.41, 50; 2006, p.313). The following summary of the seven research/evaluation practices is based on the work of both L. T. Smith and Cram:

> Aroha ki te tangata: This is about respect for research collaborators and participants.
>
> He kanohi kitea: This is about the relationships that are built between the researcher and the research participant and their community. It is about the researcher being known to, and seen around the community.
>
> Titiro, whakarongo . . . kōrero: This is about the researcher looking, listening, and observing in order to develop understanding, before speaking.
>
> Manaaki ki te Tangata: This is about the researcher looking after research participants and their community, which includes reciprocity.
>
> Kai Tupato: This is about the researcher being careful, safe, astute and reflective.
>
> Kaua e Takahia Te Mana o te Tangata: This is about upholding the mana (authority) of research participants and their community.
>
> Kia Mahaki: This is about the researcher being humble and sharing knowledge which will help to empower the community.

More recently, Kerr an evaluator from SHORE/Whāriki at Massey University (Kerr, 2012, p.8,10) reviewed the work of seven Māori theorists (Professor Russell Bishop, Dr Kathy Irwin, Professor Helen Moewaka Barnes, Dr Leonie Pihama, Professor Graham Smith, Professor Linda Tuhiwai Smith, and Dr Sheilagh Walker) to identify five key principles of Kaupapa Māori research and evaluation, as listed below in English. At this point it is important to acknowledge that many significant Māori constructs are not easily translatable into English as often English has no comparable concepts. Even though English language translations are offered here, non-Māori speakers may miss both the nuance and import of the concepts expressed in Māori.

> Control principle: Māori control/ownership.
>
> Challenge principle: Analysis and mediation of power relationships.
>
> Culture principle: Māori as normative including the survival and revival of Māori language and culture.
>
> Connection principle: Relationship-based knowledge, sharing and generation.
>
> Change principle: Transformative for Māori.

Wehipeihana, a Māori evaluator (2008, 2013) has contributed significantly to the indigenous evaluation discourse through her focus on cultural validity, defined by Kirkhart (2010) as " . . . the accuracy or trustworthiness of understandings and judgments, actions, and consequences, across multiple dimensions of cultural diversity" (p.401). Wehipeihana (2008) argues that for an evaluation of a policy or programme with Māori participants (or any other indigenous peoples or minority group) to be a quality evaluation, it must be culturally valid. And in order to be culturally valid, the evaluation must be conducted by Māori evaluators who have " . . . the necessary cultural capital - knowledge of tikanga (customs and practices), knowledge of Te Reo (language), knowledge of iwi/tribal history and contexts - in order to make sense of, and to understand what is being shared" (Wehipeihana, 2008, p. 42). Wehipeihana's identification of the role and importance of cultural capital in the evaluation of indigenous people adds to the discourse about cultural validity (Rogers & Davidson, 2013).

The indigenous evaluation discourse described above challenges the traditional role of public sector evaluation as serving the needs of government and the public management values of efficiency and cost-effectiveness of its policies and programmes. The discourse creates an alternative space in which values that are relevant to Māori shape the design, conduct and reporting of evaluations of public sector initiatives. The extent to which government agencies (other than Te Puni Kōkiri, the Ministry of Māori Development) are working (or prepared to work) in this evaluation space has not yet been examined.

Lastly, there is a small but growing group of New Zealand-based Pasifika evaluators who are developing evaluation approaches based on Pasifika pedagogies, for example Fotuali'i McGeady (2015) and Suaalii-Sauni (2015).

2.6 About the study author

I am a Pākehā New Zealander of parents who emigrated from the UK in the early 1960s in search of new opportunities beyond their working class roots. I worked in operational policy and policy roles in public sector agencies for over 20 years. In the late 1990s, some policy teams began to undertake evaluations as part of the policy function despite (as in my case) having no previous experience or formal training in evaluation. I was in the first cohort of students to complete the Post Graduate Diploma in Social Sector Evaluation Research from Massey University in 2007, the first university evaluation qualification to be offered in Aotearoa New Zealand. I became an evaluation and policy consultant in 2006.

2.7 Conclusion

This chapter has set out contextual information that forms the backdrop for the findings that follow in Part C and their subsequent interpretation in Part D. Such information includes distinctive features of Aotearoa New Zealand as a small, intimate society in which Māori are the indigenous peoples. The relationship between Māori and Pākehā (non-Māori) is one of Treaty partners, and public policy is one of the principal means by which Government meets its Treaty obligations. The chapter has introduced the Aotearoa New Zealand public sector and described how public sector has evolved in response to public sector reforms driven by the New Public Management discourse. The contribution of Māori evaluators such as Cram (2003, 2009) and Wehipeihana (2013) to the indigenous evaluation discourse has been described, building on the earlier work of Māori academics including C. W. Cunningham and Durie (1998) and L. T. Smith (1999). This information provides the context for the findings (Part C) and their interpretation (Part D) which will show how aspects of Māori epistemologies are influencing the way public sector evaluation is conceptualised and practised by some evaluation practitioners.

CHAPTER 3

METHODOLOGY

3.1 Epistemological stance

3.1.1 Epistemological stance

This section describes the epistemological stance that underpins this study. The Oxford Companion to Philosophy (2005) defines epistemology as “the theory of knowledge” (p.258) and ontology as “the science of being, embracing such issues as the nature of existence and the categorical structure of reality” (p.670). The researcher’s beliefs about the nature of reality (their ontological stance) and how it is known (their epistemological stance) determine the inquiry paradigm they deem to be valid, and therefore the inquiry methods they value. Responding to Mertens’ (2007) exhortation for researchers to be explicit about their epistemological position from the outset of any research activity, this is an interpretivist study based in part on abductive logic.

3.1.2 Interpretivist research

In seeking to understand interpretivism, it is helpful to first consider positivism. Chapter four presents a historical explanation of how positivism (defined as “the application of the methods of the natural sciences to the study of social reality” (Bryman, 2008, p.13) has had, and continues to have “an epistemologically privileged status” (Heshusius & Ballard, 1996, p.4) in the philosophy of science. According to positivist epistemology, ways of knowing that are not congruent with intellect and reason - such as personal opinion, belief and emotion - are deemed to be non-rational and subjective, and are therefore biased and unreliable (Heshusius & Ballard, 1996). Interpretivism challenges this dominant rationality by seeking to understand human action and “people’s common-sense thinking” (Bryman, 2008, p.16). Consequently, the interpretivist researcher embraces individual voice and personal idiosyncrasy - characteristics dismissed by the positivist researcher as prejudiced, and therefore invalid. Writing about social research, Letherby, Scott and Williams (2013) state: " . . . subjectivity is not something we can wish or theorise away but is inevitable in all our endeavours and, as such, is something we should - we must - engage with" (p.98). Bryman (2008) notes that interpretivism does not lend itself to a single, definitive definition, but draws on the intellectual traditions of Weber's Verstehen (the German word for understanding), hermeneutics-phenomenology defined as "The theory and method of the interpretation of human action" (Bryman, 2008, p.694), and symbolic interactionism defined as "A theoretical perspective in sociology and psychology that views social interaction as taking place in terms of the meanings actors attach to action and things" (Bryman, 2008, p.699).

The adoption of an interpretivist frame for this study is predicated on my personal beliefs about social research and the work of the social science researcher - the need for research to encompass diverse ways of knowing, the importance of context in meaning-making, the valuing of the lived experience of those who become our research participants, the need to understand and interpret participants' voice carefully and conscientiously, and humility about the researcher's role as a learner rather than detached expert.

3.1.3 Abductive logic

This study is based in part on abductive logic which is now described. The American philosopher and founder of the school of thought known as pragmatism, Peirce (1839-1914) was concerned about the forms of cognitive reasoning that " . . . help humans, in managing their everyday life, to make connections and continue with the tried and tested, or if necessary also to discover something new" (Reichertz, 2014, p. 124). Such forms of cognitive reasoning, referred to as reasoning habits, include "deducing, generalizing, inferring, inducing, sensing, guessing, recognizing and discovering" (p.124). Drawing on every day and scientific experience of inquiry, Peirce examined how new ideas, beliefs and hypotheses are created (Oxford Companion to Philosophy, 2005, p.685). Peirce argued that the structure of inference provided by deductive and inductive logics is unable to generate new ideas (Tavory and Timmermans, 2014). To address this shortcoming, Peirce proposed abduction as an additional reasoning habit, defined as " . . . an intellectual act, a mental leap, that brings together things which one had never associated with one another" (Reichertz, 2014, p.127). The Oxford Companion to Philosophy (2005) provides an expressive definition, describing abduction as

" . . . the logic of discovery: it studies how we are guided in the construction of new hypotheses from the ruins of defeated ones" (p.687). Abductive inference is not only the domain of the social scientist but is "an ordinary human cognitive and creative function" (Wolf, Peace & Brown, 2015, n.p.). It is the means by which " . . . people in their everyday life . . . produce novel generalizations about the world they live in" (Tavory and Timmermans, 2014, p.35).

Peirce's abductive logic challenged the philosophy of science discourse which prevailed in the late 19th century (Tavory & Timmermans, 2014), in particular the work of Popper (1902-1994) whose " . . . considerable reputation rests on his philosophy of science" (Oxford Companion to Philosophy, 2005, p.739). Popper differentiated the context of discovery (associated with non-logical cognitive processes such as conjecture and speculation, and deemed to be outside the realm of science) from the context of justification which was achieved through the deductive testing of hypotheses to determine whether they could be falsified (the realm of science) (Tavory & Timmermans, 2014). Peirce did not accept Popper's assertion of discovery as being non-scientific and sought to "uncover the logic through which new ideas come into existence" (Fann, 1970, as cited in Mirza, Akhtar-Danesh, Noesgaard, Martin & Staples, 2014, p.1982). Further, Peirce rejected Popper's discovery/justification dualism, stating that discovery and justification are combined in scientific inquiry. He proposed abduction as "the process of forming and exploring hypotheses when confronted with a set of unexplainable observations" (Peirce, 1903 as cited in Tavory & Timmermans, 2014, p.136). Moreover, Peirce asserted that the process of abduction is grounded in logic (Peirce, 1905 as cited in Reichertz, 2014, p.125): "It must be remembered that abduction, though it is very little hampered by logical rules, nevertheless is logical inference, asserting its conclusion only problematically or conjecturally . . . but nevertheless having a perfectly definite logical form".

Abduction is best understood through comparison with deduction and induction. Stephenson (1961), the founder of Q methodology (one of the research strategies used in this study), explains the relationship of abduction to deduction and induction as follows.

> Abduction is what one does in guessing or inventing, or proposing a theory or explanation or hypothesis: it is the initial proposition to

> explain facts. Deduction thereupon explicates the initial proposition, deducing the necessary definitions and formal hypotheses for empirical testing. Induction is the empirical establishment of the hypotheses (p.7).

A more detailed explanation of the three logics is provided below (and summarised in Figure 1).

Deductive inquiry begins with an established rule or principle which is tested against one or more cases. The rule or principle is either observed or not observed in the results and a deductive conclusion produced. A deductive inference is certain, providing its premise is true (Shank, 2008). While deductive reasoning extends existing knowledge, it does not produce new knowledge: “Deduction only conveys the old, familiar truth: it does not produce a new one” (Reichertz, 2014, p.128).

Inductive inquiry starts with one or more cases which are observed or examined. Inductive inference provides “descriptive generalisations” (Reichertz, 2014, p.130) about the shared features in the observed results, or more or less probable claims (Reichertz, 2014; Shank, 2008). The greater the number of cases, the greater the credibility of the generalisation (Tavory & Timmermans, 2014). Induction does not produce new theory, but augments what is already known: “Induction does not generate theory . . . it helps substantiate generalisations using repeated or accumulated observations . . . it strengthens or amplifies our notions of the world by broadening the database” (Tavory and Timmermans, 2014, p.14).

Abductive inquiry starts with a surprising or puzzling observation. Abductive inference makes a connection between the surprising observation and other observations or theories identified as having similar features (Tavory & Timmermans, 2014). Peirce (1955, as cited in Watts & Stenner, 2012, p.39) describes abductive inference as “hypothetical inference” by aiming to explain or theorise about the perceived connection. For this reason, abduction is referred to as “reasoning towards meaning” (Shank, 2008, p.2), and “a logic of discovery” (Watts & Stenner, 2012, p.39). Theory plays an important part in abductive inference as potential hypotheses are examined against existing theory (Charmaz, 2008): “This type of reasoning involves imaginative interpretations because the researcher imagines all possible theoretical accounts for the observed data and then forms and checks hypotheses until arriving at the most plausible interpretation of the observed data” (p.158).

Given the hypothetical nature of abductive claims, Tavory and Timmermans (2014) identify three criteria for assessing such claims: “fit (the theoretical claims are supported by the empirical materials), plausibility (the theoretical claims are stronger than competing claims), and relevance (the extent to which the theorizations matter in the broader intellectual community)” (p.131). These criteria are discussed further in chapter 9.

Peirce held that while abduction needs no justification, a hypothesis produced by abductive inquiry must be tested via deductive or inductive logics to determine its veracity (Reichertz, 2014). Accordingly, Reichertz (2014) describes induction, deduction and abduction as a three-stage logic of inquiry (rather than each being a discrete form): “Abduction searches for theories, deduction for predictions, induction for facts . . . abduction is only the first part of an empirical research strategy - research must not under any circumstance restrict itself to the separate forms of reasoning” (p.131).

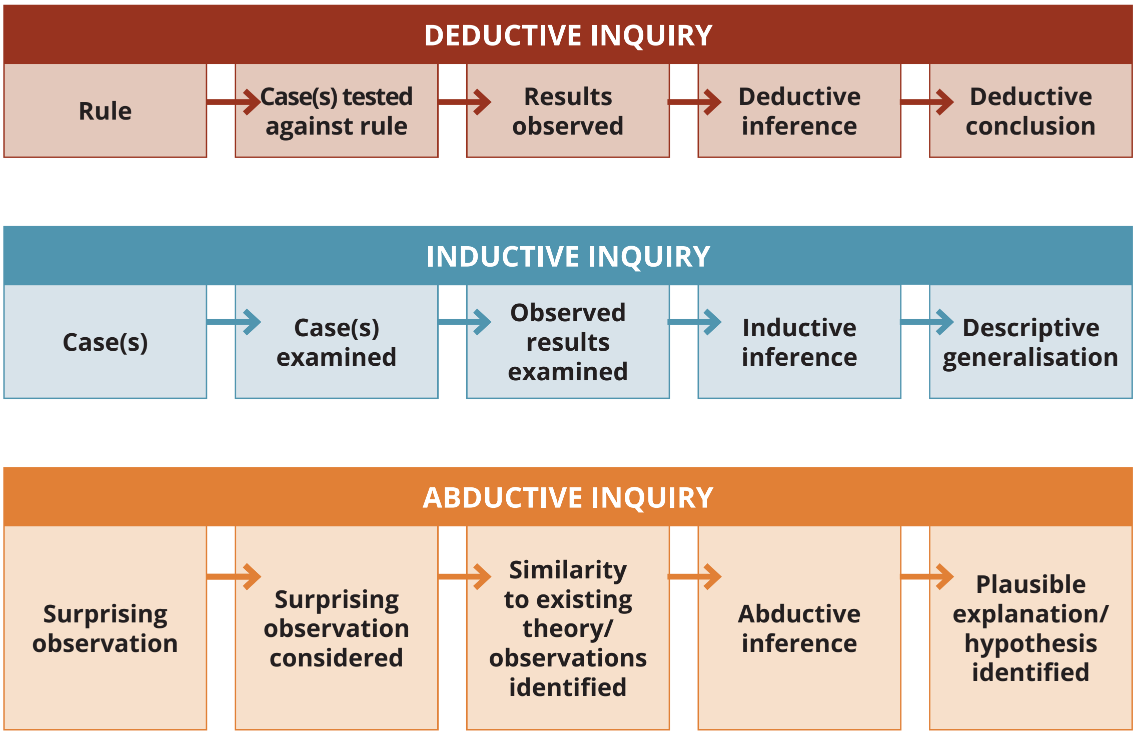

DEDUCTIVE INQUIRY
Rule
Case(s) tested against rule
Results observed
Deductive inference
Deductive conclusion
INDUCTIVE INQUIRY
Case(s)
Case(s) examined
Observed results examined
Inductive inference
Descriptive generalisation
ABDUCTIVE INQUIRY
Surprising observation
Surprising observation considered
Similarity to existing theory/ observations identified
Abductive inference
Plausible explanation/ hypothesis identified

Figure 1 Deductive, inductive and abductive reasoning

(Source: Diagram constructed from exposition in Reichertz, 2014; Shank, 2008; Tavory & Timmermans, 2014).

Abductive logic is not only used for the purposes of qualitative inquiry but is also incorporated into expert systems research and artificial intelligence research (for example, Josephson & Josephson, 1996) for application in a range of technologies such as medical science and computing.

The application of abductive reasoning to this research is now explained. An abductive approach appealed because it provided a means by which to explore my initial hunch about the importance of evaluative reasoning to the quality of evaluation. In addition, I wanted to engage with the research topic in an open-ended, inquisitive manner to maximise the opportunities for unexpected ideas to surface. The design of the research, and the way in which it was undertaken and reported are underpinned by the abductive approach (described further in chapter 3). Abduction also underpins Q methodology used in this study (explained further in chapter 6). An abductive theoretical framework supports the interpretation of the findings from the Q study, meta-evaluation, and expert interviews to answer the third research question, how evaluative reasoning practice can be improved.

3.2 Multiple method research design

3.2.1 Introduction

> We have to remember that what we observe is not nature in itself, but nature exposed to our method of questioning (Heisenberg, 1990, p.26).

Although Heisenberg wrote these words in a text about physics and philosophy, they are equally relevant for social science research. The methods that social scientists employ in their research act as lenses or filters, bringing sharp relief to some aspects and overshadowing others. This study is based on a multiple method research design (described below), enabling different lenses or filters to be applied to the research topic. A multiple method design is appropriate for the abductive approach used in this study. The use of multiple methods may be more likely to reveal surprises, anomalies and puzzles in the data (essential ingredients for abduction) than may be the case in a single method study. This deliberate approach to fostering abductive possibilities is emphasised by Tavory and Timmermans (2014): “Researchers design research to cultivate opportunities for abduction . . . . (they) foster empirical surprises” (p.123).

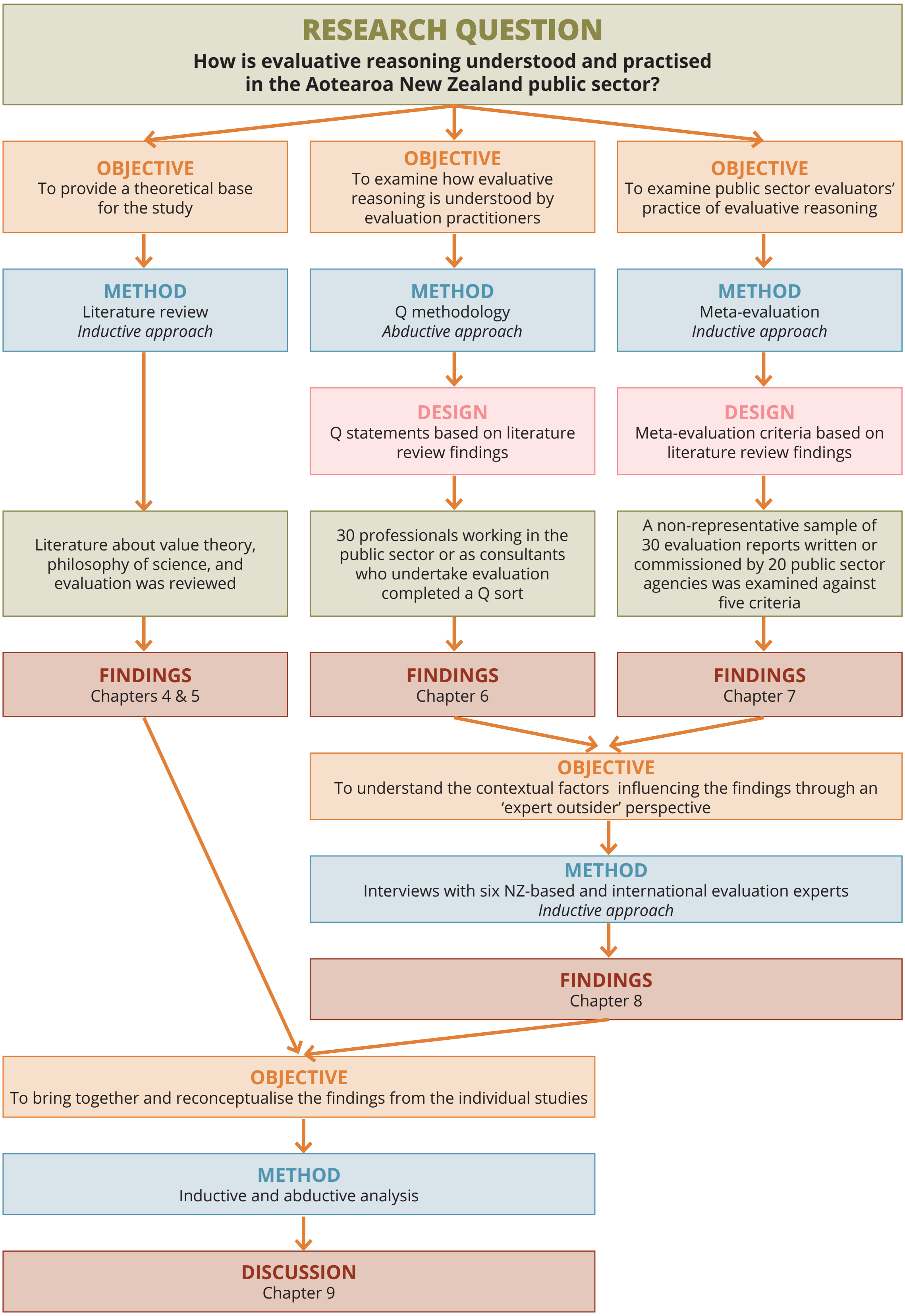
The sentiment expressed by Heisenberg is captured by the concept of triangulation (Denzin, 1978). Triangulation serves to strengthen a study (Patton, 2002a) by examining the research subject "from more than one vantage point" (Schwandt, 2007a, p.298) thereby generating different perspectives. The use of multiple methods is one of four types of triangulation identified by Denzin (1978, p.28). Denzin goes so far as to identify a "methodological principle", namely, that all research studies should use multiple methods (ibid).

3.2.2 Multiple method design

This section describes the multiple method design used in this study. A multiple method design consists of two or more standalone studies using different methods which address the same research question or parts of the same research question (Morse, 2010). The term multiple method is distinguished from the term mixed method, the latter referring to the combining of methods based on the disparate paradigms underpinning quantitative and qualitative approaches (Schwandt, 2007a). Other theorists (for example, Morse, 2010) describe mixed methods design as consisting of one project (referred to as the core project) which is a complete project in itself, and one (or more) supplementary projects that are not complete studies and use a different type of data or analysis which may be from the same or different paradigm as the core project.

An overview of the multiple method design used in this study is now provided (and summarised in Figure 2). A narrative review of literature provides a theoretical foundation for the study (reported in chapters 4 and 5). The first perspective is provided by Q methodology exploring how public sector evaluators understand evaluative reasoning. The Q study provides an abductive viewpoint (reported in chapter 6). The second perspective is provided by a meta-evaluation of 30 evaluation reports conducted or commissioned by 20 public sector agencies. The meta-evaluation, which is based on an inductive approach, examines how evaluative reasoning is practised (reported in chapter 7). The final perspective is provided through local and international evaluators acting as key informants (also an inductive approach) (chapter 8). The Q study, meta-evaluation, and expert interviews are reported as discrete studies, that is, the findings and their interpretation are contained in the relevant chapter. The final chapter (chapter 9) brings together and interprets the findings from all of the methods, using inductive and abductive analytic approaches. The unconventional report structure of the thesis reflects the iterative approach of the research.

The lines and arrows in Figure 2 show how individual methods informed other methods. As will be described in subsequent chapters, the findings from the literature review informed the design of the materials used in the Q study, and the criteria used in the meta-evaluation. The findings from the Q study and meta-evaluation formed the basis of the topics examined in the expert informant interviews.



RESEARCH QUESTION
How is evaluative reasoning understood and practised in the Aotearoa New Zealand public sector?

OBJECTIVE
To provide a theoretical base for the study

METHOD
Literature review
Inductive approach

Literature about value theory, philosophy of science, and evaluation was reviewed

FINDINGS
Chapters 4 & 5

OBJECTIVE
To examine how evaluative reasoning is understood by evaluation practitioners

METHOD
Q methodology
Abductive approach

DESIGN
Q statements based on literature review findings

30 professionals working in the public sector or as consultants who undertake evaluation completed a Q sort

FINDINGS
Chapter 6

OBJECTIVE
To examine public sector evaluators' practice of evaluative reasoning

METHOD
Meta-evaluation
Inductive approach

DESIGN
Meta-evaluation criteria based on literature review findings

A non-representative sample of 30 evaluation reports written or commissioned by 20 public sector agencies was examined against five criteria

FINDINGS
Chapter 7

OBJECTIVE
To understand the contextual factors influencing the findings through an 'expert outsider' perspective

METHOD
Interviews with six NZ-based and international evaluation experts
Inductive approach

FINDINGS
Chapter 8

OBJECTIVE
To bring together and reconceptualise the findings from the individual studies

METHOD
Inductive and abductive analysis

DISCUSSION
Chapter 9

Figure 2 Research design

I now describe what influenced the scoping of the research and choice of methods. During the scoping phase I made a decision to focus on a western philosophic orientation rather than also referencing Māori or other indigenous epistemologies. Some years earlier I had read the New Zealand historian Dr Michael King's book "Being Pakeha now" (2004) and noted the criticism of Professor Sidney Mead, a Māori academic, about Pākehā researchers: "The Pakeha are reaching into Māori culture and pulling out features with which they then try to fit into a Pakeha cultural world" (Mead, 1978 as cited in M. King, 1999, p.182).[4] Being Pākehā, I considered it more feasible and appropriate to undertake research that is located within my western ontological and epistemological understanding.

The research began with a question: how is evaluative reasoning understood and practised by evaluation practitioners in the Aotearoa New Zealand public sector context? (As noted in section 1.2. such practitioners include Māori and Pasifika evaluators who undertake public sector evaluation). Once the question was loosely articulated, the next step was to review the literature on evaluative reasoning to understand the existing discourses about evaluative reasoning in the philosophy and evaluation literatures, and therefore to provide a starting point for the research. A second purpose of the literature review was to piece together a theoretical account to help me gain greater clarity and insight about my initial hunch about evaluative reasoning being an important contributor to evaluation quality. I then searched for a data collection method that would enable me to gain insights into how evaluative reasoning is understood by evaluation practitioners working in, or for, the Aotearoa New Zealand public sector. I was intrigued by colleagues' positive comments about Q methodology as a research tool so investigated its suitability for this study. I began to understand how Q's abductive approach might assist to open up the research topic and provide insights which may not be forthcoming from an inductive method such as qualitative interviews with evaluation practitioners. A brief overview about Q methodology is provided here, while chapter 6 discusses Q methodology theory and practice in greater detail. In summary, Q methodology provides a means by which individuals' perspectives (self-referent attitudes and beliefs) on a topic of interest are sorted (using statistical methods) to reveal patterns (expressed as statistically significant factors) which the researcher then interprets to reveal points of view or

[4] Macrons in Māori terms are now commonly used in written text. However in early published work, macrons were often omitted.

perspectives about the topic of interest (McKeown & Thomas, 2013; Watts, 2011; Watts & Stenner, 2012). The focus of Q is not on the individuals, but rather on the patterns of perspectives and viewpoints that emerge (McKeown & Thomas, 2013).

The second part of the research topic, the practice of evaluative reasoning, is addressed by a meta-evaluation of published evaluation reports. The method of meta-evaluation is described in chapter 7. The rationale for examining evaluation reports as a way of understanding evaluator practice relies on the acceptance of an evaluation report being the end product of an evaluator's practice in respect of the evaluation of a particular evaluand. The report is the physical manifestation of the evaluator's response to their perception of what was required in relation to conceptualising, decision-making and deliberating towards an evaluative judgment in a particular context. In a public sector setting, the reports are conditioned by the political context in which they are produced. Viewed in this way, an evaluation report is an artefact of situated evaluation practice. It is therefore appropriate to examine evaluation reports as a way of understanding evaluators' practice of evaluative reasoning.

The final method, expert interviews, arose from a concern that I was approaching my research topic as an insider, that is, I am embedded in the professional evaluation community that is the focus of my research. The method of qualitative interview is described in chapter 8. I realised I needed outsider perspectives to expand and deepen my interpretive stance. Some overseas theorists have visited Aotearoa New Zealand in recent years to address evaluation conferences, undertake contracts for government agencies, and/or run workshops for local evaluators. I decided that interviewing a small number of these theorists and some New Zealand-based theorists with overseas experience would provide the outsider perspective I was seeking. The findings from the Q study and meta-evaluation were used as the basis of the interview, with the experts being asked for their response to, and interpretation of a summary of these findings.

Having completed and written up the Q study, meta-evaluation and expert interviews, my task was then to bring together and make meaning of the findings from these methods. Fortuitously, in late 2014 Tavory and Timmermans, American sociology academics, published a new text *Abductive Analysis: Theorizing Qualitative Research* which was recommended on the Q Methodology Network (an email group operated by Professor Steven Brown at Kent State University). Drawing on

Peirce’s theory of abductive reasoning, Tavory and Timmermans (2014) provide guidance for analysing qualitative data abductively for the purposes of theory generation: “Abduction is a systematic process of fitting unexpected or unusual findings into an interpretive framework . . . This systematic process of meaning-making aimed at theoretical generalizations is what we have termed abductive analysis” (p.123). Accordingly, I used Tavory and Timmermans’ text to guide meaning-making of the findings from the different methods. Abductive analysis is described further in chapter 9.

3.2.3 Implementing the research design

This section describes how the research design was implemented. The research design provided a useful way to undertake the study as each of the four methods provided a discrete phase of work that directed my focus and made it possible to systematically manage the research process.

I began with the literature review. The review of the evaluation literature on evaluative reasoning led me into literature that described and discussed tenets of western philosophy relating to reasoning. This was quite daunting for someone with no previous exposure to philosophy. I discovered theorists relevant to my topic who are very accessible for the non-philosopher (for example, Rescher, 1969; and Taylor, 1961) and others who are less so (for example, Hare, 1967). I found it was easy to become lost in the philosophy literature, and so for that reason I describe this study as being informed by the literature on informal logic and the philosophy of science, rather than being philosophically-based. I was excited to discover a link between the early work of Scriven (1967, 1980a, 1991, 1993) on the logic of evaluation and the work of philosophers Hare (1967) and Taylor (1961). I was fortunate to have a short conversation with Professor Scriven at the 2011 conference of the American Evaluation Association. My reading had led me to assume that advances in value theory had made it possible for reasoning involving values to become valid. In response to my question about how this had occurred, Professor Scriven advised that the emergence of evaluative reasoning was due not to advances in value theory but rather to the development of informal logic, and referred to the work of Hare and Taylor with whom he had had working relationships. The fact that evaluative reasoning is deeply grounded in philosophy will not be a surprise to evaluators with a philosophy background, but for me (and I suspect other evaluation practitioners who have come to evaluation from other disciplinary backgrounds) it was a moment of revelation with significant import. What distinguishes evaluation from other forms of systematic enquiry is evaluative reasoning. In this sense, evaluative reasoning constitutes the Theory of evaluation. According to Scriven (2013a), this theoretical foundation is what makes evaluation a transdisciplinary or an alpha discipline. For me, this realisation gave more significance to my research. Over the course of my study I continued to read new literature and added to the initial draft of the literature review. The publication of volume 133 of the New Directions for Evaluation journal (2012) titled *Promoting valuation in the public interest: Informing policies for judging value in evaluation* added considerably to the discourse, particularly with its emphasis on the role of context in valuing (discussed in chapter 5).

I chose the Q methodology as my second method because, as explained above, I anticipated that this would open up the research topic for me. Completing the literature review prior to the Q study helped me to recognise and apply the key themes from the literature to construct the sample of statements used in the Q sort. As soon as I had completed the Q study, I wrote up the results and initial interpretations. Over the following months (with the benefit of time and distance), I completed ongoing revisions of this draft as my focus moved from meaning and sense making, to critical examination (Alvesson and Skoldberg, 2009, p.273, 277).

The third phase of the design involved a meta-evaluation of evaluation reports. While I had approached the Q study with a sense of excitement (and some trepidation having not used Q previously), I initially felt less enthusiastic about the more prosaic task of examining 30 evaluation reports. It was therefore a surprise to discover that the task was more engaging than anticipated because of what it revealed about my research topic. As was the case with the Q study, I wrote up the first draft of the findings and my initial interpretations as soon I completed the meta-evaluation, and continued to reflect and develop the findings as my focus turned from sense-making to critical interpretation (Alvesson and Skoldberg, 2009).

Finally, the interviews with the evaluation experts about the Q and meta-evaluation findings provided both greater insight into the findings and some valuable provocations. The observations and comments of experts who are working (or who have worked) in New Zealand about local evaluation practice confirmed many of the key findings from the Q study and meta-evaluation.

Having completed the four methods, the challenge was then to interpret and transform the findings, using inductive and abductive analysis, into a coherent discussion of the research topic. This final stage of the research was at the same time stimulating (in respect of the similar themes revealed in the three methods), and perplexing (in respect of the variability and differences in some of the findings). Tavory and Timmermans' (2014) comment about anomalous findings was therefore reassuring: " . . . to facilitate abduction we need to find ourselves in situations in which we are puzzled by our observations" (p.123).

3.2.4 Testing findings and refining interpretations

Writing about abductive analysis, Tavory and Timmermans (2014) describe the importance of involving peers and other researchers, referred to as "the community of inquiry" (p.103), in the testing of research findings and refining of their interpretation (described further in chapter 9). This section summarises three ways in which I exposed my emerging thinking to my evaluation peers and others over the course of the research.

My two supervisors provided the primary means for me to talk about and reflect on my interpretation of the findings. Their challenging feedback and questioning often caused me to refine or reshape my ideas, or set me off on new directions of thought. A second important community of inquiry for my thinking was a group of five experienced evaluation practitioners to which I belong. We have met (for a half to whole day, usually four times a year) over a period of eight years for professional development and support purposes. Over the course of my research, these colleagues provided (individually and collectively) a listening ear for conversations about the findings, as well as offering critique and insights.

I also tested my evolving thinking at a number of fora over a four year period: the Wellington Evaluation Group (2 seminars); at annual conferences of the Aotearoa New Zealand Evaluation Association (3 workshops) and the Australasian Evaluation Society (3 workshops). After writing up the draft findings of the Q study, I invited the Q participants for feedback on the findings to test my interpretations. Three of the Q participants agreed to participate and we met to discuss the draft findings.

Taken together, these approaches exposed my evolving thinking which helped to shape, confirm and challenge my ideas. More subtlety, the processes of verbal articulation and engagement with others helped me to gain some distance from the data and therefore reflect on it in a fresh light.

3.2.5 Assessment of research design

The multiple method research design provided different perspectives and interpretive frames to examine and understand the research topic. Rich data were provided from the combination of theory (literature review), practitioner viewpoints about evaluative reasoning and its practice (Q methodology), examination of examples of evaluative reasoning practice (meta-evaluation), and evaluation experts' responses to the Q and meta-evaluation findings (expert interviews). Some of the data derived from these methods were confirmatory while others were anomalous, providing opportunities for abductive conjecture. Starting with the review of literature provided a strong theoretical foundation for the study, which is critical for abductive analysis (discussed further in chapter 9) (Tavory and Timmermans, 2014). The benefits and limitations of the individual methods used (in chronological order), and mitigation strategies are summarised in Table 2.1. A more detailed discussion about the limitations of each method is provided in the relevant chapter.

Table 2.1 Benefits and limitations of the individual methods, and mitigation strategies

Method	Benefits	Limitations	Mitigation strategies
Literature review	Articulated the philosophic origins of evaluative reasoning, thereby providing a sound theoretical foundation for the study. At a practical level, data collected for the literature review were used to inform the Q set (used in the Q study), and the criteria used in the meta-evaluation.	The evaluation literature on evaluative reasoning/ valuing is modest compared to other areas of evaluation literature such as methods and evaluation use.	Selective reading of the philosophy literature about reasoning.
Q methodology	The abductive approach helped to open up the research topic by revealing insights that may not have been forthcoming from an inductive approach.	As this was the first time I had used Q, I learnt a great deal about its use which will be beneficial for use in the future. On reflection it would have been helpful to have had previous experience using Q.	Reading of other Q studies, subscribing to and reading emails posted on the Q methodology email network, and ongoing email conversations with Dr Amanda Wolf, a New Zealand-based Q methodology expert.

Method	Benefits	Limitations	Mitigation strategies
Meta-evaluation	The collection of publically available evaluation reports revealed wide-ranging authors and evaluands, from which a purposive sample could be selected.	Limited to evaluation reports that are in the public domain. Limited to a desk-based examination of the reports. The purposive sampling approach provides a snapshot of practice. It does not allow for generalisations.	None identified. Approaching individual agencies to request evaluation reports that were not in the public domain was not considered feasible given the level of Government scrutiny on, and restructuring of evaluation teams during the time the meta-evaluation was being planned.
Expert interviews	The insightful observations and comments provided by the experts on evaluation practice in Aotearoa New Zealand and influences on it.	Since the experts are busy professionals, the duration of the interview was between 15-45 minutes which limited the depth of questioning.	The expert interviews were conducted after I had analysed and written up the findings from the Q study and meta-evaluation. This meant I was clearer about what ideas I wanted to explore with the experts.

Method	Benefits	Limitations	Mitigation strategies
Abductive analysis	A theoretically grounded approach which offers opportunities for creative and generative thinking.	Creates opportunities for methodological critique from readers who are unfamiliar with this research approach.	Systematic explanation of steps and process. Reference to recent authoritative texts.

3.3 Limitations

The evaluative reasoning discourse on which this study is based is derived from ideas put forward by informal logicians (notably Hare, 1967; Rescher, 1969; Taylor, 1961) and developed further by evaluation theorists, notably Scriven (1967). A limitation of this study is that it excludes any indigenous and other non-western epistemological framing. However, where appropriate, I have raised issues relevant to mātauranga Māori (knowledge) as suggested areas for further study.

The research focus is on the professional practice of evaluation in a public sector context. The research findings are context-specific and localised, and therefore may have limited relevance outside the public sector and beyond Aotearoa New Zealand. The research findings are also time-bound, capturing evaluation practice during the period 2011-2015. However, my research aims to highlight principles of practice in relation to evaluative reasoning and these will have wider relevance.

3.4 Research ethics

The Massey University Code of Ethical Conduct for Research, Teaching and Evaluation involving human participants was examined during the scoping phase. The researcher’s obligations and the rights of research participants were specifically noted. The proposed study appeared to meet the low risk criterion for research/evaluation involving human participants. A screening questionnaire and application was submitted to the Massey University Human Ethics Committee on 25 October 2011. The research was confirmed as being low risk and recorded on the Low Risk Database (letter from Professor J.G. O'Neill, Chair, Human Ethics Chairs' Committee and Director Research Ethics, Massey University, dated 28 November 2011). Table 2.2 shows how the principles of ethical research were applied in the study.

Table 2.2 Application of research ethics principles

Research ethics principles	**Application**
Informed and voluntary consent	Participant information sheets and consent forms comply with the compulsory statements and formats specified by the Massey University Research Office. **Q study** Participants in the Q study responded voluntarily to a generic email invitation (via the WEG email group) or a personalised email invitation. The email contained information about the study, and details about what participation in the study would involve for the participant. People who expressed an interest in participating were then sent an information sheet and invited to identify a time and place convenient to them to participate in the data collection. Another copy of the information sheet was provided at the beginning of the Q data collection, and participants were invited to sign a consent form. **Expert interviews** The experts were emailed individually and invited to participate in an interview. Those who responded were sent a two-page summary of the Q and meta-evaluation findings, and the two questions they were being asked to respond to in the interview. Verbal consent was sought at the beginning of the interview. Permission for the interview to be digitally recorded was also sought.

Research ethics principles	Application
Respect for participants	**Q study** The Q data collection was conducted at a time convenient to the participant (usually during their lunch break) at their workplace. The participants were emailed their results and a paper summarising the three orientations. As a number of participants had expressed an interest in learning more about Q, I also emailed them a two-page overview about Q with details of web links and references if they wanted to learn more. **Expert interviews** The expert interviews were conducted at a time convenient to the individual. The experts were offered a copy of their transcript.
Data confidentiality	Interview notes and other personal data were given numeric identifiers. When participants' quotations were included in the thesis, care was taken to ensure there was nothing in the text that could identify the participant or their workplace. The evaluation reports used in the meta-evaluation are not deemed to be confidential as all are in the public domain.
Data security	Electronic data were stored on a password protected computer. Hard copies were stored in a locked filing cabinet.

3.5 Conclusion

Chapter 3 has described the epistemological stance underpinning this study which informed the choice of a multiple method research design. In particular, the design allows the research topic to be viewed from differing perspectives, thereby enhancing opportunities for abductive insights. The research design is presented, noting that the individual methods are described more fully at the beginning of the relevant chapter. The unconventional unfolding and reporting structure of the thesis reflects the iterative approach used for the study. The limitations of the study are explained, the most significant being that the study is based on western philosophic thought and therefore excludes indigenous epistemologies.

PART B

EVALUATIVE REASONING - THE THEORY

Part B provides the theoretical platform for the argument that is presented in this thesis by examining the origins and key aspects of the evaluative reasoning discourse. The assertion that evaluative reasoning is fundamental to the practice of evaluation as a professional activity, and that evaluative reasoning comprises Evaluation Theory underpins this study. It is therefore appropriate to begin this study by examining the theoretical foundations of evaluative reasoning through a narrative literature review.

Chapter 4 situates evaluative reasoning within a broader discourse about the place of values in western knowledge and science. The cases for and against values as legitimate knowledge are examined. Chapter 5 presents the evaluative reasoning discourse as it appears in informal logic and published evaluation literature.
The chapter ends with the presentation of diagrammatic representations of the evaluative reasoning literature.

CHAPTER 4

VALUES IN THE PHILOSOPHY OF SCIENCE

4.1 Method: review of literature

Four literatures were examined for this study, namely, value theory, the philosophy of science, informal logic, and evaluation. The aim of the review and approach taken towards the evaluation literature differed to that taken towards the first three literatures (see Table 4.1). I approached the first three literatures in an exploratory manner with the aim of understanding the epistemological and historical context of values. In respect of the evaluation literature, my aim was to critically examine the evaluative reasoning discourse which required a more comprehensive approach.

Table 4.1 Review approach towards the literatures

	Value theory, informal logic, philosophy of science literatures	Evaluation literature
Aim	To understand the place of values in western philosophic thought. To create an epistemological and historical context for the study.	To critically examine the evaluative reasoning discourse as presented in the evaluation literature.
Approach	Exploratory.	Critical examination.
Scope	Read widely rather than in-depth.	Comprehensive.
Intended output	A historical narrative.	A critical discussion.
Presentation of findings	Chapter 4	Chapters 5 and 7
Search terms used	Values, value judgments, fact-value distinction, evaluative reasoning/thinking, objectivity, subjectivity, inference.	Values, valuing, evaluation theory, evaluation logic, working logic, evaluation judgments, inference, claim, argument/ation, warrant, backing, criteria, standards

4.1.1 Review of value theory, philosophy of science, and informal logic literatures

I started this study by reviewing the value theory and philosophy of science literatures. Since I have no background in philosophy, my initial searching focused on understanding key concepts that appeared relevant to my topic before I pursued more in-depth investigation into the literature. The Oxford Companion to Philosophy (2005) provided a useful starting point for developing an initial understanding of key concepts and identifying relevant theorists. Since the size and scope of the value theory and philosophy of science literatures are very substantial, I tried to identify key theories, concepts and ideas. I sought to stay close to my search terms using a snowballing technique to identify authors in bibliographies who appeared to be relevant, accessing works cited by authors, and by reviewing abstracts. I then accessed books and journal articles written by these authors. *The Journal of Value Inquiry* was very useful in helping me to identify theorists relevant to my topic. As I read, I wrote summary notes about the writing of key theorists, describing their work and contribution, identifying themes, areas of contention, questions and areas to explore further. This was an iterative process as I developed understanding about the place of values in the philosophy of science. As a result of my reading I had assumed that advances in value theory had enabled the emergence of reasoning about values. It was not until I had a conversation with Professor Michael Scriven at the 2011 AEA Conference that I discovered that advances in informal logic (not value theory) had enabled this to occur.[5] Professor Scriven recommended some informal logicians to read (as described in section 4.4.2), some of whom were (or had been) colleagues. It was at this point that I added informal logic to my review of literature.

4.1.2 Review of evaluation literature

The evaluation literature was more straightforward to examine given it is significantly smaller than the above literatures and is more recent. Databases such as ERIC and Web of Science were used to search for the search words (see Table 4.1). Once I established who the key authors were, I then undertook further searches by author. I followed up authors cited by key authors. All of the

[5] Professor Michael Scriven, a high profile evaluation theorist, was trained as a philosopher. He completed a masters thesis on mathematical logic, followed by a doctorate on the logic of science at Oxford University. He gained a position at the University of Minnesota where he worked in a research group led by Herbert Feigl studying the philosophy of science. Scriven's work on critical thinking led him into educational evaluation

international evaluation journals were searched. In addition, a number of theses were accessed. I also searched a range of websites such as national evaluation associations, relevant government websites (e.g. US Aid, US Government Accountability Office, UK Government), international organisations (e.g. World Bank, OECD, UNDP, UNICEF) and the websites of evaluation initiatives (e.g. International Initiative for Impact Evaluation and the International Organisation for Cooperation in Evaluation), and resource websites such as Better Evaluation. I also monitored EVALTALK (AEA's email discussion network) for discussion on evaluative reasoning topics. The importance of my research topic was reinforced to me, given the small size of the evaluative reasoning literature and its less developed state compared to other evaluation topics. I also gained new insights, firstly the benefit of evaluative reasoning theory being more accessible to evaluators, and secondly the potential benefits of bridges between evaluative reasoning theory and practice. I continued to access and read new literature over the course of the study and add insights from this reading to chapter 5. As my understanding of the topic deepened, I re-read some of the literature described in chapter 5 a number of times over a period of three years, particularly authors such as House, Fournier, Schwandt, and Scriven. Doing so provided new insights that had not been evident to me in earlier readings.

4.2 Introduction: about values

This chapter provides an overview about values in the development of the philosophy of science. It considers the case for and against values as legitimate knowledge. The examination is not philosophically driven. Rather it is informed by ideas drawn from philosophy, specifically from logic and the philosophy of science (Fournier, 1995, p.1). This chapter provides the theoretical foundation for chapter 5 about evaluative reasoning.

In considering the cases for and against values in science, it is necessary to first define the word value which is both a noun and a verb. Writing about value theory, Magendanz (2003) provides the following definitions. Used as a noun, "value is the worth of a thing . . . values are mental pictures of idealised states of affairs or models of idealised behaviour considered by the person to be of worth" (p.443). Used as a verb, "valuing is the judging of the worth of something, against some agreed or assumed standard, criterion or measure" (ibid). Values have meanings that are highly contextual and contingent (Moore, 2004; Rescher, 1969). Value meanings are influenced by temporal, social and cultural dimensions (Magendanz, 2003). The diversity of values is reflected in Rescher's (1969, p.15) classification of western value types (a type is based on the group of objects to which the value is applied) (Table 4.2). The purpose of presenting Rescher's value classification is to illustrate the all-pervasive nature of values in everyday life. This observation will gain greater significance when the exclusion of values from knowledge and science is discussed below.

Table 4.2 Classification of western values

Name of value type	Explanation of what is at issue	Example of value
Thing values	Desirable features of inert things or of animals	Purity (in precious stones), speed (in cars or horses)
Environmental values	Desirable features of arrangements in the (non-human) sector of the environment	Beauty (of landscape or urban design)
Individual or personal values	Desirable features of an individual person (character traits, abilities and talents)	Bravery Intelligence
Group values	Desirable features of relationships between an individual and his group (in family, profession etc)	Respect, mutual trust
Societal values	Desirable features of arrangements in the society	Economic justice, equality before the law

(Source: Rescher 1969)

As the term e**valu**ation suggests, values are at the heart of evaluation. Evaluation as a professional practice involves assessing value of some sort, for example, the effectiveness of a product, the responsiveness of a service, the usability of a process, or the success of a policy.

4.3 *The case against values*

This section examines the case against values by discussing the argument against values in science, and the impossibility for formal logic of inferring from facts to values. Both challenge the epistemological legitimacy of values and valuing (evaluation). While these are presented here as separate topics, they are connected through the interrelationship between formal logic and the philosophy of science.

4.3.1 *The case against values in science*

The case against values is derived from the philosophy of science, defined as "the attempt to understand the meaning, method, and logic structure of science by means of a logical and methodological analysis of the aims, methods, criteria, concepts, laws and theories of science" (Klemke, Hollinger & Kline, 1980, p.2). The philosophy of science is a collection of diverse schools of thought that explicate concepts, structure and methods of science through a particular philosophical stance. Examples of such schools include empiricism (and post empiricism), positivism (and its variants) and non-foundational schools such as pragmatism. It is not the intention to examine the numerous schools of thought that collectively contribute to the philosophy of science. Rather, this section provides an overview of the themes emerging from these literatures about the place of values in science. It is acknowledged that such an overview risks over-simplification of complex constructs. Another risk is that it may generalise significant differences among the schools, implying alignments that may not exist.

The case against values in science is based on a fundamental ontological distinction between those things that are deemed to be real, and those that are deemed to be not real. This ontological perspective created what is referred to as the fact-value distinction (Rescher 1969, Scriven, 1980b), the origins of which can be traced back to the ancient Greeks. Plato (c.428-347 BC) distinguished between the physical world as reality, and the metaphysical world of essences which included values (Frondizi, 1971). Plato identified an epistemology consisting of the "realm of knowledge" (that is, the intelligible world), and a second "realm of opinion" (which included conviction, belief and faith) where man is without knowledge (Oldroyd, 1986, p.10, 11). The influence of Plato and other ancient philosophers on western thinking has been highly influential. According to the Oxford Companion to Philosophy (2005) " . . .

(they) created and laid much of the groundwork for later philosophical debate in the fields of ontology and epistemology" (p.34).

The ontological and epistemological distinction between the physical world as reality (representing *true* knowledge), and an *inferior* non-physical world (outside of knowledge) described by the ancient Greeks was developed further by the early empirical thinkers. Bacon (1561–1626) declared that truth could only come from observation and experiment via sense experiences (M. J. Smith, 1998). Descartes (1596–1650) established the notion of dualism, the influences of which remain today as evidenced in the differentiations between mind and body, and subjectivity and objectivity. Descartes described the role of "rational and objective method" in creating knowledge (Hollinger, 1994, p.23). According to Descartes, "value claims cannot be so proven (using such methods), therefore value judgments do not constitute knowledge or belong in the realm of science" (Hollinger, 1994, p.23). Descartes also claimed that "only the quantifiable data of physical reality are objective" and asserted the need for methodological objectivity on the part of those seeking the truth (Hollinger, 1994, p.24).

The notion of values being outside scientific enquiry was developed further by the early British empiricists, notably Locke (1632-1704) and Hume (1711-1776). There was no place for values and other metaphysical concepts in their system of knowledge. Hume's dismissal of the metaphysical is dramatic (Hume, 1748: Sec XII, Part III cited in Hughes and Sharrock, 1990, p.28):

> If we take in our hand any volume of divinity or school metaphysics . . . let us ask, does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning containing matter of fact and existence? No. Commit it then to the flames for it can contain nothing but sophistry and illusion.

The inferior position accorded to values is illustrated in the structure of knowledge developed by the French positivist Comte (1798-1857), portraying pre and post Enlightenment thinking. Comte described scientific knowledge based on rational logic and reasoning as a higher level of thinking (post Enlightenment), metaphysical knowledge based on philosophy was a lower level (the Enlightenment), while "fictitious" knowledge based on faith and custom (pre Enlightenment) represented a lesser level of knowledge (M. J. Smith, 1998, p.79).

From this point on and throughout the twentieth century, a number of schools of positivism emerged including logical positivism and legal positivism (Oxford Companion to Philosophy, 2005). Theorists have identified the shared attributes of some (or all) of these schools. Kolakowski (1966, p.16,17) identifies four "rules" of positivist thought, three of which act so as to preclude values as legitimate knowledge. The first rule, phenomenalism, refers to the assumption that only phenomena that are able to be observed through the senses can be regarded as real. Putnam (2002) notes that the notion of fact as "a complex of sense qualities" (p.40) was replaced in the 20th century by the concept of verifiability. The second rule, nominalism, builds on the first rule by asserting that concepts or abstract entities have no existence or use other than as words or names (Kolakowski, 1966). The third rule is a consequence of these two rules, namely, that value judgments and normative statements do not constitute knowledge. The strength of positivists' objections to values is evident in O'Hear's (1989) comment about their "thorough-going hostility to unobservable or theoretical entities'" (p.109). These characteristics of positivism gave rise to the value-free doctrine which dominated science during the twentieth century. Its influence is evident in Scriven's (2004) description of the value-free doctrine as "inhibiting the emergence of the discipline of evaluation by some 50 years" (p.185).

The emergence of pragmatism in the United States in the latter part of the 19th century provided a sharp contrast to the dominant positivist-based discourses in Europe. The classical pragmatists, Peirce (1839-1914), James (1842-1910), and Dewey (1859-1952) rejected the fact-value dualism on ontological and epistemological grounds. For these theorists, values are part of everyday experience (Putnam, 2002). Putnam (2002) recalls reading a letter written by Dewey to James describing the all-pervasive nature of values " . . . far from being just one special corner of experience, value is something that has to do with all of experience" (Dewey, 1904, cited in Putnam, 2002, p.135). Moreover, these pragmatists asserted that "normative judgments are essential to the practice of science itself" (Putnam, 2002, p.30). This is evidenced in the scientist's selection of a particular method, as such choice expresses preference toward one or more values associated with the method.

In more recent times, the work of Campbell (1916-1996), the high-profile methodologist, reinforced the fact-value distinction. Campbell (1982) described himself as antipositivist, rejecting positivism's claims of truth and infallibility. Whereas positivists pursue "value nihilism" (p.333), Campbell acknowledged the role of values in describing facts: "We are such pervasive valuers that almost none of the facts of the world can be apprised without valuational connotations" (p.334). Despite this concession, Campbell maintained the importance of distinguishing facts from values. His argument was based on a perceived weakness in social science practice which allowed "belief assertions (to be) made in the name of science" (p.335) resulting in biased science. Campbell asserted that "belief manipulation" (p.334) behaviour is caused by "individual-competitive and clique-competitive motives" (ibid) within the research community. Campbell asserted that maintaining the fact-value distinction would highlight and therefore minimise value-based distortions, reducing the risk of scientific bias. In doing so, however, he overlooked, or chose not to acknowledge the role of values in the natural, biophysical and medical sciences, for example, judgments about " . . . research designs, . . . instruments , research quality" (Scriven, 1991, p.374), all of which involve values about merit or worth.

The fact-value distinction implicit in empiricism and positivism influenced the development of the theories of objectivity and subjectivity. For positivists and empiricists, there is a "neutral observation language" (Hughes and Sharrock, 1990, p.133) that underpins scientific method. Only statements that are "verifiable either by empirical observation or logical deduction" are scientific and are therefore objective (p.44). Statements that cannot be verified in this way are deemed to be "about personal taste or preference" (p.44) and are subjective. Further, dimensions such as beliefs, attitudes and values are deemed to be subjective because they exist in the individual and therefore lack the generalisability and impartiality required by true knowledge (M. J. Smith, 1998). As will be shown in section 4.4.1, this dualism between fact as objectivity, and values as subjectivity was challenged by philosophers in the late 20th century.

The legacy of the fact-value distinction implicit in empiricism and positivism is evident in the quantitative-qualitative methodology debates of the 1980s and 1990s, referred to as "the paradigm wars" (Patton, 2002a, p.92). Qualitative methodology is not congruent with the tenets of empiricism and positivism which underpin quantitative methodology. More precisely, qualitative methodology belongs to a different ontological and epistemological paradigm. Table 4.3

summarises some of the perceived weaknesses of qualitative methodology in light of the tenets of positivism and empiricism. These tenets had the effect of undermining qualitative methodology. Many adherents of quantitative methodology viewed qualitative methodology as lacking objectivity and deemed it to be lacking rigour and validity.

Table 4.3 Tenets of positivism and empiricism - implications for perceptions of qualitative methodology

Tenets of positivism and empiricism underpinning quantitative methodology	Implication of tenets for how qualitative methodology is perceived
Value-free, or values are strictly controlled (M. J. Smith, 1998)	Intrusion and contamination by values
Researcher detachment and distance from the object/subject being researched (Guba & Lincoln, 1989; Howe, 1992)	Researcher detachment and distance is lacking, thereby compromising objectivity
Minimisation of bias (Scriven, 1991)	Controls for bias are lacking
Verifiability/measurability of data (Guba & Lincoln, 1989; Howe, 1992)	Data is not measurable/verifiable
Generalisability of results (Guba & Lincoln, 1989)	Non generalisability of results
Context is stripped of confounding influences (Guba & Lincoln, 1989)	Contamination by contextual influences

Source: Compiled by author

The impact of the methodology debates on the evaluation profession has been significant (Donaldson & Christie, 2005). It is illustrated in a 1994 edition of the *New Directions for Evaluation* journal which aimed to address the "long standing antagonism . . . (and) . . . suppressed hostilities" between quantitative and qualitative evaluators (Reichardt & Rallis, 1994, p.1). The debate was highlighted again in 2003 when Richard Krueger, the then president of the AEA distributed a draft letter addressed to the US Department of Education to AEA members for their comment. The letter criticised the Department's decision that only evaluations based on experimental or quasi experimental designs would be funded. The draft letter was rejected by an influential group of AEA members (Donaldson & Christie, 2005). The following year, Claremont Graduate University hosted a debate to facilitate a resolution to "(the) apparent resurgence of issues reminiscent of the quantitative-qualitative paradigm wars . . . (which) has the potential to be destructive and stunt the healthy development of the discipline and profession" (Donaldson & Christie, 2005, p.70, 71). Excerpts of the addresses by Lipsey and Scriven indicate that the debate was heated (Donaldson & Christie, 2005). Subsequent literature shows that the methodological debate, framed variously as arguments for and against experimental and quasi experimental designs, and debates about what counts as credible evidence, remained unresolved (Cook, 2006).

As has been shown, the ontological distinction between facts and values has had far reaching epistemological and methodological consequences, for western science in general, but also specifically for evaluation. Most significantly, it determined what knowledge was regarded as legitimate and of worth. The fact-value distinction influenced the development of the theories of objectivity and subjectivity. It shaped what was regarded as valid scientific method. All of these factors have had fundamental epistemological consequences for values and value judgments, and their consequent exclusion from scientific knowledge and endeavour.

4.3.2 The inference problem

This section examines a second aspect of the case against values, namely, there is a fundamental problem about values and valuing (evaluation) involving the rules of inference in formal logic. Scriven (1995) describes the problem as "Whether and how one can get from empirical data to evaluative conclusions" (p.51). This is the consequence of "the impossibility proofs" established by Hume whereby formal logic does not allow a scientific or rational basis for evaluative conclusions (Scriven, 2007a; Taylor, 1961). Expressed simply by Hume: "no 'ought' can be deduced from an 'is' " (Oxford Companion to Philosophy, 2005, p.446). This logic problem supported the opposition to values in science, and to evaluation as a form of systematic enquiry. Given its role in the case against values, the inference problem is of sufficient importance to be explained in more detail.

In order to understand this logic problem, it is first necessary to define deductive and inductive inference. Deductive inference is defined as “From a given set of premises, the conclusion must follow” (Oxford Companion to Philosophy, 2005, p.194). Inductive inference refers to “one whose conclusion, while not following deductively from its premises, is in some way supported by them or rendered plausible in the light of them” (p.432).

The nature of the inference problem in formal logic as described by Scriven (1994, 1995, 2007) is as follows. An evaluative conclusion requires an evaluative premise/claim (1995). However the rules of deductive and inductive inference in formal logic do not permit either an evaluative premise/claim or an evaluative conclusion. Further, given its values-based nature, according to formal logic an evaluative premise can be regarded as “ . . . arbitrary because no evaluative premise can be established by the processes of logical, mathematical or scientific inference” (p.64). Therefore, no evaluative conclusion can be identified as being more or less true than another (1995). This means that claims about value lack the level of validity required by formal logic. Consequently value claims are deemed to be outside of scientific reasoning, a position referred to by Scriven (1995) as “the dismissive view of the status of evaluative claims” (p.64). This inference problem provided further justification for the dismissal of values as legitimate knowledge.

4.3.3 In summary: the case against values

This chapter has provided an overview of two aspects of the case against values as a form of knowledge and as a legitimate element of scientific endeavour. The case against values is based on ontological and epistemological grounds, where knowledge and science are deemed to be confined to that which is verifiable through empirical means. The long history of philosophical argument against values created a legacy of antagonism towards them. This legacy has continued to have an adverse impact on evaluation, as evidenced by the ongoing debates about the merits of quantitative and qualitative methodologies and the tendency in some quarters to prioritise random controlled methods (Davidson, 2006).

4.4 The case for values

While there is a significant body of work that discounts values, there is also theorising that attests to the epistemological legitimacy of values and valuing (evaluation). This came about following developments in philosophy and advances in logic (M. Scriven, personal communication, November 2, 2011). I identify three areas that collectively contributed to building the case for values: the challenge to the fact-value distinction, the identification of the means to assess value, and the solution to the inference problem.

4.4.1 The challenge to the fact-value distinction

A significant development in the case for values was the challenging of the long standing fact-value distinction. In order to explain this challenge, the terms formal logic and informal logic must be defined. Formal logic describes the theory of reasoning and the rules that govern it, much of which can be traced back to ancient Greek philosophy (Oxford Companion to Philosophy, 2005). Informal logic "examines the nature and function of arguments in natural (everyday) language, stressing the craft rather than the formal theory of reasoning" (p.532).

The empiricists and positivists asserted that there is an observational, descriptive language (for facts) and a separate prescriptive language (for values) (Taylor, 1961). According to formal logic, the distinctive nature of these two languages means that a descriptive statement cannot co-exist with a prescriptive statement. Informal logic provides a different perspective. In everyday discourse, words with descriptive and prescriptive meanings are used together (Campbell 1982; House 1996; House & Howe 1999). Scriven (1991) goes as far as to claim that there is "no ultimate factual language. And the more interesting side of the coin is that many statements which in one context would be clearly evaluative are, in another, clearly factual" (p.199). Scriven's comment highlights the context-imbued nature of language which can influence meaning. For example, the statement "Team Oracle (the US team) won the America's Cup against Team New Zealand in 2013" is a statement of fact. But used in particular contexts (for example, in a debate about the nature of the contest in the America's Cup), this sentence could be used to support a particular position about the value of yachting skills as opposed to computer technology (Oracle's win is attributed by some yachting commentators to the team's computerised foiling system, whereas Team New Zealand's foils were operated manually by the crew). (Example based on House & Howe, 1999, p.7).

Based on the informal logic perspective, House (1996) and House and Howe (1999) have portrayed the relationship between facts and values as a continuum. Located at one end of the continuum are "brute facts" (House & Howe, 1999, p.6) (for example, an elephant is larger than a mouse), while at the other end are "bare values" (p.7) (for example, stilettos are a key indicator of a fashionable woman). Located in the centre of the continuum are statements that combine facts and values. House and Howe (1999) state that most evaluative statements are located at the centre of the continuum because they are a blending of facts and values. House (2004b) notes: "Indeed, if you examine evaluation reports closely, you will find that facts and values are entangled so tightly it is difficult to pull them apart" (p.8).

Putnam (1926-2016) has systematically dismantled the fact-value duality, as reflected in the title of his 2002 book *The collapse of the fact/value dichotomy*. Putnam's argument is based on Dewey's rejection of philosophical dualisms. Putnam demonstrates how the fact-value dualism is no longer defensible, but does concede that in certain contexts it may be helpful to distinguish between fact and value. Putnam (2002) also argues for the validity of normative statements stating:

> . . . it is time we stopped equating objectivity with description. There are many sorts of statements - bona fide statements, ones amenable to such terms as "correct," "incorrect," "true," "false," "warranted" and unwarranted" - that are not descriptions, but are under rational control, governed by standards appropriate to their particular functions and contexts (p.33).

Putnam (2002) further observes that despite the rejection of the fact-value dualism by philosophy, it continues to be ubiquitous as evident in the on-going dismissal of normative claims as subjective: "The worst thing about the fact-value dichotomy is that in practice it functions as a discussion stopper, and not just a discussion stopper, but a thought stopper" (p.44).

The challenging of the fact-value distinction and this re-conceptualisation of the fact-value relationship were significant developments in the case for values as legitimate knowledge.

4.4.2 The means to assess value

A second factor related to the assessment of value added further weight to the case for values. During the 1960s a number of informal logicians asserted that it is possible to draw objective conclusions about values based on data about empirical properties implied by or associated with a value. Four of these theorists' views (in chronological order) Taylor (1961), Hare (1967), Scriven (1967) and Rescher (1969) are briefly discussed below.

Taylor (1961) summarises how value can be assessed as follows. It is important to note that Taylor emphasises the contextual nature of value judgments:

> In a factual assertion we claim that something has certain properties which can be discovered by empirical procedures. In a value judgment we claim that something has a certain value, but its value is not an empirically determinable property. Yet there is always a set of empirically determinable properties contextually implied by a value judgment. These are the good making and bad making (or right making and wrong making) characteristics of the evaluatum (p.241).

Scriven (1967) explicated what he described as "the general logic of evaluation". Table 4.4 contrasts Scriven's general logic with the approach explicated by Taylor (1961), Hare (1967), and Rescher (1969). Scriven's logic has four steps compared with the seven steps of the other theorists. Scriven's logic does not include the final step - justifying the norms used - identified by the other theorists. According to Taylor (1961), this final step of validating and justifying evaluative norms is critical and its absence serves to weaken the evaluative conclusion.

Table 4.4 Assessing value - comparison of approaches

Taylor (1961), Hare (1967), Rescher (1969)	**Scriven's general logic of evaluation (1967, 1980, 1991, 1994)**
1. Identify the object (X) and the value to be applied to the object 2. Identify the "class of comparison" to which X belongs (Z) 3. Identify norms for Z	1. Establish criteria of merit for the evaluand
4. Develop a set of operational statements describing levels of performance for each of the norms of Z	2. Construct standards for the criteria
5. Determine the characteristic(s) of X (the "good making characteristics")	3. Measure performance of the evaluand against the criteria 4. Synthesise and integrate data into a judgment of merit or worth
6. Compare X's characteristics with the operational statements above to come to an evaluative conclusion	
7. Justify the norms used	

The work of these informal logicians in identifying the means to explicate and measure the empirical properties associated with values (in a particular context) was significant. Their work enabled values to begin to gain legitimacy as knowledge, and for valuing (evaluation) to be regarded as a form of systematic inquiry. However, given the legacy of antagonism against values described above, the recognition of values and valuing as part of scientific discourse has been far from straightforward as is described in chapter 5.

4.4.3 The solution to the inference problem

As noted above, one of the grounds that logicians used to support the argument against values was the impossibility of inferring from facts to values using induction or deduction. Scriven (1996) describes how advances in informal logic enabled the emergence of another form of logic named probative logic. According to Scriven (1991), this is "a new, informal logic" (p.277) in contrast to formal logic such as mathematical logic. Probative logic "represents the normal logic of discourse" (p.277) and "everyday argument" (p.194). It underpins evaluative reasoning and other types of reasoning such as legal argument (Scriven, 1991). Probative inference is used to link premise and evidence with conclusions. The conclusions generated by probative inference are “prima facie . . . instead of categorical, conditional, or (quantitative) probabilistic ones” (Scriven, 1991, p.220). Elsewhere, Scriven (1980a) describes prima facie conclusions as those that are “good enough” (p.94), while House (1995) refers to them as “all-things-considered judgments” (p.40), and Schwandt (2010) defines such conclusions as “presumptive, rather than being a matter of proof” (n.p.). The emergence of probative logic (and inference) overcame another barrier in the case for values in knowledge and science.

4.4.4 In summary: the case for values

Three developments in philosophy and logic were identified as contributing to the case for values and valuing (evaluation) as a form of systematic inquiry. It is interesting to note that the emergence of values and valuing into the knowledge domain is relatively recent in the history of philosophy - the challenging of the fact-value distinction, and the work of informal logicians explicating how value may be assessed occurred in the 1960s. Despite such developments, the role of values in systematic inquiry remains disputed as evidenced in ongoing debates about the relative merits of different methods (Chelimsky, 2012).

4.5 Conclusion

This chapter has provided an overview of the cases for and against values in western philosophy of science. It provides the context for chapter 5 about evaluative reasoning within the domain of professional evaluation practice.

CHAPTER 5

EVALUATIVE REASONING

5.1 Introduction

Building on the foundations of valuing discussed in chapter 4, I now turn to the evaluative reasoning discourse in the evaluation literature. House (2004b) states "The unique contribution of evaluators is their ability to arrive at evaluative conclusions in a disciplined manner. Producing disciplined evaluative conclusions is the defining feature" (p.13). Evaluative reasoning is the discipline to which House refers. Its importance is reflected in the 2011 edition of the *Program Evaluation Standards of the Joint Committee on Standards for Educational Evaluation* which includes three standards relating to evaluative reasoning:

> Accuracy standard A7 Explicit evaluation reasoning: Evaluation reasoning leading from information and analyses to findings, interpretations, conclusions and judgments should be clearly and completely documented (p.209).
>
> Utility standard U4 Explicit values: Evaluations should clarify and specify the individual and cultural values underpinning purposes, processes and judgments (p.37).
>
> Accuracy standard A1 Justified conclusions and decisions: Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences. (p.165).

Scriven's general logic of evaluation introduced in chapter 4 (Scriven 1967, 1980a, 1991) provides the overarching theoretical framework for evaluative reasoning, with aspects of evaluative reasoning (such as valuing, evaluative argument, probative inference and evaluative judgment), fitting within the general logic. However for the purposes of this study, the general logic of evaluation is described here separately. This is followed by a discussion introducing four other elements that constitute evaluative reasoning - working logic, values selection and description, the evaluative argument, and the evaluative conclusion/judgment. Additional information about each of these elements is provided in chapter 7 in the presentation of the meta-evaluation findings.

5.2 General logic of evaluation

As described above, Scriven's (1967, 1980a, 1991) general logic of evaluation (hereafter the general logic) differentiates evaluation logically from other forms of systematic inquiry such as research, review and critique. Scriven states that this general logic applies to all fields of professional evaluation, such as product, programme, policy and personnel evaluation. The logic articulates a reasoning process whereby value terms are translated into meanings that are agreed (within a particular context), and that can be assessed or measured in a comparative manner. The notion of comparison is fundamental to evaluation (Stake & Schwandt, 2006), whether the comparison is formal and explicit (as in criteria and standards), or informal and/or implicit (as in expectations about professional behavior in a workplace setting). This is evident in the definition of evaluation provided by Rescher (1969): "A comparative assessment or measurement of something with respect to its embodiment of a certain value" (p.61).

The general logic also provides a framework for the fact to value inference to be made. Explained in simple terms, the general logic allows values to be translated into statements that can be assessed or measured against agreed criteria (for that value and in that context). Criteria are defined as "indicators of success or merit, variables that are not part of success itself (or definitionally connected to it) but rather tied to it by empirical research" (Scriven, 1991, p.111). Stake and Schwandt (2006) note that such comparisons are not limited to criteria - archetypes, models or other ideal types can be used. Data is collected about the performance of the evaluand against the criteria/standard (or other comparator) and assessed against it. Using probative inference, the resulting assessment is an evaluative statement which is an "all-things -considered judgment" (House, 1995, p.40). Table 5.1 provides an example of the application of the general logic of evaluation. For the sake of simplicity, in this example, a car (a Toyota Corolla) is being evaluated for its effectiveness (a specific value) as a family car (a specific context).

Table 5.1 Example of application of the general logic of evaluation

Steps in the general logic of evaluation	**Application to the example**
Establish criteria of merit for ‘effective family car’	*For the purposes of this exercise the criteria are:* • Safety • Room for adults, children and leisure equipment • Fuel efficiency
Identify standards of performance for each criteria (standards to be used are ‘excellent’, ‘satisfactory’, and ‘poor’)	*For the purposes of this exercise each criteria will have three levels of performance – excellent, satisfactory, poor.* **SAFETY STANDARDS:** **Excellent performance:** Six airbags **Satisfactory performance:** Four airbags **Poor performance**: No airbags
	ROOMINESS STANDARDS: **Excellent performance:** space for eight passengers and four scooters or sports bags **Satisfactory performance:** space for five passengers and two scooters or sports bags **Poor performance:** space for four passengers and no equipment
	FUEL EFFICIENCY STANDARDS: **Excellent performance:** 5 litres/100 kms **Satisfactory performance:** 7 litres/100 kms **Poor performance:** 9 litres/100 kms
Measure the performance of the evaluand against the criteria and standards	*Evidence collected about Toyota Corolla cars and examined against the criteria and standards* **Safety:** it has four airbags = *satisfactory* **Roominess:** it can carry five passengers and two scooters or sports bags = *satisfactory* **Fuel efficiency:** it does 6.5 litres/ 100 kms = *satisfactory/excellent*
Synthesise and integrate data into a judgment of merit or worth	(Using probative inference) we conclude that Toyota Corolla cars perform at a satisfactory level for safety and roominess. They perform at above satisfactory level for fuel efficiency. Overall, Toyota Corolla cars are an effective family car.

Scriven’s general logic of evaluation has been criticised by other evaluation theorists. Shadish, Cook and Leviton (1991) and Shadish and Leviton (2001) have critiqued the logic on a number of grounds. Their first criticism is that Scriven’s approach implies prescriptive valuing and rejects a descriptive approach. Shadish et al., (1991) favour a descriptive approach to valuing, whereby all stakeholder value positions are articulated (including conflicting values held by stakeholders) and no one value perspective is given greater weight than another (Shadish & Leviton, 2001). This is in contrast to a prescriptive approach which makes an evaluative judgment based on a particular “value position which is regarded as best” (Shadish & Leviton, 2001, p.184). Shadish et al., (1991) were also critical of the logic’s focus on evaluative judgments based on comparative and absolute standards of performance. While this may be appropriate for product evaluation, Shadish et al., (1991) question its appropriateness for evaluation which is being conducted for improvement or other formative purposes.

Six years after the publication of the critique by Shadish et al., (1991), twelve leading evaluators collaborated to publish a paper in the *Evaluation Practice* journal criticising Scriven’s logic (Stake, Migotsky, Davis, Cisneros, Depaul, Dunbar, Farmer, Feltovich, Johnson, Wiliams, Zurita & Chaves, 1997). The paper was based on findings from interviews with thirteen high profile theorists and practitioners about their evaluation practice. The findings indicated that evaluation standards were seldom explicitly identified, and that the level and exactness of measurement implicit in Scriven’s logic did not occur in practice. Further, the authors criticised the criterial approach inherent in Scriven’s logic as creating a narrow, limited perspective of the evaluand: “Criterial treatment of any evaluand transforms experiential knowledge of it into a knowledge of selected characteristics” (p.93).

Stake has continued to express his views about Scriven’s general logic. Some years after his original critique, Stake (2004) described the logic as a “useful heuristic” (p.17) but disagreed with Scriven’s approach of setting explicit criteria at the beginning of an evaluation. For Stake, criteria and standards are more “visions than cutting points” (p.263) and conceptualising them is an ongoing process of interpretation. Writing in a recent publication designed as a tribute to Scriven, Stake (2013) acknowledges “ . . . the superior merit and worth in the life work of Michael Scriven” (p.108) but continues to argue against the explicitness, reductionism and rationality inherent in his general logic: “Criteria, standards, validation . . . are often implicit. Professional evaluators do much of their evaluating without explicating the conceptual structure. However logical, evaluating is partly an intuitive act" (p.111). In another paper, Stake and Schwandt (2006) differentiate evaluation that uses explicit measures which they refer to as criterial thinking or "quality-as-measured" (p.407) from a significantly different approach to discerning quality, namely, "quality-as-experienced" (p.408). This approach to discerning quality is premised on practical or experiential knowledge which "is a form of non-cognitive knowing" (p.409). Eisner's (2004) view of evaluation as involving connoisseurship and criticism captures this "quality-as-experienced" approach. Eisner (2004) explains what it means to be a connoisseur by relating his experience as a young person working in a shoe store that sold high-end brands.

> I learned what to look for, and I could recognize quality when I saw it. In addition, I could give you reasons for my judgment. I became someone who, in this domain at least, could notice. This noticing ability, this ability to recognize differences that are subtle but significant in a particular qualitative display, is a pervasive feature of those who exercise connoisseurship in a particular domain (p.197, 198).

Eisner (2004) shares Stake's concern about the reductionism implicit in the criterial approach underpinning Scriven's logic, thereby narrowing the evaluator's perception of the evaluand. In contrast, evaluation by a connoisseur or expert critic incorporates "multiple perspectives" (p.199). The connoisseur or expert critic focuses on "the particular . . . how qualities that (the evaluand) possesses relate to one another" (p.199, p.201) and recognises that which is "subtle but significant' (p.198). These evaluation approaches which are based on practical or experiential knowledge as articulated by Eisner (2004), Stake (2004), and Stake and Schwandt (2006) are discussed further in chapter 9.

Scriven has been undeterred by the criticisms of Stake and other theorists, writing in a prolific matter for over more than 50 years. Stake's views have been challenged by other theorists who argue in support of Scriven's logic. For example, Davidson (2005) promotes explicit criteria and standard-setting through the use of rubrics. Greene (2011) asserts that valuing conducted in an explicit and transparent manner strengthens the validity and robustness of the evaluative judgment. G. T. Henry (2002) proposes a values inquiry method as a systematic approach to identifying relevant values.

5.3 Working logic

Scriven's general evaluation logic has been criticised for not considering factors such as context and audience (House & Howe, 1999). Such factors are addressed by Fournier's (1995) working logic. Fournier states that while Scriven's general logic underpins all types of evaluation, how the logic is applied in practice depends on four variables, namely, (i) the phenomenon being evaluated (for example, a policy or a product), (ii) the problem or issue being investigated (for example, impact or utility), (iii) the question(s) being asked and by whom, and (iv) the nature of the evaluative claim (for example, causation or performance). Drawing on formal logic, Fournier (1995) refers to these four variables as the "working logic" (p.18) of a particular evaluation. While general logic is common to all evaluations, working logic is the "logic-in-use found in everyday practice to establish and justify evaluative claims" (ibid). Further (as is described in section 5.4.) working logic explains the different ways in which the general logic is applied in individual evaluation methods/models. Scriven's general logic and Fournier's working logic informed the design of my meta-evaluation (chapter 7).

5.4 Selection and description of values

The term *valuing* is used by the evaluation profession in a number of ways, one of which is as a shorthand term to describe how values and the comparator against which they will be assessed are identified and defined (the first two steps in the general logic). House and Howe (1999) describe valuing as one of the greatest areas of contention among professional evaluators. This debate involves critical questions such as: whose values will be given priority in the evaluation? Whose values will be excluded? Who will identify the comparator such as criteria/standards? On whose values will the evaluative judgment be based? Addressing such questions is problematic given that the evaluator is surrounded by and works within a multiplicity of value perspectives, encompassing social, cultural and political values (Greene, 2011; Schwandt, 1997). House (2004b) expresses this succinctly: "Evaluators are fully 'situated' in the deepest sense: value-imbued, value-laden, and value-based" (p.7). Valuing therefore involves engaging with multiple normative logics (Schwandt, 2008a). The evaluator has to be cognisant of the (often conflicting) values of stakeholders (House 1996; Guba & Lincoln, 1989), cultural values (SenGupta, Hopson & Thompson-Robinson, 2004), and the values implicit in the larger political, social and cultural contexts within which a programme or policy exists (Schwandt, 1997). In the New Zealand context, this will include Treaty of Waitangi principles, the values of Māori as tangata whenua (Cram, 1997; Wehipeihana, 2008, 2013), Pasifika peoples and other ethnic groups. Underpinning these critical valuing questions are issues of ethics. Patton (1987, cited in Greene, 1990) sums up the valuing issue succinctly: " . . . whose interests will an evaluation serve?" (p.273). Similarly, Schwandt (2002b) stresses the importance of the ethical dimensions of valuing decisions:

> Ethical discussion aims at making us more critically aware of what we are doing. It brings us back to thinking about what it is to be a good evaluator, and to ask in whose interests should we be acting and for what purpose? These are ethical questions, and they should take precedence over technical questions about how to do evaluation (p.154).

Evaluation theorists have responded to Schwandt's (2002b) question "in whose interests should we be acting and for what purpose?" (ibid) by developing a range of evaluation methods/models premised on their particular epistemological and ethical perspectives. Examples include Democratic Evaluation (House, 1980; MacDonald & Kushner, 2005), Empowerment Evaluation (Fetterman, 2004), Participatory Evaluation (J. A. King, 1998), Responsive Evaluation (Stake, 2004; Abma & Stake, 2001), Transformative Evaluation (Mertens, 2007), and Utilization-focused Evaluation (Patton, 1997). As these names suggest, each is based on a particular value position. For this reason, N. L. Smith (2010) refers to evaluation methods/models as ideology.

The range of valuing approaches implicit in these and other evaluation methods/models are portrayed by Schwandt (1997, 2002b) who has identified three "ideal types" that describe different approaches to valuing in programme evaluation. The first type refers to "analytical value-free" approaches (1997, p.31). In this type, the evaluator does not make value judgments. Instead they describe the value positions of stakeholders and participants, giving all equal weight. Examples of evaluators who adopt this approach are MacDonald (Norris, 2015; Simons, 2015), and Shadish, Cook and Leviton (1991). The second type is referred to as the "emancipatory value-committed" type (Schwandt, 1997, p.33). According to this type, evaluation practice challenges values that serve existing power inequalities and is linked to political action. Mertens' (2009) Transformative Evaluation is an example of this type of valuing approach. The third type is the "value-critical" type (Schwandt, 1997, p.34) where the evaluator uses their expertise to add to and encourage practitioners' reflective, conversational critiques of the value commitments embedded in their practice. Patton's (2011) Developmental Evaluation illustrates this type. Davidson (2005) has identified a fourth type, noting that some evaluators chose to ignore values altogether. Scriven (1991) refers to this as "values phobia" (p.375).

Krathwohl (1980) offers an alternative typology to understand different valuing approaches. His typology is based on the locus of control of an evaluation, that is, "Who is to control the evaluation, and by implication, whose values are imposed on whom?" (p.40). He identifies three evaluation types: evaluations which have external control, via the evaluation audience and/or stakeholders; evaluations which have internal control, via the evaluator; and evaluations that balance internal and external control.

As the above indicates, valuing is fundamental to professional evaluation practice. It differentiates evaluation from other types of systematic enquiry, where values may be unacknowledged or ignored. As has been shown, the identification and description of values to be used in an evaluation involves issues of power and control. For this reason, evaluation has been described as a political activity (Greene, 1990). The evaluator must decide how they will respond to such issues according to their personal epistemological, ethical and political perspectives.

5.5 Evaluative argument

This section describes evaluative argument and demonstrates its relationship with evaluative reasoning. The generic term argument is defined by Toulmin, Rieke and Janik (1979) as: "A train of reasoning . . . the sequence of interlinked claims and reasons that, between them, establish the content and force of the position for which a particular speaker is arguing" (p.13). The notion of evaluation as argument was first proposed by House (1977) in response to the then dominant view that evaluation was principally about research methods, specifically quantitative methods. Such was the importance of evaluative argument for House that he included it in his conceptualisation of evaluative validity, described as consisting of three elements: truth (sound argument), beauty (coherence), and justice (fair politics) (House, 1980, 2014). Other than the work of Fournier and Smith (1993) and Fournier (1995) whose work draws on argumentation (described below), few evaluation theorists have written about evaluative argument.

Writing some thirty years after House's 1977 text, Schwandt (2008a) and Greene (2011) have expressed concern that evaluative argument is neglected by the evaluation profession. Schwandt (2008a) states:

> My concern is that in the press to master methods of generating data, we ignore the idea of developing a warranted argument - a clear chain of reasoning that connects the grounds, reasons or evidence to an evaluative conclusion (p.146).

In a similar vein, Greene (2011) observes: "Worrying about warrant is a core evaluator responsibility. It is because our inferences are consequential that we must have confidence that they are warranted" (p.90). Both of these authors use the word warranted in relation to evaluation argument and inference. At this point it is necessary to move outside of the evaluative reasoning discourse to explain the logic of argument, otherwise known as argumentation.

Western philosophy scholars have identified a logic of reasoning underpinning all types of inquiry that aim to build an argument (Fournier, 1995; Fournier & Smith, 1993; Mathison, 2005; Toulmin et al., 1979). This logic consists of six elements: claims, evidence, warrants, backings, conditions of exception, and qualifiers. These features work together to form a defensible argument. Three of the elements are particularly relevant to evaluative reasoning, namely, claim, warrant and backing.

A claim states what is to be taken as acceptable and legitimate (Toulmin et al., 1979). The claim has already been established as an element of evaluative reasoning (refer to the fourth variable of Fournier's working logic). A warrant is the because part of an argument. It legitimates the inference from the evidence and claim to the conclusion by appealing to an appropriate authority. Warrants are context-dependent and vary across disciplines (N. L. Smith, 1995). For example, lawyers use legal precedence as a warrant, physical scientists rely on the laws of nature (such as the law of gravity), and artists rely on expert opinion (Toulmin et al., 1979). Regardless of the particular warrant(s) used in a discipline, a warrant is one of two types (Fournier, 1995; Mathison, 2005). The first is named warrant-using. These are warrants that are well established, generally accepted and therefore are unlikely to be contested. Examples of this type of warrant used in research are those provided by sampling theory. Fournier (1995) notes that sampling theory may be regarded as a warrant because of the established state of the theory and level of agreement about it. The second type, named warrant-establishing are warrants that are not established or conventional and therefore may be contested. Use of this type of warrant requires a backing to legitimate the warrant. The backing is "added authority as to why the warrant should be accepted as legitimating the inference" (Toulmin et al., 1979, p.59). Evaluators are more likely to use the warrant-establishing type (Fournier, 1995; Mathison, 2005).

The link between the warrant, backing and evaluative reasoning is as follows. The warrant must be appropriate for the evaluative claim (the fourth variable of Fournier's working logic). For example, if the claim is about cause, then the warrant must be appropriate and relevant to causation. As noted in the previous paragraph, the warrant must also be appropriate for the particular context in which the claim is being made. For example, a warrant for a causation claim made in a medical context will be different from a warrant for a causation claim made in a legal context. N. L. Smith (1995) provides a real-life example of the relationship between context and warrant by relating the story of two separate investigations into the deaths of 47 sailors from an explosion on the US battleship Iowa in 1989. One investigation was undertaken by the Federal Bureau of Investigation (FBI), and the second by the American Psychological Association (APA). The FBI investigators used legal warrants appropriate for a law enforcement agency working in a criminal justice setting, while the APA psychologists used scientific warrants appropriate for a psychological science audience.

Evaluative reasoning can be strengthened by including warrants in the evaluation criteria and standards (steps one and two of the general logic of evaluation). For example, the criteria and standards for an evaluation of a healthy homes intervention could include house-related living standards produced by the World Health Organisation. Similarly, the evaluation of an initiative to encourage people to reduce fire risk in their home could use criteria and standards based on International Organisation for Standardization (ISO) fire safety standards. The use of credible sources (such as standards, peer reviewed literature, expert opinion) provides the backing for the warrant. This explicit link between the warrant (and backing) of argumentation and the claims, criteria and standards of evaluative reasoning is important given the contingent and interpretative nature of evaluative judgments (Stake & Schwandt, 2006). In his seminal book *The Logic of Evaluative Argument,* House (1977) argues that evaluative argument is important because evaluators rely on probative, rather than deductive or inductive inference: " . . . evaluation persuades rather than convinces, argues rather than demonstrates, is credible rather than certain, is variably accepted rather than compelling" (p.6). Schwandt (2008a) identifies four characteristics of a persuasive and credible evaluative argument. Firstly, the argument is practical (in the sense that it cannot be proven mathematically) and presumptive (rather than proven in an absolute sense). Secondly, the argument is dialectical in that it is about reasoning in a way that will address the audience's concerns about the credibility of the evaluative assessment. Thirdly, the argument is persuasive and based on inquiry. Lastly, the argument is contextual in two ways. The context influences what evidence, criteria and other aspects of the evaluation are deemed to be acceptable. The context (i.e. the particular client and stakeholders) also provide the focus for the evaluator to make their persuasive argument.

5.6 Evaluative conclusion/judgment

The final element of evaluative reasoning is the evaluative conclusion/judgment. Attention to the preceding aspects of evaluative reasoning will lead to an evaluative conclusion/judgment that is "legitimate and justified" (Fournier & Smith, 1993, p.316). Not all evaluative conclusions/judgments are the same. N. L. Smith (1981) demonstrates how evaluative conclusions/judgments differ according to the level of certainty they provide. He identifies three levels of certainty or "degrees of proof" (p.274) as follows: (i) "suggestive, where all that can be said is that X is possibly true", (ii) "preponderant, where what can be said is that X is probably true", and (iii) "conclusive, where what can be said is that X is undoubtedly true". Using the health sector as an example, N. L. Smith describes how the level of certainty required of an evaluative conclusion/judgment differs according to the evaluand characteristics (including the extent of risk to programme participants, and the extent of existing knowledge about the evaluand), the evaluation purpose and the evaluation context. For example, a formative evaluation of a low-cost programme where the potential risks to programme participants is low will tolerate an evaluative conclusion/judgement that is less precise. In contrast, the evaluation of an expensive intervention with vulnerable participants where the programme effects and side effects are not known requires greater precision and certainty. Julnes (2012b) offers a similar analysis based on the type of decision (including the level of precision required) for which the evaluative information will be used.

The making of an evaluative conclusion/judgment is not, however, a straightforward issue. There are two aspects to the debate among evaluators. The first concerns the question as to whether evaluators should make an evaluative conclusion/judgment. Shadish, Cook and Leviton (1991) and Shadish and Leviton (2001) assert that evaluators should not make a prescriptive conclusion/judgment. These theorists favour a descriptive account of value positions where the evaluation audience is left to make the evaluative conclusion/judgment.

The second aspect of the debate (among those who believe that evaluators should make evaluative conclusions/judgments) centres around whether it is appropriate to combine assessments of different evaluative dimensions into a single evaluative conclusion/judgment, and if this is deemed appropriate, how it can be done. This has become known as "the synthesis issue" (House, 1995; House & Howe, 1999; Scriven, 1994; Stake, 2004). Scriven (1993, p.72 cited in Julnes, 2012a, p.8) has wryly observed "Pulling it all together is where most evaluations fall apart". Interestingly, Scriven (1994a) has stated that synthesis into a single evaluative judgment is not required in all evaluations and that it is important to distinguish between those types of evaluations where a synthesis is required, and those where it is not.

Julnes (2012a, p.9-10) identifies four methods to aggregate assessments of multiple dimensions into one or more judgments: (i) "minimal aggregation" where the performance of individual dimensions is reported separately, (ii) "a checklist approach" in which the performance of the dimensions deemed to be important is recorded individually, (iii) "quantitative aggregation" of dimensions using quantitative calculations, such as numerical weight and sum (Davidson, 2006; House, 1995; Scriven, 1994a) (iv) "social aggregation" where social processes and evaluator intuition combine to create an evaluative judgment. House (1995), Stake (2004), and Stake and Schwandt (2006) are examples of theorists who favour this fourth approach. Stake and Schwandt (2006, p.406) emphasise the practical rather than instrumental aspect of judgment-making. As noted in section 5.2, these authors encourage evaluators to draw on practical knowledge in their evaluative sense-making, expressed in the form of perception and insight (Schwandt, 2008b). However Schwandt (2008b) notes the risks of such an approach in a political environment that values scientific and technical knowledge-making.

The synthesis issue appears to remain largely unresolved, perhaps due to the complexity it represents. Scriven (1994a) exhorts the profession to keep up their efforts to address the synthesis issue despite the challenges involved:

> While there's no silver bullet for the synthesis process we should try to get a valid rule in place whenever possible as long as we do so without distortion. Failing that, we should try for heuristics and rubrics, and failing that - as well as when we do that - we must do systematic and critical training of the judges in the remaining cases ("calibration") whether we are dealing with proposal, personnel, or product evaluation (p.369).

5.7 *Evaluative reasoning: a situated practice*

The evaluative reasoning discourse was significantly enriched by the publication of the 2012 edition of the *New Directions for Evaluation* journal which provides a range of perspectives about the valuing of programmes and policies in the public interest (Julnes, 2012a). An important theme emerging from the journal papers is the "contextually embedded and dependent" nature of valuing (Patton, 2012, p.98). Context is defined as referring to "The setting within which the evaluand . . . and thus the evaluation are situated. Context is the site, location, environment or milieu for a given evaluand" (Greene, 2005, p.83). To evaluate is to confront context. The programmes, policies and strategies we evaluate are not discrete, detached constructions but arise from and exist within a context: "Evaluands are social, political and moral constructions that embody the different (and often conflicting) interests and values of stakeholders (Schwandt, 1997, p. x). Most significantly, evaluations commissioned and/or funded by public sector agencies are determined by the priorities and interests of the government of the day.

Context determines the selection and implementation of a particular valuing approach (Julnes, 2012b). The valuing approach required will be determined by information and decision-making needs which determine the evaluation purpose and level of precision and complexity required of the information produced by the evaluation. Julnes (2012b) calls for better recognition of the multiple valuing paradigms, and improved alignment of valuing approach with contextual factors. However, Julnes (2012b) notes that his appeal for evaluators and evaluation commissioners to embrace a range of valuing perspectives " . . . seems impossible in the current milieu where embracing one value stance seems to require denigrating all others" (p.126). Julnes' observation is endorsed by Chelimsky's (2012) assertion that " . . . what we are talking about when we talk about valuing is methodology . . . " (p.78).

At a practice level, context determines the choice of criteria, how they are developed and by whom (G. T. Henry, 2002), the nature and validity of argument, and the warrants used (N. L. Smith, 1995). LaFrance, Nichols & Kirkhart (2012) emphasise the role of context in creating valid inferences, particularly when evaluating indigenous peoples: "Context is critical to valid inference; programs can be accurately understood only within their relationship to place, setting, and community" (p.59).

This chapter has presented evaluative reasoning as it is portrayed in the evaluation literature. The following section attempts to summarise this evaluative reasoning literature in a visual form.

5.8 Visual portrayal of the evaluative reasoning literature

Scriven (2012b) refers to "the logical infrastructure that makes it possible to claim that one can validate values" (p.18). As noted above, Toulmin et al., (1979) refer to "a chain of reasoning". A visual portrayal of the evaluative reasoning literature is presented in Figures 3 and 4, based on these notions of infrastructure and chain of reasoning. The figures illustrate the individual elements identified in the literature (Figure 3) and how they inter-relate (Figure 4). It is acknowledged that these diagrams have shortcomings. Most significantly, the diagrams are overly simplified. Secondly, the linearity shown in the diagrams fails to capture the recursive feature of evaluative reasoning, for example, criteria identified at the beginning of an evaluation may be subsequently revised as the evaluator's understanding of quality develops. A short explanation of Figures 3 and 4 follows.

The overarching structure for the infrastructure is provided by Scriven's general logic of evaluation (refer section 5.2), and Fournier's working logic which determines how the general logic is applied in respect of a specific evaluand, particular purpose, and type of evaluative claim (refer section 5.3). Fournier's focus on the situated nature of evaluative reasoning is reinforced by the context (refer section 5.7) being portrayed as the frame for the entire infrastructure. Contextual factors may constrain the evaluator in some way or another, thereby creating limitations for the evaluation (limitations and their effects are discussed further in sections 7.8.1 and 7.9.3). The two convex arch shapes on the left of the diagram represent lenses through which the evaluand is viewed. The first lens is provided by the comparator - whether the comparator is explicit as in criteria, or implicit as in a "quality as experienced" approach (Stake & Schwandt, 2006, p.408) (refer section 5.2). The second lens is provided by the method used in the evaluation (refer section 5.4) - the method determines the values that will underpin the evaluation and the locus of control of the evaluation (Krathwohl, 1980; Schwandt, 1997, 2002b). Moving towards the right hand side of the diagram, evidence about the evaluand is collected and analysed through the lenses provided by the comparator and method. The evidence provides the grounds to support an evaluative claim reached via probative inference. Responding to Patton's (2012) observation of this part of the reasoning process as resembling "a black hole" (p.97), this stage of the reasoning process is portrayed as occurring in a box. The inferential leap (based on probative logic) that is made between evidence and claim is portrayed by an arc of the double ended arrow (Figure 4). In order for the evaluative claim to be robust, it has to be supported by a warranted argument linking evidence to the claim (refer section 5.5). This then leads onto the final stage of the reasoning chain, the evaluative conclusion/judgment (refer section 5.6).

The orange lines on Figure 4 on page 79 show the inter-relationships between the individual elements as follows.

- The arc of the uppermost double-ended arrow indicates that an appropriate method must be used to produce the required level of precision and complexity of the evaluative conclusion/judgment (Julnes, 2012b) (refer section 5.7).

- The vertical arrow between the evaluative claim and warranted argument signals the need to use a warrant that is appropriate for the type of claim (refer section 5.5).
- The lowest double-ended arrow between warranted argument and context signals the need for a warrant to be used that is appropriate for the context (refer section 5.5).
- The longest, straight double-ended arrow between the comparator lens and warranted argument indicates that a warrant can be built into a criterion (refer section 5.5).

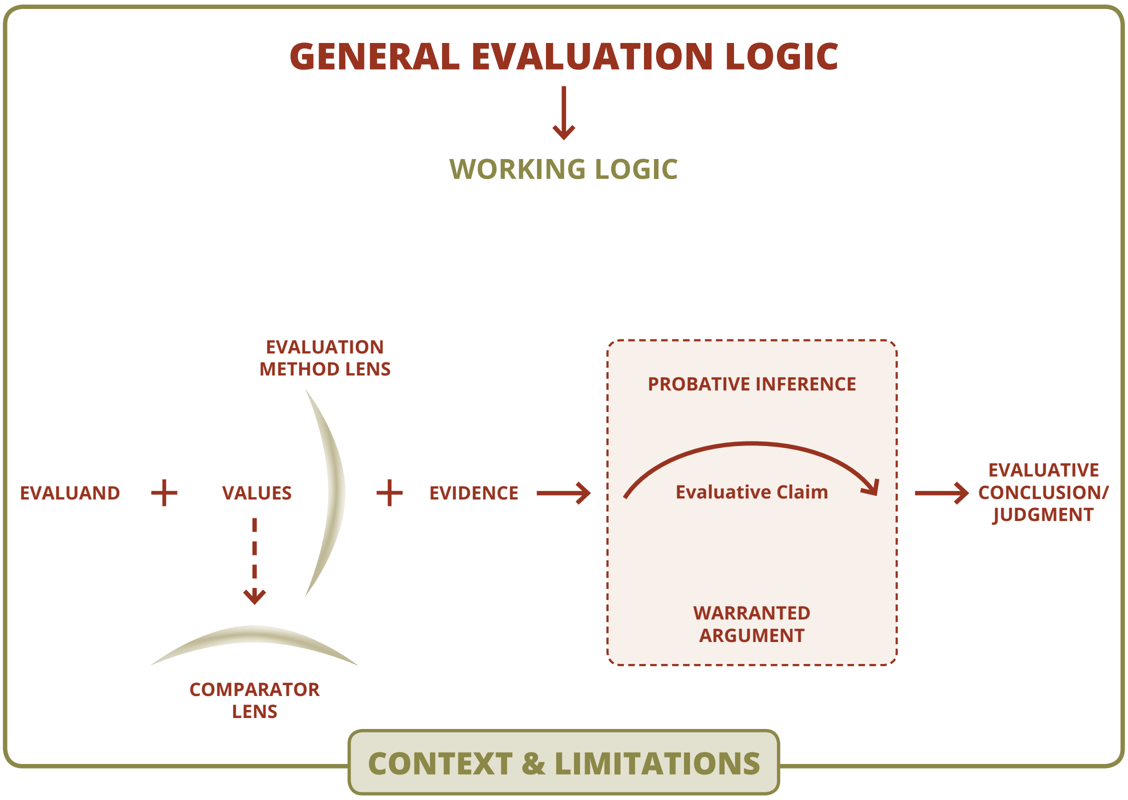


Figure 3 Evaluative reasoning as portrayed in the literature

Sources: The development of this diagram was informed by Fournier (2005), House (1977, 1980), Patton (2012), Schwandt (2002b), Scriven (1995, 2011b, 2012b), Toulmin et al. (1979).

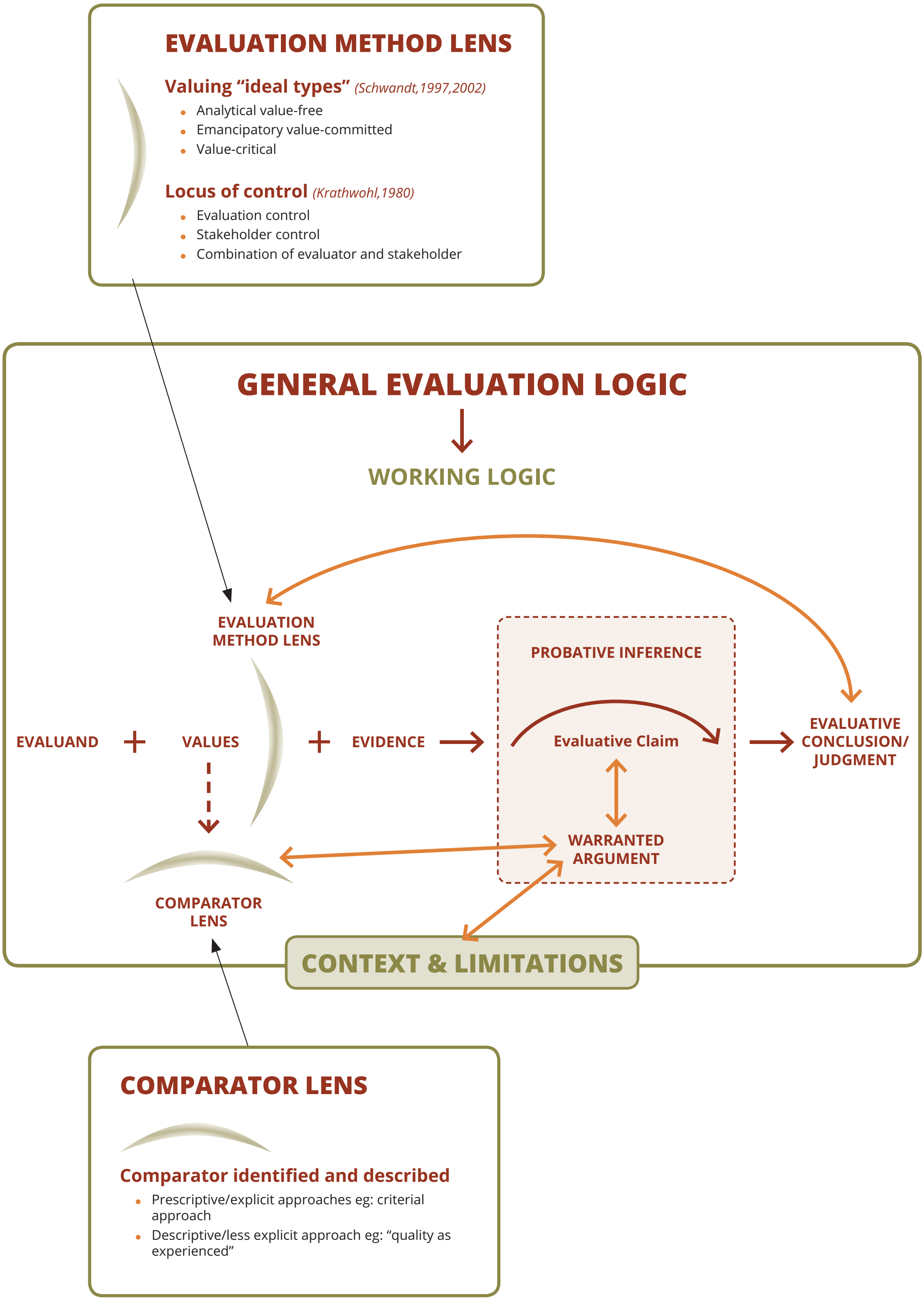

EVALUATION METHOD LENS
Valuing "ideal types" (Schwandt,1997,2002)
Analytical value-free
Emancipatory value-committed
Value-critical
Locus of control (Krathwohl,1980)
Evaluation control
Stakeholder control
Combination of evaluator and stakeholder
GENERAL EVALUATION LOGIC
WORKING LOGIC
EVALUATION METHOD LENS
PROBATIVE INFERENCE
EVALUAND
+
VALUES
+
EVIDENCE
Evaluative Claim
EVALUATIVE CONCLUSION/ JUDGMENT
WARRANTED ARGUMENT
COMPARATOR LENS
CONTEXT & LIMITATIONS
COMPARATOR LENS
Comparator identified and described
Prescriptive/explicit approaches eg: criterial approach
Descriptive/less explicit approach eg: "quality as experienced"

Figure 4 Evaluative reasoning showing inter-relationships and additional detail

5.9 Conclusion

Chapter 5 has shown that despite the centrality of evaluative reasoning to evaluation practice, the discourse has been characterised by debate rather than consensus within the evaluation profession. Underpinning such debate are fundamental questions about the purpose of evaluation, the role of the evaluator, and the role of commissioners, stakeholders, and participants involved in an evaluation. These are political questions requiring a response on the part of the evaluator. Some theorists describe this response as an ethical one (as opposed to a technical response). Other debates are about how the evaluator should undertake evaluative reasoning - in a prescriptive and explicit manner as per Scriven's logic of evaluation (1991), or in a manner that is non-prescriptive and draws on the evaluator's experience and practical knowledge of the evaluand (Stake, 2013).

The theory presented in this chapter forms the theoretical foundation for this study. It also informs the design of the Q methodology study (chapter 6) and meta-evaluation (chapter 7) that follow.

PART C

EVALUATIVE REASONING PRACTICE

Part C provides three perspectives on evaluative reasoning practice in the context of the Aotearoa New Zealand public sector. The first perspective - how evaluative reasoning is understood by evaluation practitioners - is provided by the findings of a Q methodology study (chapter 6). The second perspective - how evaluative reasoning is practised - arises from the findings of a meta-evaluation of evaluation reports written or commissioned by public sector agencies (chapter 7). The third perspective emerges from the findings of interviews with locally-based evaluation experts and international experts with knowledge of evaluation in Aotearoa New Zealand (chapter 8).

CHAPTER 6

PERSPECTIVE ONE: AN ABDUCTIVE INQUIRY

This chapter provides the first of three perspectives about evaluative reasoning and its practice in the Aotearoa New Zealand public sector context. This perspective is abductively derived through the use of Q methodology. The chapter begins with a discussion of Q methodology, firstly as theory and secondly as technique. This is followed by an account of how Q methodology was used in this study, after which the factors are presented and their interpretation discussed. Finally, the relevance of the interpretations for evaluative reasoning is examined.

6.1 Q methodology as theory

While applauding the increasing use of Q methodology (Q) as a research and policy tool, Wolf (2008/09) expresses concern about Q studies that fail to adequately account for the theory that underpins Q. Wolf's concern is significant because Q methodology is a theory with distinct epistemological features. It is therefore appropriate to begin this chapter with an explication of Q as theory.

Q methodology provides for the "systematic study of subjectivity" (Brown, 1991, p.2) where subjectivity is defined as "an individual's point of view" (McKeown & Thomas, 2013, p.ix) and "first person viewpoints" (Watts & Stenner, 2012, p.4). Q methodology is based on the premise that while subjectivity is unable to be proved in an empirical sense, it can "be shown to have structure and form" (Brown, 1980 p.6). William Stephenson (1902-1989), the founder of Q, describes Q as " . . . a mathematical-statistical key to what everyone calls 'mind' . . . it fits where nothing has before" (Stephenson, 1993/94, p.1). Watts and Stenner (2012) describe Q as "making a science of the subjective" (p.30).

The term subjectivity used in relation to Q methodology has a distinct meaning which is different to its everyday usage, as described below. Stephenson was a physicist and psychologist (of the Behaviourism School). In this school, the concept of operant describes a type of behaviour with two distinct features: behaviour is produced naturally, rather than being caused by something else; and behaviour is defined by the relationship it establishes with, and its impact on the immediate environment (Watts, 2011). Drawing on this conceptual frame, Stephenson describes subjectivity as operant in relation to the immediate environment (Watts, 2011). Such a definition rejects the notion of subjectivity being a phenomenological concept (Watts, 2011) or, expressed more colloquially, “mind stuff” (Watts & Stenner, 2012, p.32), and confirms that a person’s subjectivity (viewpoint) only exists in relation to something or someone in their immediate environment. Accordingly, a person’s subjectivity (viewpoint) is not static but may alter in relation to changes in or about the object or subject of the viewpoint (Watts, 2011). Another important characteristic of Stephenson’s subjectivity is that it is self-referent, that is, subjectivity is the “internal frame of reference” of an individual towards something or someone (McKeown & Thomas, 2013, p.2) or “that which is mine” (Wolf, 2008/09, p.10).

A second important concept in Q methodology is concourse theory. Stephenson used the term concourse to describe “the volume of discussion about a topic” (Stephenson 1980, cited in McKeown & Thomas, 2013, p.3) or “common knowledge” on a topic (Watts & Stenner, 2012, p.33). A concourse comprises words, pictures or objects about a topic, from the formal (such as academic papers) to the informal (such as cartoons), and other mediums such as music (Brown, 1991). Van Excel and de Graaf (2005) note that the terms concourse and discourse should not be confused. A concourse refers to relevant aspects of all of the discourses on a topic. Despite its centrality to Q, Watts and Stenner (2012) note that Stephenson’s discussion of concourse theory is variable, making it “a difficult concept to pin down” (p.34).

Epistemologically, Stephenson developed the conceptual frameworks for Q methodology in two important ways. Firstly, he articulated and clarified the notion of operant subjectivity and secondly, he developed concourse theory, both of which are described above. His ideas of both subjectivity and concourse depended in part on his understanding of abduction. According to Brown (1980), Stephenson viewed exploratory factor analysis used in Q (described in section 6.2) as “the technical or methodological extension of Peirce’s theory of abduction” (p.134). Unlike empirical research, Q does not start with the researcher’s “external frame of reference” (McKeown & Thomas, 2013, p.ix) as expressed in a research instrument such as a survey or interview guide. Rather, Q allows the person to express their viewpoint on the topic of interest according to statements representing the range of discourses on the topic. (This is discussed further in section 5.2). Writing about the use of Q to study political behaviour, Brown (1980) describes this feature in a descriptive manner: “[Q methodology] . . . takes a position on the frontier of behaviour, stripped of rating scales which carry their own meaning, and, shivering in the cold of uncertainty, tries to understand the political ramblings of the average citizen” (p.1).

More significantly, Stephenson rejected the dualism of “an objective natural world” (which can be scientifically investigated) and “a subjective human world” (outside of scientific enquiry) (Stenner, 2011, p.201). Stephenson (1953) asserted that “inner experience and behaviour are alike. Both are matters for objective . . . study” (p. 4). This challenge to the dominant empirical paradigm of the 1950s resulted in Stephenson’s work being subjected to on-going criticism from his peers. This is reflected in Stephenson’s account of the reaction to his initial papers about Q which “no one was prepared to take seriously” (1953, p.339). Q methodology was made more “radical and challenging” (Stenner, 2011, p.196) as a result of Stephenson’s assertion that Q sits outside of quantitative and qualitative research paradigms (McKeown & Watts, 2013): “Stephenson proposed Q methodology . . . as a fully-fledged scientific enterprise, replete with a distinctive logic of enquiry that, taken in its entirety, is tantamount to a subjective science of paradigmatic proportions” (p.73).

While some authors of Q studies describe Q as mixed methods because it uses both qualitative and quantitative methods (for example, Newman & Ramlo, 2010; Ramlo, 2016), this is not supported by other Q authors. Stenner and Stainton Rogers (2004) emphasise Stephenson’s stance that Q stands outside of existing theoretical frameworks. These authors proposed a new term *qualiquantology* “to grasp the peculiarity hybrid qualities of Q methodology” (Stenner, 2011, p.192). Further, the claim that Q is a mixed method reinforces the object/subject dualism that Stephenson aimed to dispel (Stenner, 2011).

It is this paradigmatic challenge to the traditional dichotomy of quantitative (object) and qualitative (subject) that gives Q its unique epistemology which remains challenging and controversial (for example, Kampen & Tamas, 2014). According to Stenner (2011), becoming a Q convert involves a "high epistemic cost" (p.192). This is evidenced in Q researchers' posts on the Q methodology Network reporting difficulties in getting their studies published in quantitative-focused journals. This is attributed to journal editors failing to understand Q's distinctive epistemology despite its use of the quantitative methods of correlation and factor analysis (Ramlo, 2016). For those researchers who have become Q converts, Q offers an abductive approach to "capturing and understanding personal viewpoints and attitudes on a topic of interest, that is both versatile and novel" (Ramlo, 2016, p.28). Given its abductive feature, Q offers opportunities for new perspectives and insights that can be used for theory generation.

6.2 Q methodology as technique

Having outlined the key theoretical features of Q methodology, this section examines Q as technique. Given Stephenson's stance of Q being outside of existing methodological paradigms, Q theorists refer to Q as a technique rather than a method (for example, Brown, 1991; Watts & Stenner, 2012; Wolf, 2012).

Stephenson was a research assistant for the British psychologist Charles Spearman who (with Karl Pearson) developed the method of correlation used in regression analysis (R) and was greatly influenced by him (Stephenson, 1993/94). Factor analysis is a data reduction technique whereby "a number of tests (variables) are applied to a sample of persons" (Stephenson, 1953, p.15) to determine whether the variables are inter-related. The underlying relationships (which are not dependent on each other) are referred to as factors (Bryman, 2008). Stephenson (1953) reconceptualised Spearman's and Pearson's method to enable by-person factor analysis, that is, "an experiment (is designed) in terms of people . . . to assess qualities of performance with respect to each person in turn, and then to make correlations between people" (p.16). Stephenson used the letter Q to distinguish his approach.

Put simply, in R, variables are characteristics (test scores, traits) of a person and factor analysis looks for which characteristics go together. In Q, variables are the people who participate in a Q study. Factor analysis is used to group participants together according to some underlying dimension of commonality in their viewpoints (Wolf, 2012). This enables shared meanings or viewpoints to be identified (referred to as *orientations*) and the extent of each participant's association with a particular orientation. (This correlation is analogous to the step in R that correlates, for example, tests results on maths and music achievement) (Wolf, 2012).

Despite sharing the same statistical techniques, R and Q are based on very different epistemological paradigms and have dissimilar purposes. Drawing on the literature, the key differences between R and Q are summarised in Table 6.1.

Table 6.1 Differences between R and Q

	R	**Q methodology**
Purpose	The objective analysis of a topic of interest. Identifies the structure of opinion or attitudes in a population of interest (Spearman, 1904). "The purpose is to reduce and/or eliminate the qualitative and subjective" (Stenner, 2011, p.198).	Reveals differences in points of view, attitudes, opinions about a topic of interest. The focus of Q is on "the constructions, rather than the constructors (participants)" (Stainton Rogers, 2005, p.180). "The purpose is to maximize the qualitative and subjective" (Stenner, 2011, p.198).
Logic	"Hypothetico-deductive" (Stephenson, 1953, p.17).	"Postulatory-dependency" (Stephenson, 1953, p.17).
Participant selection	A representative sample of the population of interest.	One or more individuals who have been selected using purposive sampling.
What is being collected	The degree to which a person has a certain trait/characteristic (assessed one at a time)	During a Q sort, the participants "put meaning upon and draw meaning from the statements" in the sort (Wolf, 2008/09, p.27). In this way, Q provides access to "the unrestricted viewpoint of its participants" (Watts, 2011, p.45).
How data is being collected	By use of a standardised data collection instrument, such as a test or survey with fixed categories.	By a multi-item comparison and ranking/scoping on a grid.

How the data are treated statistically	In R, the measuring units are "objectively scorable traits" (Brown, 1980, p.19). In R, correlation summarises the relationships among the traits, and factor analysis identifies the clusters of traits.	Unlike R, "there is no common unit of measurement other than the person's self-referential viewpoint" (McKeown & Thomas, 2013, p.48). This means that Q involves the correlation and factoring of people - correlation summarises the views among the people, and factor analysis identifies the clusters of people with shared views (McKeown & Thomas, 2013).
Outputs	The outputs of R describe the characteristics of the sample population that are statistically associated with the topic of interest.	The outputs of Q reveal factors or clusters of viewpoints about a topic of interest.

The rest of this section gives a brief overview about how a Q study is undertaken to provide a context for the following section describing how I conducted my Q research.

The first task for the Q researcher is to collect a large number of statements representing the range of discourses about the research topic (pictures, objects or sounds may be used as an alternative to written language). Using a matrix (or other framework) of themes/subthemes in the discourses, the researcher sorts the individual statements by theme/subtheme and then systematically selects statements which together provide "a representative miniature" of the larger concourse (Brown, 1991, p.6) (referred to as the *Q set*). Wolf (2012) notes that if the Q researcher is already very familiar with the research topic, they may create the matrix (or other framework) first and then collect statements that fit within it.

Recruiting participants for a Q study is done by purposive sampling to provide a differentiated sample (Wolf, 2012). Participants (referred to as the P set) are asked to rank each of the items in the Q set according to a specific instruction (referred to as the *condition of instruction*) and an ordinal ranking scale (such as from - 4 to +4) (Watts & Stenner, 2012). Participants are further guided by the requirement to sort the statements according to a predetermined distribution (the degree of flatness or steepness of the distribution is referred to as the *kurtosis*) (Brown, 1980). This sorting and ranking of the Q set by an individual is referred to as a Q sort.

Through their ranking of the items, participants express their viewpoint on the topic (McKeown & Thomas, 2013), “tell a story” (Stainton Rogers, Stenner, Gleeson & Stainton Rogers,1999, p.249) or provide “a picture (of their) . . . conception of the way things stand” (Brown, 1980, p.6). A participant’s subjectivity is expressed in how the items are understood and how they are ranked (Brown, 1991).

The Q sorts undergo correlational and Q factor analysis (using Q software, in this case PQMethod) to identify statistically significant patterns of associations (referred to as *factors*) and the extent of each participant’s association with a particular factor. Stainton Rogers (2005) describes the distinctiveness of a factor: “Each factor represents a fully alternative understanding of the topic of interest” (p.191).[6] The researcher’s task is then to interpret each of the factors expressed as a factor array. Wolf (2012) identifies two broad stances researchers may use to interpret patterns from Q factor analysis. Firstly in a person-centred Q study, the researcher enquires into the ways in which people view a matter from their perspective and the underlying predispositions that may influence a person’s response to the items in the Q sort. In a discourse-centred Q study, the researcher is interested in the discourses with which participants align. My approach is person-centred, examining how evaluative reasoning is understood by professionals undertaking public sector evaluation.

6.3 Design and conduct of the Q study

This section describes the design and conduct of the Q study under the following headings: concourse development, item selection, participant selection, Q sort preparation, Q sort administration, factor results, factor description, and interpretation of results.

6.3.1 Concourse development

Using the literature review as a starting point, I collected succinct statements about evaluative reasoning, including statements from value theory, evaluative reasoning theory, and articulations of evaluator practices in relation to evaluative reasoning. I then reviewed the approximately 300 statements against Figure 3 (refer chapter 5)

[6] Expressed in statistical terms, the factors derived from a Q study are orthogonal, that is, independent and at 90 degrees to each other (Stainton Rogers, 2005).

(developed from the review of literature) to ensure the key theoretical positions about evaluative reasoning were included. This review proved very useful as it highlighted omissions in both the statements and the conceptual framework. It also revealed duplicative ideas in the statements.

6.3.2 Item selection

A workshop was held with five evaluation colleagues (who are experienced practitioners and familiar with my doctoral study) to select statements for the Q set. This collaborative approach helped to ensure the Q set was representative and comprehensible. The importance of context in textual comprehension was demonstrated in this collaborative approach to statement selection. While I understood the statements from having read them in situ, some had become less understandable as a result of being extracted from their context. The selection group began by sorting the statements into identified themes as shown in Table 6.2.

Table 6.2 Themes for statement selection

Themes	Sub themes
1. How evaluation is defined and its purposes	
2. How values get privileged in evaluation	Contextual factors
	Methods
	Worldviews
3. Who does/is involved in valuing	Evaluator only
	Stakeholders only
	Both
4. Evaluative criteria: explicit/implicit/non existent	Explicit - prescriptive/external
	Explicit – descriptive/emergent/consensual
	Implicit
	No criteria
5. Evaluative judgments: how evaluative judgments are arrived at, and by whom	Emphasis placed on evaluative judgments
	Who is involved in judgment-making
	Logic/warranted argument
	Intuitive approaches
6. The stances/roles/behaviours adopted by evaluators in relation to thinking about values and practising evaluative reasoning	

The selection group then reviewed the statements by theme/subtheme. The aim was to select 35 or 43 statements. I chose to have 35 statements to enable participants to complete a Q sort and participate in a short discussion about the way they had sorted the statements within their sixty minute lunch break. Any ambiguous or duplicated items were rejected. The items that best addressed the theme/subtheme were then selected. Any aspects about the theme/subtheme (e.g. a specific theoretical position) that was missing from the statements were identified, and one or more other statements were subsequently included. Following discussion, themes one and six were excluded from the Q set as it was assumed they would be implicit in, or emerge through the factors (orientations).

The selected statements (hereinafter referred to as *items* or the *Q set*) were subsequently reviewed by my two supervisors and New Zealand-based Q methodology expert Dr Amanda Wolf, for their clarity and readability. I was advised to re-write some of the items, replacing academic terms with more accessible language to make it easier for the participants to self-reference (so as to avoid feeling they were undertaking a cognitive task) (A. Wolf, personal communication, 23 May 2012). The Q set was then trialled with an experienced researcher (an ex-academic and evaluation practitioner) who suggested some minor wording changes to some items and gave feedback on my administration of the Q sort.

6.3.3 Q sort preparation

I drew the distribution shown in Figure 5 onto large (122.5 x 91.5 centimetre) heavy-duty cardboard. A velcro tab was glued in the centre of each of the boxes in the matrix. Each of the statements was written onto a card, the cards were laminated and a velcro tab attached to the back of each card. This enabled participants to easily attach (and detach) the cards on the board as they wished. The shape of the distribution, was guided by the advice of Watts and Stenner (2012). They advise that a more spread-out distribution is appropriate for Q participants who are familiar with the topic of interest. Whereas a steeper shape with more boxes in the middle columns is preferable for participants who are less informed about the topic, providing sorters with more opportunity to express uncertainty.

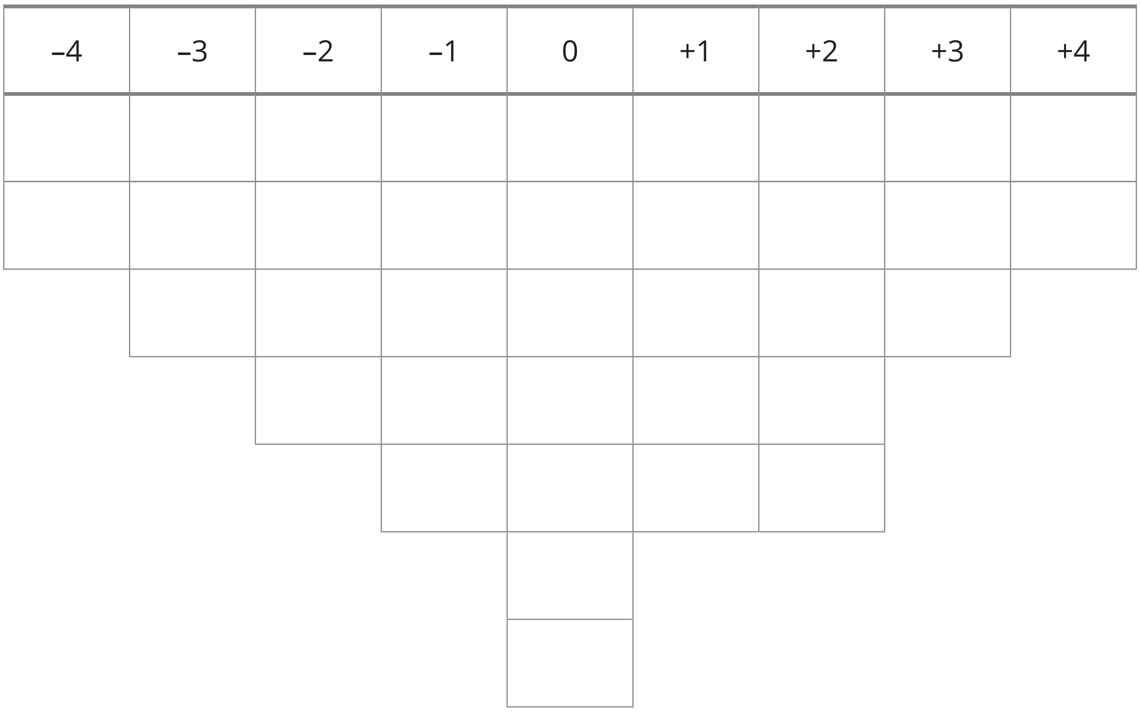


Figure 5 Q distribution matrix

6.3.4 Participant selection

My aim was to recruit a differentiated group of 30 participants, including evaluators and evaluation commissioners who are government employees and consultants, and New Zealand European, Māori and Pasifika. (Note: In Q, the term sample refers to the Q items, not the participants). This number of participants was recommended as an appropriate size for a study of this nature (Wolf, 2012). I prepared an email about the Q study which was sent to 300 or so people who subscribe to the WEG email list. My rationale for this approach was that since subscribers are Wellington-based, they are likely to be undertaking public sector evaluation, either as a public servant or contractor. (The majority of evaluations undertaken or commissioned by government agencies are done so from head offices which are based in Wellington where the government is located). Although 21 people expressed an interest in participating, only 15 subsequently completed a Q sort (the others did not respond when asked to identify a suitable time for the Q sort). I then undertook some targeted recruitment (using professional networks) to ensure a differentiated sample of 30 participants, in particular Māori and Pasifika evaluators (of whom there are few). This involved a visit to evaluators based outside the Wellington region who undertake work for government agencies. (While the term evaluator is used in this chapter, it should be noted that one of the participants in the Q study is a policy analyst who commissions evaluations and a second participant undertakes evaluations but does not refer to themselves as an evaluator).

Information about the study (Appendix A) and a consent form (Appendix B) were emailed to participants before the agreed time for the Q sort. The participants (P set) consisted of: public sector employees (18) working in 10 government agencies; consultants (11) (either working independently or as an employee in a private research organisation or university); and one person who works for a non-government organisation. The ethnicity of the P set is as follows: New Zealand European (17), Māori (5), Pasifika (3), other (2), not specified (3).

6.3.5 Q sort administration

The Q sorts were conducted in a room at the participant's workplace, usually during their lunch hour. Participants were given another copy of the information sheet and then invited to sign a consent form. Participants were asked to provide a unique identifier to enable them to identify their results when the results were subsequently emailed to them, and so individual sorts could be referenced anonymously. Participants were then given written instructions about how to do the Q sort (Appendix C), which included the following condition of instruction: *"Please sort the statements to reflect your point of view as a professional evaluator, +4 being the two statements that are most similar to your views, and -4 being the two statements that are most different to your views."* The instructions also suggested an approach to sort the items, as follows: *"Sort the items initially into three piles (i) items aligned with your point of view, (ii) items not aligned with your point of view, (iii) items you need to think about, do not understand, or about which you do not have a view. Then proceed to the detailed sorting of the items."* It was also suggested that participants begin by placing items onto the board at the far left and far right ends, and to work inwards towards the centre.

The sort board was placed upright on a table. It was large enough for me to sit behind so I was not visible to the participant. The board appeared to provide a space for participants to think, and I was surprised to hear a few people talk to themselves as they sorted the items. On completion of the sort, I asked the participant an open ended question about their reasons for selecting the items placed at +4, +3, -3, -4 and recorded their responses on paper (Wolf, 2012).

I was also interested in participants' choice of items in the zero column. For some participants, the items in this column were the taken-for-granted aspects of evaluation practice, whereas for other participants these were the items about which they did not have a view, or did not understand. I recorded participants' comments about their sort on paper and subsequently referred to them to aid interpretation of the three factors that emerged (discussed in sections 6.3.6 – 6.4.5).

Q methodology was a new experience for all but one of the 30 participants. The majority of participants made unprompted comments about enjoying the task of sorting the items and arranging (and rearranging) them on the board. Some also expressed interest in learning more about Q and how they might use it in their work. This prompted me to write a two-page overview about Q methodology which I emailed to interested participants after the sort (Appendix D). I also asked participants who expressed interest in my research whether they would be interested in being part of a discussion with other participants about the Q results. While most signalled their willingness to do so, only three responded to an email invitation (some months later) to a discussion (discussed below).

6.3.6 Factor results

The data from the 30 Q sorts were inputted into PQMethod software (version 2.33, December 2012). For each Q sort, the data comprises the number allocated to each statement representing where each card was positioned in the distribution matrix, and the unique identifier for the participant. The software undertakes the following statistical procedures (McKeown & Thomas, 2013): (i) correlation of each Q sort with each other Q sort, (ii) the intercorrelational matrix is factor analysed (to identify clusters of common meaning), (iii) the factor scores are rotated (using Varimax rotation), and (iv) factor arrays produced (identifying people who are statistically associated with a factor).[7]

The unrotated factor matrix showed that factors one, two and three have eigenvalues of greater than 1 (factor one explains 47% of the variance, factor two

[7] Factor analysis in Q is based on Centroid Method as this was the preferred statistical approach of Stephenson for theoretical reasons (McKeown & Thomas, 2013). Stainton Rogers (2005) describes Varimax rotation as the usual approach used by Q methodologists, but there is disagreement among some Q practitioners over this.

6% and factor three 4%).[8] Table 6.3 shows the results of the rotated (Varimax) factor matrix, with the three factors accounting for 51 percent of the variance. I then calculated the significant factor loading for my study (0.43) to identify the sorts that are statistically significantly associated with one or more of the three factors (referred to as a *defining sort*) as shown in Table 6.3.[9] There are 28 defining sorts. Twenty-two of the 28 sorts are associated with one factor, and six sorts are associated with two factors. Factors one and three are strongly correlated (0.8485), whereas the correlation between factors one and two is 0.5895, and between factors two and three is 0.5705.

Table 6.3 Summary of factor results

	Percentage of study variance accounted for	**No. of participants significantly statistically associated with one or more factors**
Factor 1	19%	14, of which 3 are also associated with factor 3
Factor 2	12%	6, of which 3 are also associated with factor 3
Factor 3	20%	14, of which 3 are also associated with factor 1, and 3 with factor 3
Total	51%	28 [10]

6.3.7 Analysis of factor arrays

The factor arrays are shown in Appendix E. As described above, there are two approaches to interpreting the factor arrays produced from Q factor analysis: person-centred or discourse centred. This study is the former. Watts and Stenner (2012) identify two steps in analysing the factor arrays. The first step involves cross factor item comparison, followed by a comparison of items within a factor.

[8] Eigenvalues are produced from a statistical procedure which determines whether or not a factor is significant. Eigenvalues greater than 1.00 are considered significant (McKeown & Thomas, 2013, p.53). This is referred to as the Kaiser-Guttman criterion (Watts & Stenner, 2012).

[9] A significant factor loading is calculated which determines the point at which a Q sort is deemed to be statistically significant in respect of the factor.

[10] As six sorts are associated with two factors, this column does not add up.

In regards to the second step, they emphasise that the holistic nature of a factor is fundamentally important to Q. This is achieved by examining the interrelationship of items within the factor, rather than simply focusing on distinguishing statements and items that have been ranked at the outer ends of the scale. (A *distinguishing statement* is an item that has been ranked in a significantly different way to its ranking in the other factors (Watts & Stenner, 2012, p.217)). I followed the advice of Watts and Stenner, but added a third step as described below.

Step 1: *Cross factor item comparison.* I heeded Watts and Stenner's (2012) advice to cut the factor arrays in different ways in order to examine the data from diverse perspectives. I also followed their advice to interpret the results in their anonymised form before examining the results of individual participants so as to allow the data to speak for itself. Accordingly, I identified (i) items ranked the same across the three factors (ii) items ranked almost the same across the three factors (iii) distinguishing statements and +4 items from each of the three factors (I paid less attention to the - 4 items because they were almost identical across the three factors).

Step 2: *Comparison of items within a factor.* For each factor I identified (i) the items ranked by score from +4 to -4 (ii) items ranked +4, items ranked higher in the factor than in any of the other two factors, items ranked lower in the factor than any of the other two factors, and items ranked -4.

Step 3: *Examination of each sort and qualitative comments.* At this point, I grouped the 28 participants with a defining sort according to their factor result. Taking each factor in turn, I then examined each participant's sort with their qualitative comment. I sought to understand the viewpoint expressed in each sort in its totality (as expressed in the scoring and qualitative comments). I then compared the individual sort with other sorts in the factor to identify whether any common themes existed.

Step 4: *Discussion with colleagues and supervisors.* I presented my initial interpretations of the factors to the group who had assisted with the item selection and supervisors. The feedback from these colleagues indicated that the interpretation of factor 2 was less developed than the other two factors, and needed further examination.

6.3.8 Interpretation approach

As noted above, Q provides an abductive approach to exploring a topic of interest. A significant aspect of this abductive inquiry lies in the researcher's interpretation of the statistical results of a Q study, that is, the Q researcher's interpretation of the subjective viewpoints of the Q sorters. I was therefore keen to explore the extent to which my interpretations of the factors would be corroborated by other evaluators and to test my hunches (albeit, in their formative stage) on them. (From this point on, I refer to the interpreted factors as orientations).

Firstly, I sent each participant their Q sort result, together with a seven-page paper summarising the three orientations. I invited participants (who had previously signalled their interest) to meet to discuss the orientations. Three participants responded to my invitation and a group discussion took place responding to the questions in Table 6.4. Secondly, I presented the orientations in workshops at the ANZEA Conference (attended by approximately 25 people) in July 2014, and the Wellington Evaluation Group (attended by approximately 30 people) in September 2014 where I posed the same questions. While the length and quality of the discussion varied across the three events, it was helpful in that it made me be more transparent about the assumptions and inferences implicit in my interpretations of the orientations, and opened up new ideas and insights for further thought.

Table 6.4 Questions about the orientations

1	Do the common themes across the three orientations surprise you? If yes, why? If no, why not? What do you think is the reason(s) for this commonality across the three orientations?
2	Do the three orientations make sense to you? If yes, why? If no, why not?
3	Is there anything that you might have expected to be included in one or more of the orientations that wasn't included?
4	Have other ideas emerged for you from the orientations? If yes, what?
5	Is there anything in the orientations that doesn't make sense to you?

6.4 The three orientations

6.4.1 Introduction

This section begins by identifying the commonalities across the three orientations. The three orientations, written as narratives and named after their main themes, are then presented. The narratives include qualitative comments made by participants associated with the orientation after completing their sort. The item score for the orientation is shown in brackets (orientation one is abbreviated as F1, orientation two is F2, and orientation three is F3. Therefore 'F1 -4, F2 +2, F3 0' means the item has a score of -4 on orientation 1, +2 on orientation 2, and zero on orientation 3). Some of the items are expressed in the negative. Therefore a negative score represents disagreement with a negative stance, or in other words, agreement with a positive stance. The narratives for each factor may not include references to items scored +1, 0, -1. As noted above, participants identified different reasons for their choice of items in the centre of the matrix distribution. Therefore there may be no obvious explanation for items placed in these positions. Distinguishing statements are identified with an asterisk.

6.4.2 Common themes across the three orientations

Despite the orientations articulating statistically different perspectives, six items have identical or similar scores across the three orientations. In describing these six consensus items, I am mindful of Wolf's advice that items with identical scores across factors do not necessarily have the same meaning for each factor. Rather, the meaning has to be interpreted in the context of the rest of the items in the factor (Wolf, Q Methodology Network, 2 December 2014). I therefore looked to participants' comments in the post-sort discussion about their placement of the items at the outer edges of the matrix (-4, -3, +3, +4) to provide insights about these scores. In light of participants' comments, I am confident that most consensus items represent shared meanings across the orientations.

First, the notion of the evaluator needing to be detached from the evaluand in order to provide an independent and objective assessment (item one - see below) is rejected in all three orientations. A participant associated with orientation two commented: "Evaluators aren't god-like beings floating above their subjects. Detachment is not desirable or achievable" (755). The same participant described the evaluator as being: " . . . part of a network of relationships including stakeholders, policy analysts and programme implementers" (755). A participant associated with orientation three was dismissive of the notion of a value-free perspective: "Providing a distanced view is a myth. Going in pretending you are providing a value-free judgment is silly" (824).

> ***Item 1:*** *Evaluators need to maintain a detached stance from an evaluand so they can provide a distanced view. This requires minimum interaction with staff involved with the evaluand. It's the only way to ensure an independent and objective assessment of the evaluand (F1 -4, F2 -3, F3 -4).*

This rejection of the idea of the evaluator keeping their distance from the evaluand is reinforced in the negative scores for item two about evaluators needing to keep stakeholders at arms' length. The three orientations endorse stakeholder involvement in the evaluation process.

> ***Item 2:*** *Stakeholders should not have any input into the evaluation process. Assessing the performance or quality of an evaluand is the sole responsibility of the evaluator (F1 -4; F2 -4; F3 -4).*

The reasons for stakeholders to be involved in the evaluation process were explained at length by participants in the post-sort discussion about their placement of the items. The first reason is the need for culturally-appropriate practices when working with Māori and Pasifika peoples. A participant associated with orientation one cited a commonly used phrase about Te Ao Māori (the Māori world): "He aha te mea nui o Te Ao? He tangata, He tangata, He tangata" (What is the most important thing in the world? It is people, people, people) (195). Another participant associated with orientation one said:

> If you are evaluating from a Te Ao Māori (Māori world view) perspective of delivering to Māori, how could you not involve providers? How would you capture the material if you didn't involve them? If you are an economist looking at data, then this would be OK. But because our work is about services, this approach wouldn't work. The Māori worldview is that if something is going to work and if you are evaluating it, you need dialogue. It sits with kanohi ki kanohi - face to face (10).

This people-centred approach to conducting evaluation is also reflected in a comment made by a participant associated with orientation three: "Evaluation is about doing it with people, not doing it on people" (384). Participants value the knowledge stakeholders bring to an evaluation, as expressed by a participant associated with orientation two:

> Stakeholders are the experts about the evaluand, not the evaluator. The evaluator needs to draw on their knowledge about the evaluand. No evaluator can be an expert in everything. The richer source is the people who the programme affects (755).

Other reasons provided by participants have an instrumental focus. The first concerns the utility of the evaluation: "The evaluator needs to find out what matters to stakeholders so the evaluation findings will be meaningful and useful for them" (55). A participant associated with orientation one said:

> To understand stakeholders' needs, their input is essential. Therefore to contribute meaningfully to evaluation as an evaluator, to make a difference, a transformative change, stakeholders need to see themselves in the evaluation process (1840).

Participants evaluating community-based service providers described stakeholder involvement as encouraging learning about evaluative thinking to strengthen the quality of service delivery. These participants also described the anxiety of some providers about evaluation due to government agencies using evaluation to cut providers' funding and service provision. Such anxiety requires the evaluator to build trust and confidence by working alongside and involving providers in the evaluation.

Second, the three orientations endorse the context-dependent nature of merit and worth (item 20). Consequently, the identification of evaluative criteria (standards) is an important part of evaluation in which stakeholder involvement is critical:

> ***Item 20:*** *The process for identifying the criteria and standards to evaluate an evaluand is a critical aspect of evaluation. It involves thoughtful dialogue among diverse stakeholders. For what constitutes a 'good' or 'quality' evaluand in a particular context is often a matter of much debate (F1 +3; F2 +3; F3 +3).*

A participant associated with orientation two explained that the contextual nature of evaluation requires them to talk with stakeholders in order to understand and identify relevant criteria:

> Whoever you are doing work for, and regardless of the objectives of the evaluation, you will be operating in a context. How you define what success looks like in that context is essential and it needs to be identified through dialogue with stakeholders (2).

In light of the importance placed on evaluative criteria (standards) to define quality, it is unsurprising that all three orientations reject (to a greater or lesser extent) the notion of the evaluator interpreting data in an intuitive manner rather than relying on criteria and standards (item 29: F1 -2, F2 -4, F3 -1).

Third, the three orientations endorse the purpose of evaluation as being to produce explicit evaluative judgments (item 4) and confirm the role of the evaluator in providing such judgments, rather than stakeholders (item 5). A participant associated with orientation one described evaluative judgments as “the added value that evaluation provides” (345), while a participant associated with orientation three said: “What’s the point of doing an evaluation if the evaluator doesn’t provide an evaluative judgment?” (5).

> ***Item 4:*** *It is the primary responsibility of stakeholders, not the evaluator, to make evaluative judgments. The evaluator should only describe and report the various perspectives about the evaluand and make descriptive statements such as ‘if you value A, then B is the case’ (F1 -2, F2 -3, F3 -2)*
>
> ***Item 5:*** *The evaluator should not provide any assessments of an evaluand’s quality or performance. Instead she should give the information she has gathered about the evaluand to those who want to assess its operations or achievements (F1 -3; F2 -3; F3 -3).*

Lastly, the three orientations express (to a greater or lesser extent) the need for evaluators to be aware of how their personal values influence their perceptions of the evaluand, its context and stakeholder perspectives (item 16). According to a participant associated with orientation two: “It’s about always being humble about who you are and what you bring to a piece of work, recognising your biases and constantly checking yourself” (146).

> ***Item 16:*** *As evaluators, we need to understand how our value lens influences our perceptions of what we’re evaluating, its context, and how we understand stakeholders’ perspectives (F1 +2; F2 +3; F3 +3).*

6.4.3 Orientation one: the CONTEXT RESPONSIVE EVALUATOR with an eclectic approach

Orientation one describes an eclectic approach to evaluation and an evaluator who is flexible and responsive to the different contexts in which they work. The orientation one evaluator has a toolkit of evaluation methods and valuing approaches which they use according to context, people and circumstances (item 34 +4). Similarly, there are a variety of evaluator valuing roles which have different implications for the way an evaluation is conducted (item 33 +3). In the words of a participant associated with this orientation: "There is no one right way to do evaluation" (327).

The orientation one evaluator is both an idealist and a pragmatist. The evaluator wants their work to make a difference by helping to create a better world. However this desire has to be moderated by obligations to political masters, clients, stakeholders and informants. The balancing between aspiration and reality can be "tough" (item 6 +2*). A participant associated with orientation one said about item 6: "This is the motivation for me being an evaluator. But in reality, (evaluation) evidence isn't always as influential as it could be. This is the context we work in" (713). This obligation to multiple evaluation audiences expressed in item six is reinforced by the orientation one evaluator's sense of responsibility to the powerless who are the recipients of public services, and the general public (item 7 +4). For the orientation one evaluator, the impact of politics on public sector evaluation cannot be ignored, neither can the politics associated with the evaluand. A participant associated with orientation one described the evaluator's wide ranging and potentially conflicting responsibilities as follows:

> All the work we do is about evaluating government policy which usually includes people who don't have a lot of power. So we have a responsibility to everyone, you have to be working with everyone. You're responsible to multiple audiences - government is the client but the clients of the policy are the important ones, plus the general public, the taxpayer (332).

The orientation one evaluator places more emphasis than the other two orientations on the cultural norms, values and ways of knowing that are part of the context in which evaluators work (item 11 +3). A participant associated with orientation one commented on the need to think about culture in broad terms:

“We work in the midst of cultural environments, but it’s not just about ethnicity. It’s also about different organisational cultures” (332).

Similarly, the orientation one evaluator emphasises the importance of criteria and standards used in an evaluation to be made explicit (item 23: F1 -3). Like the orientation three evaluator, the orientation one evaluator rejects the idea for an evaluand to be measured against its objectives (thereby avoiding stakeholder debate about what quality means in relation to the evaluand) (item 19: F1 -2, F2 -2, F3 -1). For the orientation one evaluator such differences need to be made explicit.

The orientation one evaluator (like the orientation three evaluator) feels some constraint in their work due to the political context in which they work (item 10 +2). According to a participant associated with orientation one: “The real skill of an evaluator is managing these constraints” (327).

Of the three orientations, the orientation one evaluator provides the greatest rejection of the notion of the evaluator needing to be analytic, an empiricist, logician and dispassionate in order to arrive at an evaluative conclusion (item 28: F1 -3, F2 2*, F3 -2). Of the three orientations, the orientation one evaluator places less relevance on the need for public sector initiatives and their evaluation to focus on effectiveness and efficiency (item 18: F1 -2, F2 2*, F3 -1). Orientation one also places less emphasis than the other two orientations on defensible evaluative conclusions (item 25: F1 2, F2 4, F3 4).

In summary, the dominant themes that emerge about the orientation one evaluator is that they are contextually-sensitive and responsive. The orientation one evaluator has an eclectic approach. They use their toolkit of evaluation approaches and methods to respond to context which includes political influences, different evaluation participants, stakeholders and audiences, and other factors associated with a particular evaluand. The orientation one evaluator’s idealism is moderated by the political context in which they work and their obligations to political masters, clients, and stakeholders.

6.4.4 Orientation two: the ANALYTIC EVALUATOR focused on building a convincing case

Orientations two and three describe one of the evaluator’s core responsibilities as being to produce evaluative claims that are legitimate and justified (item 25: F2

+4, F3 +4). Both orientations reject the notion of evaluators providing judgments that are softly framed to enable readers to make their own interpretations (item 32: F2 -2, F3 -2). However, what is required to produce such judgments is described differently in the two orientations. The orientation two evaluator is cerebral. The task of developing evaluative conclusions requires them to be analytic, an empiricist, logician and dispassionate (item 28: F1 -3, F2 +2*, F3 -2). This is in contrast to orientations one and three which reject this description of the evaluator role. Given the evaluator's role as analyst, factor two strongly rejects the notion of data being interpreted in an intuitive manner (item 29: F2 -4).

The orientation two evaluator is concerned to build a convincing evaluation case using thick description (item 2) and argument (item 26). These are required to persuade an audience that the findings are plausible and reasonable (item 26: +2*). A participant associated with orientation two likened this to building a legal case: "It's like building a case like a lawyer . . . it has to stand up to rigorous inspection and debate" (228). Such argument engages the audience's reason and understanding (item 26: +2*). This focus on engaging and convincing the evaluation audience is evident in the endorsement of the use of thick description in an evaluation report (item 3: +2*). A participant associated with orientation two described such description as strengthening the "transparency of the evaluation case" (146).

Like orientations one and three, evaluative criteria and standards are deemed to be important and are defined through dialogue with stakeholders (item 20: F2 +3). The orientation two evaluator rejects the notion of the evaluator doing this by themselves (item 21, -2*): "You can't make a judgment without the stakeholders' input. In my view there's no way that you can avoid having stakeholders' input" (2). The orientation two evaluator endorses the approach of allowing criteria and standards to emerge during an evaluation as the evaluator builds their understanding of quality in relation to the evaluand and context (item 22: F2 +3).

The orientation two evaluator places importance on the evaluator surfacing implicit values and assumptions associated with an evaluand, such as those associated with an organisation (item 9: +4 distinguishing statement). Orientation two (like orientations one and three) rejects the notion that an independent and objective assessment requires the evaluator to be detached (item 1: F1 -4, F2 -3, F3 -4). For the orientation two evaluator, "an independent and objective assessment" described in item one and “a neutral, external perspective” described in item 9 require implicit values and assumptions to be surfaced (item 9: +4*). A participant associated with this orientation said:

> The idea of surfacing stuff that might not be explicit for people, or are not tangible, or are diffuse, but that still have a powerful influence – this is the opening of the door to designing evaluations. There has to be some sort of surfacing (2).

Another participant associated with orientation two explained that surfacing implicit values is important to enable “the non-captured perspectives to be put into the (evaluation) picture” (146).

Accountability-focussed evaluation of public sector initiatives is important for the orientation two evaluator. This is achieved by evaluations focussed on measuring effectiveness and efficiency (item 18: +2 distinguishing statement). A participant associated with orientation two expressed this viewpoint in the following way:

> Because you are a public sector evaluator, you have to be pragmatic. You are evaluating tax payers’ dollars so you need to provide information that people want, like value for money even though personally you might prefer a softer, qualitative approach. We shouldn’t shy away from putting a quantitative value on an evaluand, even though it is hard (713).

In contrast to orientations one and three where evaluators feel somewhat constrained by context and politics, the orientation two evaluator does not feel constrained (item 10: F1 2, F2 -2*, F3 2). Further, the orientation two evaluator does not feel responsibility to audiences with unequal power dynamics to the same extent as orientations one and three evaluators (item 7: F1 4, F2 1, F3 4). The orientation two evaluator places less emphasis than the other orientations on the need for a variety of valuing approaches to match with people and context (item 34: F1 4, F2 -1*, F3 1).

In summary, there are three defining features of the orientation two evaluator. The first is their analytic, empirical and dispassionate approach which provides the greatest point of difference with the other two orientations. The second is the emphasis on building a convincing and defensible case through argument which will engage evaluation audiences. A third point of difference is the orientation two evaluator's focus on accountability measures such as cost benefit analysis which are given less importance by the other two orientations.

6.4.5 Orientation three: the JUDGMENT-CENTRED EVALUATOR using inclusive practices to create defensible judgments

The orientation three evaluator places more emphasis on judgment-making as being fundamental to the practice of evaluation than the other two orientations. The orientation three evaluator provides the strongest rejection of the statement that evaluation involves describing, explaining and informing rather than judging (item 35: F1 -1, F2 0, F3 -2*). Two participants associated with orientation three described the centrality of judgment-making to evaluation: "Judging is the essence of evaluation . . . it's the critical thing that evaluation does" (824). "The point of evaluation is to attribute value. It's a deliberate process. Some suggest it's a matter of describing. It's not, it's about judging" (589). Another participant associated with orientation three made the following comment about the contribution of evaluative judgments: "Making descriptive statements (about value) is a cop out. Evaluative judgements are about shining light and providing value" (345).

Orientations two and three place equal importance on evaluators producing evaluative claims that are legitimate, justified and defensible (item 25: F2 +4; F3 +4). Such claims require transparency as highlighted by a participant associated with orientation three: "You have to be able to show how you got to where you've got to" (65). Another participant stressed the importance of transparency for the credibility of the evaluation and their team's work:

> As an evaluator I need to be able to defend the claims and conclusion I have come to. People need to be able to see clearly what is underlying those claims and be able to see I have followed a logical approach based on the evidence . . . it's important for the credibility of our work and the (evaluation) unit (384).

While orientations two and three place equal importance on evaluators producing evaluative claims that are legitimate, justified and defensible (item 25: F2 +4; F3 +4), orientation three identifies different requirements for defensible judgments to those identified in orientation two. The orientation three evaluator stresses the need for inclusive practices. For an evaluative conclusion/judgment to be defensible, it must be based on the perspectives of multiple stakeholders through the use of inclusive processes (item 27: F3 +3). Such processes are described by two participants associated with this orientation: "An evaluation has to be able to draw on a range of stakeholders to be able to draw conclusions" (18). "Multiple perspectives help to come to a defensible judgment. The defensible word is important for (name of government agency) because the evaluation report is going out to multiple audiences and has serious implications for those being evaluated" (195).

The evaluator's role is to produce the evaluative judgment, not stakeholders (item 3: F3 -3*; item 4: F3 -2; item 5: F3 -3). Having considered the perspectives from multiple interests, the evaluator examines them against the criteria to produce an all-things-considered judgment (item 30: F3 +2*).

While all of the orientations endorse (to a lesser of greater extent) the need for embedded values to be made explicit, orientation three places more emphasis than the other orientations on the value-laden nature of evaluation theory, methods and practice (item 13: F1 +1, F2 0, F3 +2), (item 14: F1: 0, F2 +1, F3 +2). Referring to item 13 about the value-laden nature of evaluation, a participant associated with orientation three reflected: Evaluation stands with me - I am one of the key components and players in the evaluation. So my values, training, the methods I am comfortable with, will all influence how I approach an evaluation" (384).

As noted above, factors one and three are strongly correlated (0.8485). Orientation three shares three themes with orientation one. Evaluators' responsibilities to multiple audiences with unequal power dynamics is emphasised in both orientations (item 7: F1 +4; F3 +4). Orientation one and three evaluators share a feeling of constraint in their work, as a result of context and politics (item 10: F1 +2; F3 +2). The third common theme is the need for the standards used in an evaluation to be made explicit (item 23: F1 -3; F3 -3).

In summary, the orientation three evaluator is distinguished from the other two orientations through their emphasis on the centrality of judgment-making to the practice of evaluation and the need for such judgments to be defensible. For the orientation three evaluator, such defensibility is achieved through inclusive processes that gather the views of multiple stakeholders.

6.5 Discussion of the orientations

6.5.1 Introduction

This section examines the orientations as a whole to identify what the Q results suggest about evaluative reasoning in public sector evaluation in Aotearoa New Zealand.

I was initially concerned about the three items with identical scores and the three items with similar scores across the three orientations. I questioned whether this was the result of insufficient choice of items for participants and/or that the participant group lacked heterogeneity. My concern was allayed as a result of re-visiting the work of Abma (2006) and Abma and Widdershoven (2008, 2011) on social relations in evaluation, discussing the orientations with three Q participants who responded to my invitation to discuss the orientations, and feedback from evaluation colleagues during the presentation of the findings at the ANZEA and WEG workshops.

6.5.2 A shared epistemology

This section will argue that the commonality across the three orientations can be explained by a shared epistemology within which there are differences in evaluator understanding and practice of evaluative reasoning, as represented by the three orientations.

Abma's (2006) work on social relations between the evaluator and others in the evaluation setting provides a theoretical framework for interpreting the orientations. Abma's (2006) schema shows how social relations differ in the four major contemporary evaluation traditions of postpositivism, utilitarian pragmatism, critical participatory theory, and interpretivism. In the postpositivitist tradition, the evaluator is the expert and "social relations are irrelevant or sources of error . . . (consequently) the evaluator adopts a distanced relationship" (Abma, 2006, p.192). Turning to the Q results, the six items with identical or almost identical scores across the three orientations can be interpreted as a rejection of postpositivist evaluation. This is evidenced by the rejection of statements stating that an objective assessment requires the evaluator to be detached from the evaluand (item 1: F1 -4, F2 -3, F3 -4), and stakeholders should be kept at arms' length from the evaluation process (item 2: F1 -4, F2 -4, F3 -4). Rejection of the postpositivist approach is also evident in the endorsement of items 16 and 17 (item 16: F1 2, F2 3, F3 3; item 17: F1 1, F2 0, F3 0) describing the contextual nature of valuing, and the need to surface and articulate values in the evaluation process.

Referring again to Abma's (2006) schema, the relational nature of evaluation practice expressed across the three orientations aligns to the interpretivist evaluation tradition. According to this tradition, social relations are "no longer a means to another end . . . but rather are intrinsically valuable" (p.195) and "evaluation is not only a technique but also a social practice" (p.196). The evaluator is not an "expert" but is a "facilitator" (p.195). These features are evident across the three orientations, specifically in the endorsement of stakeholders' contribution to the evaluation process (item 2) and the confirmation of a dialogic approach to developing evaluative criteria and standards (item 20). The interpretivist perspective is also demonstrated in orientations one and three which endorse evaluators' responsibility to audiences with asymmetric power dynamics (item 7). It is also evident in Q participants' comments (made after the Q sort) explaining why stakeholder involvement in the evaluation process is important (reported in section 6.4.2). Most significantly, the intrinsic value of human relationships underpins Te Ao Māori and Pasifika worldviews. Q participants do not regard themselves as the expert in respect of the evaluand, but instead value stakeholders' knowledge of the evaluand and its context.

This positioning within the interpretivist tradition provides an overarching epistemological framework within which differences in evaluator understanding and practice of evaluative reasoning exist, as represented by the three orientations. The correlation scores between the three orientations indicate that the differences expressed in the orientations are about focus, nuance and emphasis, rather than dissimilarities of a more fundamental nature. Orientation one emphasises the importance of context-appropriate evaluation approaches, while orientation three stresses the centrality of judgment-making to evaluation. For the orientation three evaluator, evaluative judgments are made defensible through the use of inclusive practices. For the orientation two evaluator, such defensibility comes from an argument that will convince the evaluation audience. Of the three orientations, orientation two provides the greatest points of difference through its description of the evaluator role as a dispassionate analyst, emphasis on building a convincing evaluative argument, and accountability focus.

Feedback from evaluation colleagues during the presentation of the findings at the ANZEA and WEG seminars confirmed the relative homogeneity of evaluation practice in Aotearoa New Zealand suggested by the Q results. A public sector evaluator who attended one of these seminars observed: "The evaluation community is small, with only one degree of separation between us. As a result, ideas flow through the profession very quickly". At another seminar, Dr Jane Davidson, a high profile New Zealand evaluator described New Zealand evaluators as generalists due to limited work opportunities in specialist areas which has the effect of reducing diversity of practice.

6.5.3 Evaluative reasoning endorsed across orientations

The aim of the Q study was to gain insights into how professionals who undertake public sector evaluation understand evaluative reasoning. Before undertaking the Q study, I thought there may be at least one orientation which demonstrated less understanding or placed less emphasis on evaluative reasoning than other orientations. This was not the case. The three orientations endorse the purpose of evaluation and the evaluator role as making defensible evaluative judgments (items 4, 5 and 25) (albeit the endorsement is stronger in some orientations than others). Other elements of evaluative reasoning are evident across the three orientations, particularly recognition of the "values-imbued" (House, 2004b, p.7) nature of evaluative activity. Each of the orientations give emphasis to a particular value that influences evaluation: cultural values (orientation one), institutional values (orientation two), values implicit in evaluation theory (orientation three). The three orientations place equal importance on the role of evaluative criteria/standards in creating evaluative judgments and the importance of dialogic processes in their development (item 20), while orientations one and three identify the need for explicit standards (item 23). Taken overall, the three orientations can be regarded as being attuned to evaluative reasoning and its implications for practice.

6.5.4 Other insights from the Q sort

Lastly, two additional insights emerged from participants' comments after completing a sort. Firstly, some participants spoke about the role of context when considering some items. They described their difficulty in sorting some items without a context in which to place the statement. They made comments to the effect that they would agree with the item in a certain context, but disagree in another context. Such comments are reminiscent of Patton's observation that "valuing must be understood as contextually embedded and dependent (2012, p.98). The second insight is a reminder about the importance of language in our meaning-making. For a Māori participant, the culturally embedded meanings of the words argument (item 26) and defensible (item 27) raised negative connotations of power and control. A Pasifika participant said that a good evaluator should not argue (item 26) as this implies conflict, but should rather discuss.

6.5.5 Towards a hypothesis

The differences evident in the three orientations require further examination - what are the possible reasons for such difference? To assist in explaining such differences, orientation one is referred to as the *context responsive evaluator,* orientation two as the *analytic evaluator,* and orientation three as the *judgment-centred evaluator.*

Before discussing possible reasons for the differences among the three orientations, it is first necessary to revisit Q methodology. As noted in Table 6.1, in Q a sample of items is sorted by a collection of people. Q involves the correlation and factoring of people to reveal factors of similar viewpoints on a particular topic or "attitudes of mind held in common" (Wolf, Q Methodology Network, 9 March 2015). Therefore (unlike R), the focus of Q is on the observed variation in the sample of items, not on the Q participants. This means that "factors are not groups of people, and a person cannot be said to 'belong' to a factor" (Wolf, Q Methodology Network, 9 March 2015).

However, the Q researcher may look to the demographics of the people who are associated with a particular factor to aid understanding of the factor. Watts and Stenner (2012) caution that this should be done at the end of the interpretation stage so that "each factor array is approached on its own terms and prevents our succumbing to the temptations of preconception and expectation" (p.157). Therefore the use of participants' demographics is not a primary interpretative tool as in R.

Examination of the demographic information about the Q participants who are statistically significantly associated with one or more orientations (n=28) may provide some insight into the observed differences among the three orientations.

Consultants are more likely to be associated with orientation one (the context responsive evaluator) (n=6) than any other orientation (orientation two n=2; orientation three n=1). (Consultants include professionals working independently or as an employee in a private research organisation or university). Turning to orientation three (the judgment centred evaluator) - thirteen of the 18 government employees are associated with this orientation (orientation three only n=7; orientation three and one n=3; orientation three and two n=3). It is hypothesised that the differences in nuance and emphasis observed in orientations one and three may be explained by where the evaluator is located, inside or outside government. This is now explained.

Turning first to the context responsive evaluator - this evaluator is eclectic, contextually sensitive and responsive. They use their toolkit of evaluation methods and valuing approaches based on what will work best in the situation at hand. This is congruent with what is involved in working as an evaluation consultant - a professional who is contextually-sensitive, flexible, and multi-skilled. Obligation to political masters, clients, stakeholders and the powerless recipients of government programmes is more pronounced for the orientation one evaluator than the other two orientations. The evaluation consultant role involves balancing multiple responsibilities and obligations - to the government agency as the evaluation funder, the policy makers or other immediate users of the evaluation, and the communities and programme recipients participating in an evaluation.

While the context responsive evaluator and the judgment centred evaluator have areas in common, the latter has different priorities. Making judgments about value is central to the practice of evaluation for the judgment centred evaluator, and the transparency and defensibility of such evaluative judgments are achieved through stakeholder involvement and inclusive practices. Their focus on judgment-making can be seen as congruent with the role of public sector evaluation in assessing public policy implementation and its impacts. The judgment centred evaluator’s emphasis on transparency and defensibility is also congruent with the public sector context, given the far-reaching consequences of decisions based on evaluative findings and the political and public scrutiny of such decisions. The judgment centred evaluator’s focus on stakeholder involvement and inclusive evaluative practices may reflect the intent of government agencies to include stakeholders in policy development and its assessment.

It is therefore hypothesised that the differences in nuance and emphasis evident in orientations one and three are associated with where the evaluator stands in relation to the public sector - either as a consultant outsider, or as a government employee insider. The orientations suggest that each has different priorities for, and emphasises in their evaluative reasoning practice.

Orientation two does not suggest any role-related influences. (Of the six participants associated with orientation two, three are government employees, two are consultants, and one works for a non-government organisation). As noted in section 6.3.6, the correlations of orientation two with orientation one (0.5895) and orientation two with orientation three (0.5705) are less than the correlation between orientations one and three (0.8485). This indicates that the analytic evaluator (orientation two) is significantly different to the context responsive evaluator and judgment centred evaluator, with evaluator role having no obvious influence. It is hypothesised that this difference is about their analytic and systematic approach to evaluation. While the analytic evaluator and the judgment centred evaluator both emphasise the need for legitimate and justified evaluative claims, the analytic evaluator is more explicit about what is involved in producing such claims, namely, through building an argument-based evaluative case that will withstand scrutiny. This suggests an evaluator who approaches their work in a considered and careful manner with a strong emphasis on sound reasoning.

6.6 Conclusion

This chapter has provided the first of three perspectives about evaluative reasoning and its practice in a public sector context. Q methodology, an abductive approach, enables statistically significant factors (orientations) to be identified which describe shared viewpoints. Three orientations were revealed which had themes in common. It was argued that this commonality can be explained by a shared epistemology within which there are differences in evaluator understanding and practice of evaluative reasoning. Variations in the way evaluative reasoning is described in the three orientations are about emphasis and nuance, rather than significant differences. It is hypothesised that the differences observed in orientations one and three may be explained by where the evaluator stands in relation to the public sector - either as a consultant outsider, or as a government employee insider. Taken overall, the three orientations can be regarded as being attuned to evaluative reasoning and its implications for public sector evaluation practice.

CHAPTER 7

PERSPECTIVE TWO: META-EVALUATION

7.1 Purpose

While the first perspective provided by the Q methodology study (reported in chapter 6) aimed to gain insight into how evaluative reasoning is understood by evaluators undertaking public sector evaluation, this chapter provides a different perspective, namely, a snapshot of how evaluative reasoning is being practised. It is based on a meta-evaluation of 30 evaluation reports in the public domain which were conducted or commissioned by 20 central government agencies in Aotearoa New Zealand during the period 2010-2013.

Chapters 4 and 5 set out the key elements identified in informal logic (Hare, 1967; Rescher, 1969; Scriven, 1967; Taylor, 1961) which make it logically possible to reason about values. These elements have been further explicated by a number of evaluation theorists since the early 1980s, predominantly from the United States, including Davidson (2005), Fournier (1995), Fournier and Smith (1993), Greene (2011), House (1977, 1980, 1996), House and Howe (1999), Julnes (2012b), Scriven (1980, 1994, 1995, 2007a), Schwandt (1997, 2001, 2008b), and Stake (2003, 2004). The work of these theorists has built a body of knowledge about what is required to reason from a value to an evaluative conclusion that is legitimate and robust, referred to as evaluative reasoning (House & Howe, 1999, p.xvi).

7.2 Research objective

The research objective for this part of the study is to understand the practice of evaluative reasoning, as evident in a sample of public sector evaluation reports written by Aotearoa New Zealand evaluators, by examining the presence of the key elements of evaluative reasoning in the reports. My aim is not to examine the quality of these elements in collectively building a robust evaluative case, but rather to find evidence of their presence. Given my interest in identifying a range of practice, I decided to look for evidence of the key evaluative reasoning elements in a large sample of reports rather than examining a fewer number of reports in greater detail.

The term *evaluator* is used in this chapter to describe a professional who has conducted an evaluation for a public sector agency, whether or not the professional identifies as an evaluator.

7.3 Meta-evaluation

Meta-evaluation is a method for assessing evaluation quality. The term meta-evaluation was first used by Scriven who defined it as any evaluation of an evaluation (or set of evaluations), evaluation system, or evaluation device (Scriven, 1969, cited in Stufflebeam, 2001b; Scriven, 2004). In a more recent paper, Scriven (2009) describes meta-evaluation as being “the consultant’s version of a peer review” (p.iv). The purpose of a meta-evaluation is to assess (usually by an objective third party) an evaluation report to determine (i) how well it meets the requirements of a sound evaluation (merit) and, (ii) the extent to which it meets the audience’s needs for evaluative information (worth) (Stufflebeam, 2001b). Meta-evaluations may be formative (to improve an evaluation while it is being conducted) or summative (to identify the strengths and weaknesses of a completed evaluation) (Cooksy & Caracelli, 2009). The 2011 edition of the Joint Committee standards has two standards relating to meta-evaluation. The first encourages evaluators to evaluate their own work. The second is targeted at groups who commission evaluations and evaluation stakeholders, and promotes the conduct of meta-evaluations.

> Evaluation accountability standard E2 - Internal Meta-evaluation. Evaluators should use these and other applicable standards to examine the accountability of the evaluation design, procedures employed, information collected and outcomes (p.237).
>
> Evaluation accountability standard E3 - External Meta-evaluation. Program evaluation sponsors, clients, evaluators, and other stakeholders should encourage the conduct of external meta-evaluations using these and other applicable standards (p.245).

Stufflebeam (2001b) favours the Program Evaluation Standards (Joint Committee, 1994) and the Guiding Principles of the American Evaluation Association (1995) as the evaluative criteria for meta-evaluations, and bases his meta-evaluation checklist on them (Stufflebeam, 1999). Other criteria may be used such as published standards or principles (Stufflebeam, 2001b) or evaluation checklists (for example, Scriven, 2007b). Scriven (2009) recommends that the meta-evaluator uses a different evaluation checklist than that used by the original evaluator. Cooksy and Caracelli (2009) examined eighteen meta-evaluations conducted in the United States. They found that only five used the Joint Committee standards, while seven used emergent criteria, three used criteria specifically identified for the evaluation, and three were assessed against the trustworthiness criterion of confirmability and dependability. Davidson (n.d., cited in Scriven, 2009) expresses caution about the use of checklists as meta-evaluation criteria. She points out that their use assumes their validity as a measurement tool (through having undergone appropriate calibration and testing for measurement purposes). Even if a checklist is deemed to be valid, its perfunctory application may fail to identify a significant issue about an evaluation.

Given the diverse ways in which evaluation quality may be understood, it is important to involve the evaluation client and stakeholders in defining the meta-evaluation criteria (Cooksy & Caracelli, 2005; Stufflebeam, 2001b). Based on their experience of conducting a meta-evaluation of 87 evaluation reports produced by an agricultural research organisation, Cooksy and Caracelli (2005) emphasise the need for meta-evaluators to be cognisant of the political and cultural context in which a meta-evaluation takes place. They advise that “the criteria selected be tailored to the purpose of the meta-evaluation and to the culture and sensibilities of the meta-evaluation’s stakeholders” (p.35).

Finally, the important role of meta-evaluation in evaluative practice is stressed by Scriven and Stufflebeam. Scriven (2004) asserts that meta-evaluation is part of the reflexive nature of evaluation practice. In a paper entitled "The meta-evaluation imperative" Stufflebeam (2001b) describes meta-evaluation as "a professional obligation of evaluators" (p.183).

7.4 Meta-evaluation criteria

The meta-criteria used in this study are derived from Scriven's general logic of evaluation (1980), Fournier's working logic (1995), the work of Fournier and Smith (1993) on evaluative argument, and N. L. Smith's (1981, 1987) work on claims and judgments. Initially, I identified four meta-criteria as follows: (i) criteria or other comparator, (ii) standards for the criteria, (iii) warranted argument, and (iv) evaluative conclusion/judgment. (Each of these meta-criteria is explained in greater detail in section 7.8). I had assumed that the evaluation reports would have at least one **evaluative** evaluation objective (or, in the absence of objectives, at least one evaluative evaluation question). In this context an evaluative evaluation objective/question is distinguished from other types such as causal, explanatory or descriptive objectives/questions. However on the first reading of the reports I realised an additional meta-criteria was required, that is, one or more evaluation objectives that are evaluative (or in the absence of objectives, at least one evaluative evaluation question). (Hereinafter, the term *elements* is used rather than meta-criteria, so as to avoid confusion with the term criteria).

As Figure 6 illustrates, these five elements are interconnected and together build a transparent chain of reasoning (Scriven, 1976) or coherent case from which an evaluative conclusion/judgment can be drawn that is legitimate and defensible (Fournier, 1995; Fournier & Smith, 1993). Each of these five elements is required in an evaluation regardless of the working logic or evaluation approach/model used (Fournier, 1995). The interconnected nature of these elements is described in summary as follows (section 7.8 which describes the meta-evaluation findings illustrates this interconnectedness in greater detail). The evaluation objectives articulate the focus of the evaluation, and determine the type of claims that will be produced (N. L. Smith, 1987). They also determine the choice of the criteria (or other comparator) against which the evaluand is to be assessed. The criteria determine the evidence to be collected and the type of warrant used (Fournier, 1995). Standards are required for the criteria (or other comparator) to identify and describe levels of performance. A warranted argument is required to support and strengthen the evaluative claims about the performance of the evaluand (i.e. the evidence) in relation to the criteria and standards (Fournier & Smith, 1993). The warranted argument sets out and builds the case from which an evaluative conclusion/judgment can be drawn that is legitimate and defensible (Fournier & Smith, 1993). The five elements are located within a context. The role of context in evaluative reasoning is described in section 7.5.

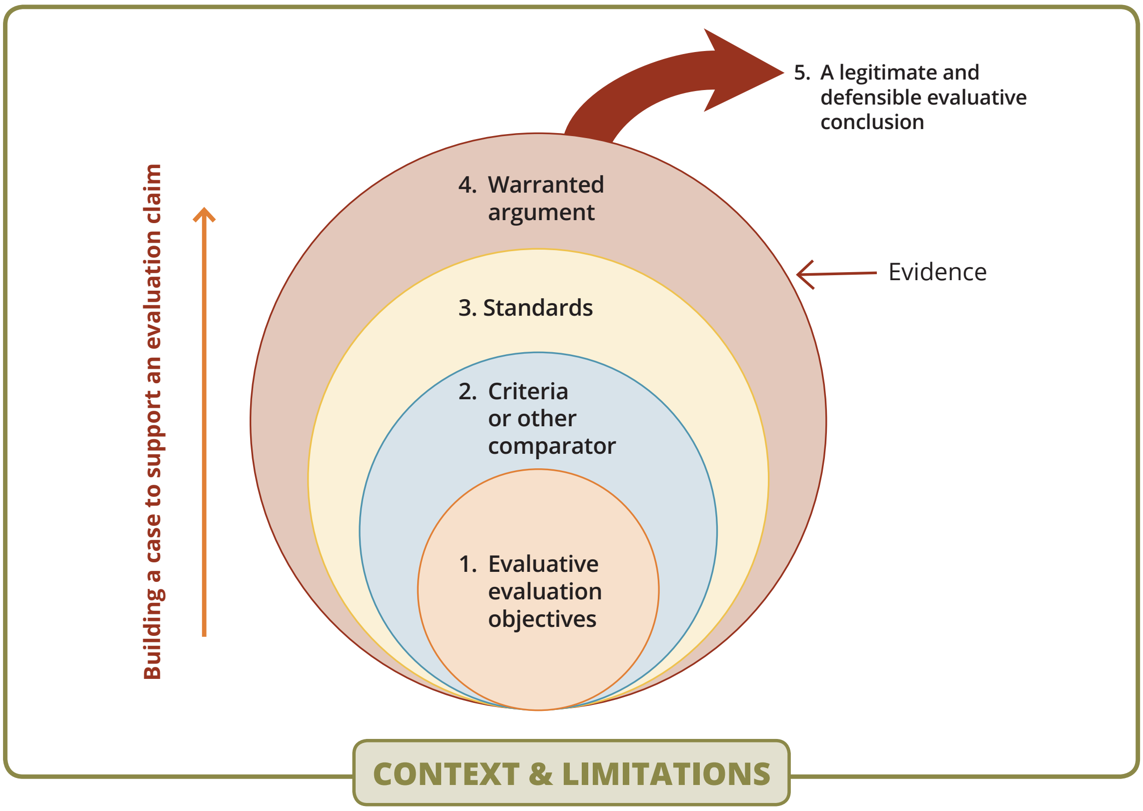

5. A legitimate and defensible evaluative conclusion
Building a case to support an evaluation claim
4. Warranted argument
Evidence
3. Standards
2. Criteria or other comparator
1. Evaluative evaluation objectives
CONTEXT & LIMITATIONS

Figure 6 Building a case to support an evaluative claim - elements of evaluative reasoning

The five elements and their standards are shown in Table 7.1. As noted above, my aim was to find evidence of the presence of the five elements in the evaluation reports (therefore excluding such things as an assessment of the quality of these elements; the appropriateness of the evidence for the criteria; and the appropriateness of the warrant for the evaluation audience and context). Therefore the standard for each of the elements is whether there is evidence of the element in the report, or not. While this assessment was straightforward for elements 1, 2, 3 and 5, it was less straightforward for element 4: warranted argument. The definition of warranted argument used in this study is based on that of Booth, Colomb, and Williams (2008) who describe a research argument as consisting of five components: (i) a claim, (ii) reasons that support the claim, (iii) evidence that supports the reasons, (iv) an acknowledgment of and response to alternatives/complications/objections, and (v) a principle which makes the reasons relevant to the claim (a warrant). I examined the argument in each report to determine whether these components were addressed.

Table 7.1 Elements and their standards

Element	**Standard**
1. Evaluative evaluation objectives (or questions if there are no objectives)	Evidence of one or more evaluation objectives that are evaluative.
2. Criteria or other comparator	Evidence of criteria (or another form of comparator) against which the evaluand is assessed. Evidence of a justification for the criteria (or other comparator).
3. Standards	Evidence of standards or benchmarks of performance for the criteria (or other comparator).
4. Warranted argument	Evidence of a warranted argument to support the evaluative claim(s).
5. Evaluative conclusion/ judgment	Evidence of one or more evaluative conclusions/judgments. If there is a synthesised conclusion/judgment, an explanation of how the conclusions/judgments were synthesised.

It is important to stress that in using these five elements as the meta-criteria, it is not my intention to simplify or reduce evaluative reasoning to an easy to master technique. The complexity that is evaluative reasoning is acknowledged. At the least, it involves careful listening (Abma, 2006), perceptive consideration of stakeholder perspectives (Stake, 2004), critical thinking (Schwandt, 2002b), considered deliberation and argumentation (House, 1977), reflexivity about personal values and their impact on the deliberative process (Greene, 2011), sensitivity to bias (Denzin & Lincoln, 2011b), and astute judgment-making (Scriven, 1994a). However the five elements I have identified provide a framework around which such complexity is built.

7.5 Contextual information

During the design of this study, I was mindful of the influence of context on evaluative reasoning as described in chapter 5. The impact of context on evaluation practice is reflected in Shadish's notion of contingency theories of evaluation practice (1998), an area which appears to remain undeveloped. Shadish (1998) describes evaluation practice as involving trade-offs, requiring evaluators " . . . to make choices based on the contingencies of the situation" (p.8). He notes that most evaluation theorists do not address contingency issues in the explication of their theory. This means that a particular evaluation theory expresses an ideal, rather than addressing the realities of applying the theory in real-life situations. At a practice level, Bamberger, Rugh and Mabry (2006) provide insights into the trade-offs that may occur doing what they refer to as "real world evaluation" (p.xxix). Such trade-offs arise from budget, time or data constraints, and political influences, and include, for example, compromises about evaluation design, sample size, and the way in which methods are used.

Given the importance of context in evaluative inquiry, I scrutinised each of the reports for contextual information to inform my understanding of the factors that influenced the design and conduct of the evaluation. Such information included the evaluation purpose, the audiences for and intended uses of the evaluation findings, information about the evaluand and context, and the methods used. I also examined information about limitations in the report to understand the contextual factors that may have impacted on the evaluation and its valuing approach. I recorded this information about individual reports to provide insight about the elements and their application.

7.6 *Sampling strategy*

My original intention was to collect the sample of evaluation reports by requesting copies from evaluation managers in government agencies. However while conducting the Q study, I realised that the political scrutiny on public sector agencies at that time (2013) might not support this approach. The then Government had stated its intention to downsize and rationalise the public service, and the restructuring of research and evaluation teams in several agencies had been signalled or was already underway. My impression from talking with the 18 Q participants who are public sector employees was that evaluation teams felt under the spotlight and may be reluctant to participate in research that would put their work under external scrutiny. Consequently, I decided to focus on evaluation reports in the public domain.

I examined the websites of 63 central government agencies, local body agencies (such as city and regional councils), and other government organisations (such as district health boards). I restricted my web search to reports dated 2010-2013 to ensure their currency. Evaluation reports were available on the websites of 22 agencies, while the websites of 41 agencies either had summaries of evaluation reports, reports dated pre-2010, or contained no reports.[11] Since I was able to find only two reports from local body and other non-central government agencies, I removed these organisations from the sample frame.

From the 52 evaluation reports collected, I chose 30 reports based on the following criteria:

(i) The reports were commissioned and/or funded by central government agencies.

(ii) The reports were written by New Zealand-based authors.

(iii) No author appears twice in the sample.

(iv) The report is a complete evaluation report rather than a summary of key findings. (This is discussed further below).

[11] I was unable to locate evaluation reports on the website of the New Zealand Qualifications Authority which manages the New Zealand Qualifications Framework and undertakes quality assurance of non-university education providers. I was subsequently advised that evaluation reports have been on the Authority’s website since 2009.

(v) The sample includes a range of evaluand-types (for example, policy, programme, media campaign, strategy) and evaluation approaches (for example, development, developmental, economic, implementation, outcomes, impact).

(vi) There are no more than two reports per agency. (An exception is the Ministry of Health. There are three reports funded by the Ministry in the sample, two of which are health impact assessments commissioned by another agency).

The sample consists of evaluation reports commissioned and/or funded by 20 public agencies (three of these agencies have since been disestablished - two have been combined into a new agency, and another subsumed into an existing agency). Eleven reports have authors who are employed by the agency (referred to as internal authors. It is assumed the authors of the seven reports with anonymous authors are also agency employees). Nineteen reports have authors who work outside of the agency (referred to as external authors). The sample of reports is summarised in Appendix F.

There is a range of evaluand-types in the sample. Social and educational programmes are the most common (17 reports), policies (4 reports), interventions (4 reports), an aid strategy (1 report), a media strategy (1 report), governance arrangements (1 report), a road construction project (1 report), and research use (1 report).

The sample includes five economic evaluations - two are value for money assessments (reports 30, 41), two are cost benefit analyses (reports 21, 42) and one is a cost effectiveness analysis (report 14). In two of these five evaluations (reports 14, 41), the economic assessment is one of two or more methods used in the evaluation.

The sample also includes two health impact assessments funded by the Ministry of Health (reports 43, 44). Health impact assessment (HIA) is defined as "a combination of procedures, methods and tools by which a policy, program or project may be judged by its potential effects on the health of a population, and the distribution of those effects within the population" (Fehr, Viliani, Nowacki & Martuzzi, 2014, p.89). It consists of four steps, of which the final step is an evaluation of the completed HIA (R. Cunningham, Signal & Bowers, 2010).

I assigned an evaluation orientation (Chelimsky, 2006) to each of the 30 reports based on its evaluation purpose statement or other information, namely, accountability (the measurement of results or efficiency), development (the provision of evaluative help to strengthen institutions), knowledge (the acquisition of a more profound understanding in some specific area or field), and management (for oversight and/or improvement purposes).

The authors of the 30 reports in the sample work in a range of professions including civil engineering, economics, health and management. Authors include education consultants, psychologists, health professionals, academics, and evaluation practitioners.

In two reports (8, 43) the person responsible for designing and/or implementing the programme also evaluated the programme.

7.7 Coding and analysis approach

I developed a recording sheet consisting of the headings of the five elements and the areas of interest about an evaluation report described above (Appendix G). As I read a report, I wrote qualitative comments about each element on the report's recording sheet, including excerpts from the report illustrating particular points of interest, questions for further consideration, and any discrepancies. Once I had read all of the reports I did a second reading, this time focussing on evaluative reasoning as a coherent "chain of reasoning" (Schwandt, 2008a, p.146) unfolding through the report.

I then coded these comments onto an Excel spreadsheet using a primary coding framework based on the five elements. As I coded, I created sub-codes within the primary codes to record particular aspects. I usually use NVivo for coding qualitative data but I decided not to use it for the meta-evaluation because I felt I needed to eyeball the whole dataset in a visual way. (Excel allows the researcher to do this in a way that NVivo does not). Using the Excel spreadsheet as my primary source of analysis, I examined each of the evaluative reasoning elements to understand how the authors had applied (or not applied) them. During this process I revisited individual reports as necessary.

Taking each of the elements in turn, I first analysed the findings about the element to assess whether, and how, it had been applied, before grouping the reports into three groups: reports with five elements of evaluative reasoning, reports with three or four elements, and reports with two or fewer elements. I examined the reports in each group to determine whether any patterns were evident.

7.8 Findings

This section begins by describing the evaluation contexts for the 30 evaluation reports in the sample. The findings for 28 of the 30 reports are then presented for each of the five elements of evaluative reasoning.[12] A short description about each element precedes the findings which build on the overview of the elements presented in Chapter 5.

7.8.1 Evaluation contexts

The five elements of evaluative reasoning provide a coherent framework to guide the evaluator's work. However as any experienced practitioner will attest, applying these elements in a practical context can be less than straightforward. This section summarises the contexts of the 30 evaluations. It illustrates the diversity of contexts in which public sector evaluation takes place and therefore the range of influences and constraints on the evaluator's work, which in turn may influence evaluative reasoning practices.

Seventeen evaluations were conducted in community settings in Aotearoa New Zealand, including communities with high Māori populations, people living in temporary accommodation whose homes had been destroyed in the Canterbury earthquakes in 2010-2011, vulnerable parents in their homes, and in marae-based courts, workplaces, therapeutic communities and schools. Two evaluations of New Zealand-funded aid projects were conducted in community settings in Pacific Island countries. Such community-based settings often involve challenges for the evaluator, such as issues of respondent accessibility, time constraints, and resource availability. Compromises and trade-offs may have to be made which may have had an impact on the evaluation, and therefore on evaluative reasoning

[12] This section includes the two value for money assessments (reports 30, 41) and the cost effectiveness analysis (report 14). It does not include the findings of the review of the two cost benefit analyses (21, 42) which are reported separately in section 7.9.7.

practice. This is illustrated vividly in report 12, an evaluation of a New Zealand-funded aid strategy in a Pacific state. Arrangements had been made for the New Zealand-based evaluator to work with an in-country evaluator but this person was re-assigned at the last minute. The evaluator discovered other aspects that affected their ability to do the evaluation, such as poor quality documentation requiring a significant amount of investigative and back-filling work. These issues were compounded by the evaluator having only two weeks in-country. Such constraints require flexibility and adaptability on the evaluator's part to ensure the optimal quality of the evaluative reasoning.

In contrast to the 19 community-based evaluations, six evaluations were primarily desk-based using either secondary data (four reports), or secondary data with minimal additional qualitative data (2 reports). The remaining five evaluations were done in organisational settings.

Based on Chelimsky's (2006) evaluation orientations, the majority (19) of the 30 reports have a management orientation, 10 reports have an accountability orientation, one has a development purpose, while no reports have a knowledge orientation. The emphasis on instrumental purposes is not surprising given the current Government's focus on efficiency and effectiveness of public expenditure in the post-global financial crisis era.

I examined the information about limitations in the reports to understand the impact of contextual and other constraints which may have impacted on the evaluator's work. Half of the 28 reports contained no information about the limitations associated with the evaluation. The absence of such information restricted understanding of the contextual and other factors that may have influenced the design and conduct of the evaluation. In a university context it is expected research practice to identify the limitations associated with an inquiry. It is possible that some of the authors of the reports in the meta-evaluation are not trained in research methods.

7.8.2 Element one: Evaluation objectives

The evaluation objectives focus the inquiry by providing the foundations for how the evaluation is designed, conducted and reported. They determine the values to be examined, the criteria and standards to be selected, the data to be collected, and the nature of the argumentation required to support the evaluative judgment.

Davidson (2005) distinguishes between two types of evaluation objectives: evaluative (those containing a value word) and non-evaluative (those containing descriptive or explanatory language). While it is appropriate for an evaluation to contain some non-evaluative objectives, a predominance of them means the inquiry is likely to be research rather than evaluation.

Findings: On examining the reports, it was found that some did not have evaluation objectives. In such cases the evaluation questions were examined instead. Four of the 28 reports were found to have no evaluation objectives or questions (despite having the word evaluation in their title). Of the 24 reports with evaluation objectives/questions, seven reports contain only evaluative objectives/questions, seven reports contain only non-evaluative objectives/questions, while 10 reports have a combination of evaluative and non-evaluative objectives/questions. Therefore 17 reports contain evaluation objectives/questions that provide a foundation for evaluative reasoning.

As might be expected in public sector evaluation, many of the values expressed in the evaluation objectives are associated with public values, such as effectiveness, efficiency, relevance, sustainability, value for money, and cost-effectiveness.

Report 7 is one of seven reports containing descriptive evaluation objectives/questions only, as expressed in the statements: “Did the project meet its delivery targets, and if not why not? Were the homes safety audited, and what did the safety audits find? What were the changes made to the houses, and what equipment was provided?” Despite its descriptive focus, the report ends with an evaluative conclusion: “The preliminary evidence presented in this report strongly indicated good progress”. (The inclusion of evaluative conclusions/judgments in reports without evaluative evaluation objectives/questions is discussed further in section 7.8.6).

Report 8 is one of the four reports without evaluation objectives/questions. The author states that it is a formative evaluation of a residential community-based therapeutic programme. The author of the report appears to be the same person who led the programme design. Other than the statement that the evaluation is a formative evaluation, there is no information about the aim or intent of the evaluation. The report also lacks a methodology section. The absence of an evaluation purpose, objectives and questions, together with any explanation of the methods raises questions about the evaluation capability of the agency and author.

7.8.3 Element two: Criteria or other comparator

To evaluate is to compare: "The fundamental idea of conceptualising quality is through comparison, direct or even vaporously indirect" (Stake & Schwandt, 2006, p.412). Criteria provide the most explicit approach for such comparison, defined as "the aspects, qualities or dimensions that distinguish a more meritorious or valuable evaluand from one that is less meritorious or valuable" (Davidson, 2005, p.91). Criteria provide the grounds on which the evaluator reasons towards an evaluative judgment (Fournier, 1995; Valovirta, 2002). Fournier (1995) explains the critical contribution of criteria to the reasoning process as follows:

> Criteria selection affects the validity of conclusions because it influences the reasoning used in establishing them. The reasoning is affected because the source of the criteria commits us to look for certain kinds of evidence and to appeal to certain kinds of warrants in order to justify resulting claims. In other words, how evaluators reason towards an evaluative judgment depends on how value is defined (p.22).

Stake and Schwandt (2006) exhort evaluators to be rigorous in their approach to selecting criteria, for example, by combining stakeholder perspectives, relevant literature and personal experience of the evaluand: "Majority opinion (of stakeholders) should not be considered sufficient . . . standards of quality generated by representative groups and quotations from learned papers are but starting points" (p.412). Other authors (Hurteau, Houle & Mongait, 2009) advise that the validity of criteria (and therefore of the evaluative judgment) is strengthened if the criteria are justified, for example, by reference to their source. Two standards for identifying and selecting criteria have been identified by G. T. Henry (2002), namely, transparency of the criteria selection process, and minimisation of bias. Such guidance about criteria and their selection is highly pertinent given Valovirta's (2002) observation that "the grounds (criteria) on which evaluative judgments have been made form the basis of one of the most common forms of debate about an evaluation report" (p.63).

As described in chapter 5, Stake and Schwandt (2006) identify two approaches whereby an evaluand is assessed against a comparison: a criterial-based approach (as described by Davidson, 2005), and a "quality-as-experienced" approach (p.408), for example, the assessment of wine by the expert viticulturist, or the valuation of fine art by the professional art appraiser. There were no examples of the quality-as-experienced approach in the study sample.

Findings: Values are defined in the form of criteria in 11 of the reports, in either a rubric (J. King, McKegg, Oakden & Wehipeihana, 2013) (six reports), descriptive textual definitions (three reports), or expressed as indicators (two reports). The reports were examined to identify whether (and how) criteria were justified. The authors of nine of the 11 reports with criteria provided one or more justifications for their criteria. The evaluation commissioner and/or stakeholders were involved in defining the criteria in five evaluations. Relevant legislation, literature, policy and programme documentation informed the criteria in three evaluations. This is illustrated by a desk-based evaluation of an environmental policy statement proposal (report 11). The criteria to be used in evaluations of environmental policy statement proposals are specified in the Resource Management Act 1991 (section 32 (1), (2)).

Two evaluations use existing criteria. Child maltreatment prevention criteria are used in an evaluation of an early childhood parenting intervention (report 6). An evaluation of a New Zealand regulatory policy framework (report 14) uses OECD Development Assistance Committee (DAC) criteria. The rationale for using development criteria in a non-development context is not explained.

The evaluators developed the criteria used in two reports. The authors of one report explained this was required because no existing benchmarks were found that were deemed relevant for the evaluand, a post-disaster temporary housing project (report 2). The other evaluation (report 30) was conducted by a group which included an expert in the economics of problem gambling interventions.

Scriven (1994a) and Davidson (2005) note that individual criterion can be weighted according to importance or some other aspect. All 11 evaluations which use criteria give equal weighting to the criteria.

As described above, evaluators may prefer to use a comparator other than criteria. Three reports (all of which are evaluations of capability building and therapeutic interventions aimed at individuals) include relevant academic literature (reports 6, 18, 38). In all three cases the literature is used to provide a broader context about the topic within which the findings about the evaluand are presented. Used in this way, the literature acts as an indirect comparator (rather than specific findings being compared directly to relevant topics in the literature).

Of the 11 reports with either no evaluation objectives/questions or only descriptive evaluation objectives/questions, nine reports refer to one or more values in the body of the report and/or in an evaluative conclusion/judgment (such values are not defined). Such inconsistencies are illustrated as follows (examples of such evaluative conclusions/judgments are provided in section 7.8.6). Report 15 (a programme evaluation) has an evaluative purpose but no evaluation objectives. The findings are discussed under headings that are evaluative (for example, "Appropriateness and effectiveness of engagement strategies"), yet these value terms are not defined anywhere in the report. Similarly, despite having descriptive evaluation objectives, a health impact assessment evaluation (report 44) uses headings in the discussion section such as "Was the health impact assessment process effective?"

7.8.4 Element three: Standards

While criteria define quality, they require accompanying standards to explicate how quality is discerned in relation to better quality and poorer quality. Standards act as benchmarks (Arens, 2005) against which the evaluand can be compared and ranked. Standards may be expressed quantitatively (for example, by a number grade or rank), or qualitatively (in a range such as from excellent to poor) (Davidson, 2005).

Findings: Of the eleven reports with criteria, six reports include standards of performance that are defined. In three reports, the definitions of standards are tailored to the evaluand, while three reports use standards based on generic definitions of performance (examples of tailored and generic standards are provided in Appendices H and I). The remaining five reports include references to standards of performance but the standards are not defined.

7.8.5 Element four: Warranted argument

As has been argued in Chapter 5, argument is an essential element of evaluative reasoning because it articulates the inference that links evidence to a normative claim (Fournier & Smith, 1993). According to these authors “building a justifiable argument is the crux of evaluation practice” (p.316). Argument assumes greater importance in evaluation than in other forms of systematic inquiry because of the type of inference used, namely, probative inference. As described in Chapter 5, this type of inference leads to a conclusion that is not certain (in a deductive sense), but rather is an “all things considered” judgment (House, 1995, p.40) or a “prima facie conclusion” (Scriven, 1991, p.277). A well-constructed and supported argument builds the plausibility of the claim (Booth et al., 2008, p.112). The role of argument in arriving at an evaluative judgment is explicated by Schwandt (2001b):

> . . . what evaluators should be doing in offering their professional service is not simply summing up empirical evidence and delivering a report of their ‘findings’ as it were. Rather, they should be engaging in a process of deliberation - using reasons, evidence and principles of valid argumentation to combine statements of fact and value to reach a reasoned judgment (p.266).

The strength of the evaluative claim (and argument) is strengthened by the inclusion of a warrant that is both reliable and relevant to the claim. The warrant establishes legitimacy through appealing to an authority or general principle (Booth et al., 2008). The warrant “legitimises the inference or reasoning step we make” (Fournier & Smith, 1993, p.318). As noted in Chapter 5, a warrant that is not established or conventional requires a backing to legitimate the warrant (Toulmin et al., 1979). The backing justifies why the warrant should be accepted.

Findings: Seventeen of the 28 reports contain an argument, that is, the author interprets the evidence to produce one or more evaluative claims (relevant to the evaluation objectives/questions) that are supported by reasons and evidence. In 10 of the 17 reports with an argument, the argument is located in a separate section to the findings, while in seven reports the presentation of the evidence is combined with the argument. At least half of these seven reports contain text where it is difficult to differentiate between evidence and the authors’ interpretation of the evidence. This has the effect of weakening the argument.

In contrast, 11 of the 28 reports either do not contain an argument (8 reports) or have text that is ambiguous, that is, it is not clear whether the text refers to evidence or argument (three reports). The authors of the eight reports without an argument summarise the evidence. This is followed by a short section usually headed conclusion which contains the authors' claims about the evidence. This section may finish with an evaluative conclusion/judgment.

Some of these reports give the impression of the author as the evaluation expert whose claims should be trusted despite a lack of transparency about how they were arrived at. Other reports give the impression of the author as a narrator, reporting the views of different stakeholders. The author then *changes hat* and becomes an evaluator, issuing an evaluative claim. In both cases, there is a lack of explicit interpretation of, and argument about the evidence. As a result, the inferential leap between evidence and claim is left to the reader to work out. This is illustrated in an evaluation of the effectiveness of a teacher development programme aimed at Māori teachers (report 35). The report has a combination of descriptive and evaluative objectives. None of the evaluative terms used (such as effectiveness, quality) are defined in criteria or by another comparator. A nine-page section titled *finding*s reports the results of an on-line survey of programme participants, together with key informant views about the programme collected in face to face interviews. Five short descriptive case studies complete the findings section. This is followed by a section headed conclusion (one and a quarter pages in length) containing the evaluator's claims. The absence of an argument means there is no interpretive bridge connecting the evidence presented in the findings section with the claims in the conclusion section.

Thirteen of the 17 reports that contain an argument use one or more warrants in either an explicit or implicit manner. Relevant academic literature and other documentation is the most frequently used warrant (six reports). For example, report 30 which is a value for money evaluation of a specific type of health-related intervention compares New Zealand's experience to that of four overseas jurisdictions. Cultural warrants are evident in four reports. Three evaluations of initiatives targeting Māori are authored by Māori evaluators who describe how the evaluation is based on Kaupapa Māori principles (reports 1, 17, 25), while an evaluation of an aid initiative in a Pacific state is co-authored by a New Zealand-based evaluator from the Pacific state (report 19).

A developmental evaluation of a sport and recreation programme targeted at Māori (report 1) provides extensive information about how Kaupapa Māori principles were applied in the design and conduct of the evaluation. The Māori author of an evaluation of a professional development programme for Māori teachers (report 25) provides a more succinct explanation of the Kaupapa Māori research methodology used:

> While the above western scientific research methods give the project a robust research process that permits the systematic collection of data, a range of Māori research principles and culturally safe practices were applied. The Māori principles and practices implemented in this research include, but were not limited to, Te Reo Māori (Māori language), manaakitanga (hospitality), aroha (love, regard), karakia (prayer), koha (gift), mātauranga (knowledge, comprehension, or understanding of everything visible and invisible in the universe) (report 25, p.9-10).

Six evaluations use expert warrants - either the evaluator is a subject expert or subject experts were involved in the interpretation of findings. For example, the authors of an evaluation of an early childhood parenting intervention (report 6) involved child health experts in the data analysis. The authors of an evaluation of a therapeutic residential intervention (report 18) are academics working in a university-based centre for addiction studies. In addition to these three evaluations, one could also argue that the Māori authors of the three evaluations based on Kaupapa Māori principles described above (reports 1, 17, 25) can be regarded as subject experts in Kaupapa Māori approaches, thereby providing an additional warrant for these evaluations.

An authority warrant is implicitly associated with two evaluations published by the Education Review Office (report 35, 37), given its statutory role as the agency responsible for evaluating the pre-compulsory and compulsory education sectors. Lastly, the authors of four reports included methodological warrants to strengthen the legitimacy of their claims. Such warrants include the use of participatory approaches in evaluation design and governance (report 19), rubric development (reports 14, 17), meetings to validate the analysis (report 30), and a diversity of stakeholders in the respondent sample (report 17). Such methodological warrants are illustrated in the following extract from report 19, a Pacific-based evaluation of an aid project:

> A participatory methodology was designed where stakeholders and participants had an increased level of participation and ownership

> of the evaluation as shown through the evaluation plan stage, the steering group mechanism and feedback workshops . . . a total of 111 stakeholders participated in the evaluation (report 19, p.14).

In four reports, the authors have directly or indirectly asserted the validity of their claims by referring to the quality and authenticity of their evidence. The authors of report 17 describe the wide range of stakeholder perspectives included in the evaluation which they claim increases the robustness of the conclusions:

> The stakeholder analysis carried out by the evaluation team in consultation with agency informants reflects the cross-section of stakeholders who participated in the evaluation. Overall, this broad coverage provided a good cross-section of feedback for reaching robust evaluative conclusions (report 17, p.21).

As noted above, a warrant that is not established or conventional requires a backing to legitimate the warrant (Toulmin et al., 1979). There was no evidence of backings for any of the types of warrants described above.

7.8.6 Element five: Evaluative conclusion/judgment

The evaluative conclusion/judgment is the intended *end-point* or *destination* of the reasoning process. Stake and Schwandt (2006) describe judgment-making as being fundamental to the evaluation profession: "Making judgments of quality constitutes a core professional responsibility of evaluators" (p.416). Despite this, House and Howe (1999) note that "there are no clear professional rules" (p.30) available to the profession about how to do so. According to these authors, judgment-making cannot be reduced to a set of standardised procedures. Rather, judgment-making requires the evaluator " . . . to take relevant multiple criteria and interests, and combine them into all-things-considered judgments in which everything is consolidated and related" (p.29). Writing some ten years after House, Alkin, Vo and Christie (2012) note that judgment-making remains an undeveloped area of evaluator practice: "Careful review of the program evaluation literature turns up only a few resources that describe value judgments and operationalise the ways in which they are reached" (p.29).

Despite this shortcoming, the literature identifies what makes an evaluative judgment robust. For an evaluative judgment to be legitimate and defensible, there must be coherent and transparent connections across the evaluation objectives, criteria (or other comparator) and standards, claims, argument, and judgment (Fournier, 1995; Fournier & Smith, 1993). Finally, the contingent nature of evaluative judgments must be stressed. Stake and Schwandt (2006) remind evaluators that all evaluative judgments regardless of their level of precision are “perspectival, temporal and conditional” (p.412).

Findings: Twenty-four reports contain one or more evaluative conclusions/judgments. Examination of them indicates three types: (i) comparator-based, that is, the conclusions/judgments are based on criteria or other comparator (13 reports), (ii) undefined values, that is, conclusions/judgments about value terms that are referred to in the evaluation objectives/questions and body of the report but are not defined (six reports), (iii) unreferenced values, that is, conclusions/judgments about value terms that are not referred to anywhere else in the evaluation report (five reports).

I first examine the group of reports that are comparator-based (type (i)). The presentation of the evaluative judgment in the 13 reports with criteria or other comparator is either by individual dimension (eight reports), or by individual dimensions which are also synthesised into an overall qualitative judgment (five reports). Report 41 (a report of an evaluation of a fund for primary industries initiatives) provides an example of a synthesised judgment which is located at the beginning of the executive summary: “The evaluation found that the (name of fund) is good value for money and makes a worthwhile and valuable contribution to primary industries and rural communities” (p.7). None of the reports with a synthesised judgment explain how the assessments of individual dimensions were aggregated.

I now examine the group of reports that have undefined values (type (ii)). Report 5, an evaluation of a training programme, is an example of this type. This report has as its evaluation objective: “This report examines the extent to which the intent of the (name of programme) was met and identifies what went well and what could be improved going forward” (There are no evaluation questions). There is no definition or description of what the programme *working well* would *look like*. The report consists of a findings section identifying four aspects of the programme that have been successful and three aspects requiring improvement.

This is followed by a two-page conclusion which refers to overseas literature about education to employment initiatives. There is an evaluative judgment in the introduction to the report identifying the successful elements of the programme: "The evaluation found the following elements of the initiative were successful: community leadership, scholarship model, academic support, community pastoral care" (report 5, p.2). The report makes statements and a judgment about success without defining this term nor building an argument to support the claim. Report 36 is another example of type (ii). This evaluation of governance arrangements of a government-funded educational institution contains only descriptive evaluation objectives (and no evaluation questions). The combined findings and discussion section includes headings such as "Improved quality of interactions between council and management" (p.24). The report has two evaluative judgments identifying improved quality of council/management interactions and enhanced council operations:

> The enforcement of the competency model has seen an improvement in the quality of interactions between council and management . . . Overall, there was agreement the change has enhanced council operations and brought greater consistency to the performance of the *(sic)* governance (report 36, p.28).

Despite the repeated use of the term quality and other value terms, they are not defined anywhere in the report. The final group of reports, unreferenced values (type (iii)) consist of five reports which contain evaluative judgments despite having descriptive evaluation objectives or no evaluation objectives (or questions).

7.8.7 Findings about the cost benefit reports

The sample includes five economic analyses - two value for money (VFM) analyses (reports 30, 41), two cost benefit analyses (CBAs) (reports 21, 42), and a cost effectiveness analysis (CEA) (report 14). The way the report authors have applied the VFM and CEA methods aligns with the five elements used in this meta-evaluation. As a result, these reports are reported with the non-economic evaluation reports above (in sections 7.8.1 to 7.8.6). For example, the cost effectiveness analysis (report 14) and one of the value for money analyses (report 41) use criteria and standards in rubric formats. The other value for money analysis (report 30) uses indicators which are each compared against (up to) four comparators. All three reports use the economic analysis results to argue a case, which in the case of reports 14 and 41 includes data sourced from other methods.

As the distinct features of the CBA method do not align with the five elements in a similar way, the findings from the two CBA studies are reported separately in this section. It is not the intention of this section to explicate the theoretical foundations of CBA or to describe CBA method in full, but rather to understand the extent to which the two CBA reports demonstrate evaluative reasoning as outlined in the elements used for this meta-evaluation.

Cost benefit analysis involves valuing, whether the valuing refers to the environment, water quality, time or a human life (Boardman, Greenberg, Vining & Weimer, 2006). These authors describe CBA as “a policy assessment method that quantifies in monetary terms the value of consequences of a policy to members of society” (p.2). Cost benefit analysis also involves another fundamental aspect of evaluation, that is, comparison - costs are compared to the value (expressed in financial terms) of the benefits arising from that expenditure.

The first step in a CBA involves identifying the costs and benefits associated with a particular intervention which are deemed to be relevant and material to the analysis. The analysis parameters of the CBA are also identified, for example, the type of adjustments to be applied and the time period to be used. Snell (2011) describes these decisions as involving value judgments. For this reason, Snell (2011) emphasises the need for the analyst to be explicit about whose point of view is being applied in the analysis and to explain the assumptions underpinning the study. He is critical of economists who ignore these explanatory requirements. The costs and benefits to be examined and the analysis parameters can be regarded as criteria. The standard applied in CBA studies is one (that is, the net benefit to cost ratio is above one). It is at this point that CBA appears to depart from the elements used in this study. It is assumed that a benefit to cost ratio of any figure higher than one is of value and is therefore acceptable. This score is left to speak for itself without further elaboration or justification. That is, it is assumed that the audience will understand the value of the benefit to cost ratio identified. In order to be explicitly evaluative, value for money evaluation must address the question of whether the outcomes achieved are worth the resources used (Davidson, 2005).

In the two CBA studies examined for this study, a positive benefit to cost ratio of greater than one was produced but no explanation of its value or acceptability was provided. For example, report 42 (a study of a health intervention) concludes with the authors' assessment that the benefit-cost ratios (ranging between 2.6 and 4.6 with a central - 4% discount rate – benefit-cost ratio of 3.9) are acceptable and therefore that the intervention "has been well justified in terms of positive net benefits" (p.26). The authors are relying solely on the cost-benefit result as the basis of their claim - a result which has not been examined in comparison with the CBA results of similar health interventions, or other relevant information.

In conclusion, cost benefit analysis is evaluative in two respects - it is a valuing exercise based on values identified as being relevant and important. Secondly, it is an exercise in comparison. The expressed ratio result of a CBA implies a comparison with the CBA standard of greater than one. However CBA does not articulate a level of justification of worth as described by Davidson (2005). Therefore in the context of this study, the two CBA reports fail to demonstrate all of the elements of evaluative reasoning.

7.9 Interpretation of the findings

This section begins with an overview of the evaluative reasoning practice evident in the 28 reports reviewed. (The two CBA studies are excluded from this discussion given the differences in approach described above). I then provide some possible explanations for the findings. Finally, some implications of the findings for evaluation practice in Aotearoa New Zealand are suggested.

7.9.1 Evaluative reasoning practice

As described above, the purpose of the meta-evaluation is to gather insights into evaluative reasoning practice, rather than to assess the quality of evaluative reasoning. For the purposes of this meta-evaluation, evaluative reasoning is deemed to be demonstrated by the presence of five key elements, namely, (i) evaluation objectives (or questions) that are evaluative, (ii) criteria (or some other comparator against which the evaluand is assessed) that are defined, (iii) standards that are defined, (iv) a warranted argument linking evidence and claims, and (v) a judgment that is evaluative. The meta-evaluation findings indicate that eight of the 28 reports have evidence of all five key elements (Table 7.2).[13] (Please note: Table 7.2 excludes the two CBAs). All but one of these evaluations was written by external authors. Eleven reports demonstrate three or four of the key elements. The most common omission is that values referred to in the report are not defined, for example, by criteria, in a descriptive textual definition or indicators (seven reports). There is no significant difference in authorship of these 11 reports - six were authored by external authors and five by internal authors.

The final group is made up of nine reports which lack three or more of the five key elements. Surprisingly, three of these reports end with a conclusion/judgment that uses evaluative language despite an absence of most or all of the preceding elements.

Table 7.2 Results by author type

	No. of reports with five elements of evaluative reasoning	**No. of reports with three-four elements of evaluative reasoning**	**No. of reports with two or fewer elements of evaluative reasoning**	**Total**
Internal authors (named & unnamed)	1	5	5	11
External authors (named & unnamed)	7	6	4	17
Total	8	11	9	28

The three groups of reports shown in Table 7.2 were analysed to ascertain whether any patterns are discernible, for example, by evaluand-type, evaluation orientation or approach. No patterns were apparent among these dimensions. However I noticed that the authors of the seven named reports with five elements of evaluative reasoning are active members of ANZEA and/or AES (these people are known to me). I then attempted to establish whether the authors of the remaining 16 named reports had individual membership of a professional evaluation association, such as ANZEA, AES, AEA[14] or are affiliated to WEG. Since some reports

[13] Although all the reports in the meta-evaluation are in the public domain and are therefore discoverable, for the purposes of protecting the identity and reputation of the authors as much as possible the reports are not reported individually according to the number of elements demonstrated in the report.

[14] Information about individual membership of AES and ANZEA is available. Information about members who are part of a corporate membership is not available.

had up to seven authors, I examined whether the principal author (the first name appearing in the list of authors) had an individual membership of, or was affiliated to, one (or more) of the above evaluation networks (Table 7.3). Fourteen reports appear to have principal authors who did not have individual membership of, or affiliation to, one of these professional evaluation networks. Even assuming that some of these authors were part of a corporate membership, there were more principal authors who appear to be working outside the umbrella of a professional evaluation network, than authors who are members of, or affiliated to one of them. Potential implications of this observation are examined in chapter 9.

Table 7.3 Evaluation type by author type

Evaluation type	**No. of named reports with principal authors who are part of a professional evaluation network**	**No. of named reports with principal authors who do not appear to be part of a professional evaluation network**
Programme/policy/strategy evaluation	9	10
Economic evaluation[15]	0	2
Health Impact Assessment	0	2
Total	9	14

[15] There are four economic analyses with named authors. In two reports, the economic analysis was part of a programme evaluation. These reports are therefore included in the programme evaluation row in the table.

7.9.2 Suggested explanations for the findings

The meta-evaluation findings do not claim to be representative of public sector evaluation practice in Aotearoa New Zealand. Rather, the findings can be regarded as providing a *snapshot* of evaluation practice, and as such offer insights for further consideration and investigation. The findings suggest there may be variable practice in evaluative reasoning among authors of public sector evaluations. This section offers a possible explanation for the observed variability, namely, evaluators' focus on practice, in particular on methods and evidence. I suggest this focus has distracted evaluators from becoming engaged with evaluation theory, and therefore with evaluative reasoning.

In his presidential address to the 1998 conference of the American Evaluation Association, William Shadish (p.1) stated "Perhaps because evaluation is primarily a practice-driven field, evaluation theory is not a topic that causes evaluators' hearts to flutter". Shadish's observation about the US evaluation community in the late 1990s is topical for the Aotearoa New Zealand situation. Evaluation practitioners have developed through doing evaluation rather than learning about the theory underpinning it. This emphasis on the practice of evaluation is demonstrated by two developments in Aotearoa New Zealand in the early 2000s. The first was the formation of the Social Policy Evaluation and Research Committee (SPEaR) in 2001. This was a government-funded body tasked with two objectives - building a knowledge base about government investment in social research and evaluation, and increasing the capacity and capability of the social sector to deliver evidence-informed advice (SPEaR, 2008). During its nine years of operation, SPEaR ran practice-based workshops and bi-annual conferences, and produced good practice guidelines on topics such as contracting research and evaluation, research and evaluation ethics, research and evaluation involving Māori and Pasifika. In examining SPEaR's activities, its response to improving "capacity and capability to deliver evidence-based advice" (n.p.) did not appear to include any requirements to focus on the theoretical underpinnings of evaluative reasoning.

The second development was the publication in 2003 of the only book that has been written about evaluation in Aotearoa New Zealand titled "Evaluating Policy and Practice: A New Zealand reader". In its introduction the editors state "Our goal was always to create a text which would become essential reading for those seeking to understand the theory and practice of evaluation in New Zealand" (Lunt, Davidson & McKegg, 2003, p.ix). The five chapters that are described as providing “an overview of theoretical and conceptual considerations raised by evaluation” (p.xii) examine topics such as evaluation approaches and methods, mixed methods evaluation, use of administrative data in evaluation, and evaluator competencies. Despite a foreword written by Professor Michael Scriven, the book lacks any substantive engagement with theoretical perspectives.

This focus on practice has contributed to the evaluation community being very receptive to new methods promoted by overseas evaluators visiting Aotearoa New Zealand described in chapter 2. As Shadish implied in his AEA address, new methods appeal to evaluation practitioners in ways that theory does not. Schwandt (2008a) makes a similar observation about the attractiveness of methods to evaluators and the impact on evaluative reasoning: “My concern is that in the press to master methods of generating data, we ignore the idea of developing a warranted argument - a clear chain of reasoning that connects the grounds, reasons or evidence to an evaluative conclusion” (p.146).

The emphasis given by successive governments to the role of evidence in public sector decision-making may have also contributed to the evaluation community’s pre-occupation with subjects other than evaluation theory. The terms evidence-based practice (Nutley, Davies & Walter, 2003) and evidence-informed policy making (Gluckman, 2013) describe this discourse. New Zealand’s focus on evidence has followed that of other countries such as the United Kingdom where the new Labour Government released a white paper titled *Modernising Government* (United Kingdom Government, 1999) endorsing evidence-based policy making, and Australia where the Australian Productivity Commission published a book (2009) of proceedings from its conference on evidence-based policy making, including a chapter about the New Zealand experience (Scobie, 2009). The focus on evidence is evident in other conferences held during this time. The theme of the first SPEaR Conference was *Evidence-based policy and practice in the social sector*. This was followed in 2005 by a conference with another evidence-related theme, namely, *What Works?* The theme of the 2009 conference of the Australasian Evaluation Society (which I and many other New Zealand evaluators attended) was *Gathering and Using Evidence.* The appointment of Professor Sir Peter Gluckman as the Chief Science Advisor to the Prime Minister in 2009 has provided new impetus for the evidence discourse in New Zealand. Professor Gluckman (2013) has reinforced the role of evidence in public policy making, stating that evidence requires research quality and integrity: “The fundamental challenge is to ensure that the scientific evidence is placed appropriately within the decision-making chain. Without proper initial assessment of the evidence, there are risks of inappropriate policy development at senior political level” (p.11).

In conclusion, I suggest that this focus on practice, in particular on methods and evidence as described above, has preoccupied Aotearoa New Zealand evaluators for some ten years, without the benefit of any critical perspective from a university sector focused on developing theoretical underpinnings for practice. This observation does not undervalue the importance of these topics in evaluative enquiry. However they have had the effect of distancing evaluators from theoretical engagement and reducing demand for universities to provide theoretical capacity to the field. Arguments supporting the value of evaluative reasoning as one of the key theoretical underpinnings of the field have not been developed.

7.9.3 Implications of findings for practice

This section identifies three implications for evaluative reasoning practice arising from the findings.

Implication no.1: Warranted argument requires greater attention

Of the five elements of evaluative reasoning examined in this meta-evaluation, warranted argument is the element where there appears to be the least attention. This is despite Greene’s (2011) advice that “warranted argument is a professional responsibility of evaluators” (p.90). Eleven of the 28 reports either do not contain an argument or contain text that is ambiguous (that is, it is not clear whether the text refers to evidence or is the author’s argument). A further seven reports combine evidence and the author’s interpretation of the evidence (argument), of which three reports contain text where it is difficult to differentiate evidence from argument. Consequently, half of the 28 reports lack an argument or have text that is ambiguous. This is a significant shortcoming that undermines the defensibility of the evaluative conclusion/judgment, exposing the evaluation to criticism about its quality. This is discussed further in chapter 9.

Implication no. 2: While endorsing the evaluator's role to provide evaluative conclusions/judgments, some authors do not demonstrate capability to do so

Twenty-four of the 28 reports contain one or more evaluative conclusions/ judgments, despite 11 of these reports lacking either an evaluative evaluation objective/question that links to the conclusion/judgment, or a definition for the value word used in the conclusion/judgment. This suggests that while the authors of these 11 reports endorse the evaluator's obligation to provide an evaluative conclusion/judgment, they do not demonstrate capability to do so.

Implication no.3: Articulating the limitations associated with an evaluation supports an evaluative conclusion/judgment by being explicit about its shortcomings

As a result of undertaking this study I have come to view any text in an evaluation report describing its limitations with new respect. Such text enables the evaluator to articulate the contextual and other factors that have had an impact on their evaluative reasoning. The limitations information can be regarded as acting in a similar way to the warrant - just as the warrant adds strength to the argument and therefore to the evaluative judgment, articulating the limitations of the evaluation supports the evaluative conclusion/judgment by being explicit about its shortcomings. It is therefore of concern that half of the authors of the 28 reports do not specify any limitations associated with their evaluation.

7.9.4 Does it matter that we do better?

Given the issues associated with public sector evaluative reasoning practice evident in this study, this section considers the question: Does it matter that we do better?

As described above, the five elements are interconnected and together build a coherent case from which an evaluative conclusion can be drawn that is defensible (Fournier, 1995; Fournier & Smith, 1993). Therefore the absence of, or weakness in one or more elements undermines the evaluative claim and therefore weakens the evaluative conclusion/judgment. More fundamentally, it compromises the logic underpinning the evaluation and therefore the evaluation report. Such flaws in logic are illustrated in some of the reports that contain either no evaluation objectives or questions (three reports) or only non-evaluative evaluation objectives (seven reports). Of these 10 reports, seven end with one or more conclusions/judgments that use evaluative language. Put simply, such reasoning is logically impossible and therefore seriously flawed.

Further as described in chapters 4 and 5, I assert that evaluative reasoning is evaluation Theory (the emphasis being on the capital T). It is what makes reasoning from values to an evaluative judgment valid according to informal logic. It is what distinguishes evaluation from other forms of systematic enquiry. If evaluation Theory (aka evaluative reasoning) is at the core of evaluation, it is critically important that those of us who conduct evaluations practice rigorous evaluative reasoning. Reports based on poor evaluative reasoning have the potential to discredit evaluation as a legitimate and useful tool for government agencies.

Referring to the question in the title of this sub-section, based on the snapshot offered by the meta-evaluation findings it appears that some evaluators need to do better, and it does matter that they do. Given my premise that evaluative reasoning is a critical aspect of evaluation quality, there appears to be need for greater diligence about reasoning to an evaluative conclusion.

7.10 Limitations

There are a number of limitations associated with the meta-evaluation. Firstly, the sample is not, nor does it claim to be, representative of public sector evaluation reports. Only evaluation reports in the public domain were able to be considered for inclusion in the sample. There is no requirement in New Zealand for public sector agencies to make evaluation reports and other official documents publicly available, other than via a formal request made under the Official Information Act 1982. As noted above, of the 63 agency websites examined, only 22 websites contained evaluation reports (not summaries) dated 2010-2013. No reports were available on the websites of some large agencies. Secondly, while there are many programme evaluation reports posted on websites, there are noticeably fewer policy evaluation reports. The reports of some large-scale, national policy initiatives posted on agency websites (such as *Working for Families* and *KiwiSaver*) are summary reports of high-level findings written for a general audience. Such reports were excluded from the sample because they lack sufficient detail about the evaluation approach required for the meta-evaluation. Lastly, the review focuses only on central government agencies as there were an insufficient number of reports available on local government agency websites.

A further limitation is associated with desk-based examination. There are a range of influences that determine the design and conduct of an evaluation and a desk-based examination is unable to collect such information. As noted above, political influences are both inherent and significant in the public sector context. What is presented in an evaluation report represents the requirements of the commissioner irrespective of whether the report has been produced internally or externally. An evaluation may have been poorly scoped by the commissioning agency, leaving the external evaluator to do the best they can with a poorly considered brief. An evaluation may have had a restricted budget, resulting in an argument that is unable to be supported by expert or literature-based warrants. These and other important contextual factors are not visible in an evaluation report, unless they are included as information about limitations. Lastly, the contents and format of an evaluation report may reflect the requirements of the commissioner. A commissioner may request a report presenting high level findings only, and may place lesser value on a robust argument supporting such findings.

A final limitation is that the examination of the reports was undertaken by one researcher who knew some of the report authors, creating the possibility of researcher bias. A larger meta-evaluation could be undertaken involving a small team of evaluators thereby allowing for inter-rater reliability to be established.

7.11 Conclusion

This study seeks to understand how evaluative reasoning is understood and practised in a public sector context in Aotearoa New Zealand. While the Q study reported in chapter 6 provided insights into how evaluative reasoning is understood in this context, this chapter has presented a different perspective by examining (using meta-evaluation) evaluative reasoning practice as demonstrated in evaluation reports written or commissioned by public sector agencies. The meta-evaluation indicated significant variability of practice among a non-representative sample of 30 reports written by New Zealand-based authors in the period 2010-2013 for 20 public agencies. The snapshot of current practice provided by the meta-evaluation findings suggest there is scope for evaluators working in or commissioned by public sector agencies to improve their evaluative reasoning practice, particularly in respect to building a strong evaluative claim through warranted argument. This chapter has argued that improving the practice of evaluative reasoning is not an option for evaluators. Rather, robust evaluative reasoning is at the core of what it means to undertake evaluation, particularly in the public sector context. This argument is discussed further in chapter 9.

CHAPTER 8

PERSPECTIVE THREE: EVALUATION EXPERTS

8.1 *Rationale*

This chapter provides a third perspective on evaluative reasoning in the Aotearoa New Zealand public sector through engaging international and New Zealand-based evaluation experts with the findings from the Q methodology and the meta-evaluation. As I am embedded in the professional evaluation community which is the focus of this study, I was concerned that I may be blind to assumptions associated with local evaluation practice or immune to nuances in the findings. In qualitative research this is referred to as an insider (emic) perspective, as opposed to an outsider (etic) perspective (Patton, 2002a). Patton (2002a) emphasises the importance of including both insider and outsider perspectives in qualitative studies: "Understanding different perspectives from inside and outside a phenomenon goes to the core of qualitative inquiry" (p.335). I therefore decided to seek input from overseas evaluation experts with knowledge about Aotearoa New Zealand to provide an outsider perspective, as well as local evaluation experts with overseas experience to provide an insider/outsider perspective.

8.2 *Qualitative interviews*

Interviews are an important tool for social enquiry, that is, inquiry that "aims at understanding the meaning of human action" (Schwandt, 2007a, p.248). Qualitative inquiry is based on the epistemological position that knowledge claims about a topic of interest must be based on people's experience, perceptions and meanings: " . . . what is in and on people's minds . . . their stories" (Patton, 2002a, p. 341). Qualitative interviews provide descriptive and explanatory data variously described by Miles and Huberman (1994) as "well grounded" (p.1), "holistic" (p.6), and capturing “raw experience” (p.9). Qualitative interviews allow for the probing and clarification of responses, as well as the capturing of rich contextual information to assist sense making (Patton, 2002a). Schwandt (2007a) identifies three types of qualitative interview based on the research task, type of knowledge sought, and epistemological influences. The first interview type is referred to as “fact-finding” (p.164) where the research task is to produce factual data. This type has its roots in logical empiricism and logical positivism which require knowledge claims to be verifiable through observation or logic (Schwandt, 2007a). The second type is referred to as “stories of experience” (2007a, p.164) which aims to elicit accounts of lived experience through in-depth data. The influences of this interview type include naturalistic inquiry, referred to as “aexperimentalism” (Lincoln & Guba, 2013, p.20) to indicate its juxtaposition to experimentalism. Whereas experimentalism aims to eliminate context through the controlling of variables, naturalistic enquiry focuses on understanding human action through “the meaning, character, and nature of social life . . . in a particular setting . . . . from the point of social actors” (Schwandt, 2007a, p.206, 207). The third type is referred to as “interactional encounter” (p.164) derived from ethnomethodology and social constructionism which are grounded in an emic or insider perspective of a phenomenon (Schwandt, 2007a). This interview type aims to understand not only the substance or content of the interview, but also how meaning making occurs.

Patton (2002a) illustrates how the interviewer’s role and the interviewer-interviewee relationship may differ in qualitative interviews. In a structured interview the interviewer acts as a collector of information according to a standardised interview schedule. There is minimal interaction between the interviewer and interviewee, other than that relating to the interview schedule. A second role is where the interviewer conducts an interview according to a guide identifying questions or issues to be explored. There is greater interaction between the interviewer and interviewee as the interviewer has flexibility to explore emergent areas of interest or topics of importance to the interviewee. Another role, sometimes referred to as ethnographic interviewing, is based around unstructured interviewing where the interviewer engages the interviewee in a conversation on the topic of interest with the interviewer being “ . . . free to go where the data and respondents lead” (p.343). The choice of interviewer role and interview instrument will be determined by the purpose of the inquiry, the nature of the information sought (for example, data that is comprehensive and comparable, or data that is personalised and sensitive to context), the number of interviewees, and the time available (Patton, 2002a).

Research ethics assume heightened importance given the highly personalised and potentially identifiable information collected from qualitative interviews. Sound ethical practices include informed consent, research participants having the option to withdraw from the study at any time, interviewers' duty of care responsibilities (for example, ending an interview prematurely if it is causing stress to the interviewee), interviewee anonymity, data confidentiality and security (Australasian Evaluation Society, 2013).

8.3 Approach

Six local evaluation experts with international experience and seven international evaluation experts (from Australia, UK and USA) who have either worked in, or recently visited Aotearoa New Zealand were identified. From this list, I approached (via email) three local experts and three international experts to request an interview. All agreed to be interviewed.

The aim of the interview was to test the findings for their trustworthiness in portraying local evaluation practice, and to identify contextual factors to explain the findings. I produced a two-page summary of the Q and meta-evaluation findings (Appendix J), which included two open-ended questions for the experts to consider about the findings, as follows. (I also emailed a copy of a published paper about the meta-evaluation findings as further information, if the experts chose to read it).

> (i) Are the findings surprising? If yes, why? If no, why not?
>
> (ii) Is there anything about the findings that is unique to New Zealand? If yes, what? why? If not, why?

There are a number of reasons for using only two open-ended questions. Firstly, the expertise of the interviewees means that they do not need direction or prompting as is provided by structured questions. Secondly, I wanted the experts to interpret the summary of findings in their own way, rather than prescribing their interpretation through focused questions. Lastly, as busy professionals I thought the experts might be more responsive to my request if the interview did not involve too much time and preparation.

8.4 The evaluation experts

The six evaluation experts are currently working as a private consultant, academic or employee. All have published on evaluation-related topics. In the discussion of the findings that follows, the New Zealand experts are identified as NZ1, NZ2, NZ3 and the overseas experts as OS1, OS2, OS3.

8.5 Interviewing, coding and analysis

Three experts were interviewed in person and three via Skype. All gave verbal consent for the interview to be recorded. The interviews lasted between 15 and 45 minutes (the 15 minute interview was 40 minutes in duration but had two significant interruptions which I did not record. Unfortunately the disruptions affected the flow of the interview). I took notes during the interviews and the interviews were also transcribed. I did not engage in any in-depth discussions with experts about their responses (other than probing for clarification purposes) due to the short length of time available for the interview.

This section describes the coding and analysis process. As the six interviews were conducted over a period of five months, I started by reading the transcripts and interview notes a number of times to re-familiarise myself with each interview. The interviews were coded using NVivo10 based on a coding frame consisting of themes of interest. New codes were added as the interviews were being coded. Questions and comments about the data were written up in note form during coding. The analysis of the data was an iterative process - the data were examined to identify similarities (themes) and differences (questions to explore further), and re-examined some days later to ensure that the full meaning of the interviews had been captured. Notes were written on both occasions recording thoughts and questions about the data. This process of examining and re-examining the data over a number of days helped the reflective process.

Despite the open ended nature of the interview questions, distinct themes emerged across five of the interviews particularly about contextual influences that have shaped local evaluation practice as discussed below. The data were also examined to identify any differences in responses from New Zealand-based experts compared with those from the overseas-based experts, and the three experts who are consultants compared to the three evaluation academics and employees. No differences were found except that the overseas-based experts were generally more positive about evaluation practice in Aotearoa New Zealand than the local experts. The small number of interviewees and the fact that the experts are either personally known to me or I was familiar with their work made the coding and analysis of the data more straightforward. I was able to become immersed in and remember the content of each interview in a way that may not occur when interviewing a larger number of unfamiliar interviewees.

8.6 Limitations of expert informant interviews

The findings are based on interviews with six experts who were asked to respond to a summary of the Q study and meta-evaluation in a short, unstructured interview. While the experts gave insightful comments and observations about evaluation practice in Aotearoa New Zealand, the interviews were not designed to provide an in-depth examination of the way in which evaluation is conceptualised and practised. The interviews sought corroboration of, or challenge to, the Q and meta-evaluation findings, rather than to gather primary data. Further, there is a level of abstraction and generalisation in experts' comments and observations. Another potential limitation is that the New Zealand-based and overseas experts are members of one or more of ANZEA, AES, AEA or EES. These experts are therefore not representative of the meta-evaluation authors who are not affiliated with a professional evaluation organisation as reported in chapter 7.

8.7 Findings

8.7.1 Overall response from five experts

The experts who work or have worked in New Zealand were generally not surprised by the summary of the Q study and meta-evaluation findings, describing them as being congruent with their experiences. As noted above, the experts with work experience in New Zealand were generally more critical of local evaluation practice, while the overseas-based experts provided more positive comments. One of these experts observed that the weaknesses in evaluative reasoning identified in the meta-evaluation are no different from reports authored by some evaluators in their country. The sixth expert did not respond to the findings, but rather critiqued the approach and methods used to address the research question. This critique is reported in section 8.7.3.

8.7.2 Influences on evaluation practice

Five experts provided insights about and explanations for the summary findings. The following influences were identified as contributing to the shaping of evaluation practice in Aotearoa New Zealand.

Heritage-influenced evaluation practices

Four experts who are working or have worked in Aotearoa New Zealand identified aspects of evaluation practice that can be regarded as reflecting its heritage. At this point it is necessary to take a brief diversion to explain this heritage. New Zealand's early stories are about Polynesian peoples arriving in the thirteenth century and thriving in their new home with its rich soil, bird life and fisheries (M. King, 2003). Europeans arrived from the 1830s, initially ex-convicts from Australia and traders, followed by English and French missionaries and their families, and migrants looking to escape nineteenth century Europe (M. King, 2003). Māori were presumably important in imparting practical survival skills to the Pākehā newcomers to help them become accustomed to their new life in a geographically isolated country. While its development has not been formally documented, a notion emerged over the next century of Aotearoa New Zealand as a country of people " . . . whatever their cultural backgrounds (who are) . . . practical and commonsensical" (p.520) and as having "the versatility of practical men" (Mulgan, 1993 as cited in M. King, 2003, p.514). This notion is expressed in a colloquialism which is still in use: “No. 8 fencing wire ability” (p.513).[16] The meaning of this colloquialism is that New Zealanders can turn their hands to anything, requiring only minimal resources to do so. This has led to the present-day notion of New Zealanders as a DIY-ers (DIY meaning ‘do it yourself’) (M. King, 2003).

According to one expert, this pragmatic heritage is reflected in the way some people start doing evaluation with the attitude “this evaluation stuff can’t be too hard”, and without any training or reading. This expert wryly observed: “Sometimes research isn’t the worst they can do. They can do evaluation” (NZ1). A second expert observed that our pragmatic heritage is evident in the way we value practical knowledge more than theoretical knowledge, and evaluation practitioners more than evaluation academics. Comparing evaluation practice in Aotearoa New Zealand with other countries that place greater emphasis on evaluation theory, a local expert described our practice as being “informal . . . less systematic and lacking rigour . . . We have a much less rigid notion of what is involved in an evaluation” (NZ2). Given these observations, the shortcomings identified in the meta-evaluation findings were unsurprising to these experts.

While the aspects of evaluation practice identified in the preceding paragraph can be regarded as being less favourable consequences of New Zealand’s pragmatic heritage, there are also positive consequences. The orientation one evaluator described in the Q study - *the idealist, pragmatic evaluator with an eclectic approach* - resonated with two experts. One of the experts described local evaluators as not being committed to particular methods or approaches, but seeking out the best method for a specific evaluation “We like to mix it up a bit, use our toolkits . . . we’re pragmatic” (NZ3). A second expert observed: “It’s not just a question of pulling a methodology off the shelf . . . there is much more thought put into it” (OS1). Experts contrasted this approach to that in the UK and US. An expert described public sector evaluation in the UK as being based on a technocratic approach, while another experts said that evaluation practice in the US “ . . . revolves around the theorists” (NZ3), with evaluators tending to specialise in specific approaches and methods.

[16] Number eight wire is used in the building of farm fences (Te Ara, n.d.).

Māori and Pasifika influences

Just as New Zealand's European heritage has influenced evaluation practice, so too have Māori and Pasifika heritages. According to an overseas expert who has worked in New Zealand, the influence of Māori and Pasifika on New Zealand society is profound, resulting in a New Zealand way of being:

> What I think is underestimated is the influence that Māori and Pacific cultures have on mainstream New Zealand. It is profound, absolutely profound . . . but it's not fully recognised . . . leaving Pacific to one side for the moment, (people) talk like it's a dichotomy - you have got Māori and you have got Pākehā but it isn't that simple. There's a massive overlap. There is a New Zealand way of being which is hugely influenced by a Māori way of being (OS1).

Two experts who are based overseas described how they perceive local evaluation practice has been shaped by Māori ways of being. One expert described values as being prominent in evaluation practice due to the Treaty of Waitangi:

> . . . values are really high up on the list in New Zealand. There is a lot of talk about values because of the cultural dimensions. It happens more broadly in the evaluation community internationally but it is particularly emphasised in New Zealand . . . There is the Treaty. You have got a basis for doing it. Other countries don't have that. New Zealand is unique (OS1).

The same expert described New Zealand as being a leader in culturally responsive evaluation approaches with indigenous peoples: "New Zealand has led on this . . . it's way ahead (of other countries)" (OS1). A second expert noted that Māori evaluators' contribution goes beyond culturally responsive approaches through their articulation of the importance of culture for evaluation validity from an Aotearoa New Zealand perspective:

> The idea of weaving in cultural values . . . embedded deep into the evaluation questions . . . not because it is the nice thing to do but because it is about validity. And being hard-nosed about doing it, not just because somebody said so . . . it's about getting the damn thing right. This is a huge strength of what we've got in New Zealand (NZ3).

Size-related influences on evaluation practice

As noted in chapter 2, Aotearoa is a small country demographically with a population of 4.5 million. Five of the experts identified influences on evaluation practice that are related to this small population size. Three experts described evaluators as needing to be generalists because specialist opportunities are infrequent: “We have to be kind of Jack or Jill of all trades. I have seen people go into content areas where they have not worked before, yet content knowledge would be a fabulous thing” (NZ1). Another expert reflected on how being a generalist evaluator is a skill and “New Zealanders have got it” (NZ2). A consequence of the small public sector is that evaluators are likely to have relationships with policy-makers and evaluation commissioners:

> We are closer to the policy-makers and clients in general. Being a small country, things happen on a smaller scale which means we are more likely to have ongoing relationships. In the US there are huge government contracts and everything is more impersonal (NZ3).

Salmond’s (2008) description of Aotearoa New Zealand as an intimate society is reflected in the comments of one expert who referred to the “connectiveness” of New Zealand society. Describing social and professional networks, another expert observed: “We tend to know one another or we know people who know each other” (NZ1). This expert surmised whether such connectedness increases the quality of relationships: “I wonder if that makes us have a little bit more value or integrity (in relationships)” (NZ1). At a professional level, an expert commented that evaluators have to take care in their practice because of their visibility. (The Māori word *mana* is used in this context to refer to an individual’s good reputation).[17]

> You can’t easily be anonymous in New Zealand so you have to take care. You can’t just blunder around because everybody knows you. Your mana has to be preserved whereas in larger countries you could blunder around as much as you like. You won’t see the same person twice (OS1).

A further reflection was whether such intimate networks make New Zealanders reluctant to criticise others and do not like being criticised, which may impact on the way evaluators give evaluative feedback:

[17] Mana refers to authority, control, influence, prestige and power (Moorfield, 2005, p.76).

> I wonder if . . . it is linked to an idealist position which says we should be encouraging of one another . . . We are not very good at giving or receiving bad news . . . so what we've done as evaluators is we've tried to move around that so we don't have to give people bad news (NZ1).

There is also a cultural dimension about not being critical of others, namely, mana. An expert talked about the importance of evaluation maintaining or enhancing an individual's or organisation's mana. Therefore negative evaluation findings may be regarded as criticism which undermines an individual or organisation.

An overseas expert commented on the ease with which "things can get done in New Zealand" (OS2) because of its small size and nationally-based public policy-making (unlike larger countries with federal and state government structures). She illustrated this by describing a meeting with evaluation officials from a number of government departments during one of her visits to New Zealand: "You can literally get the key players in a room . . . some things are easier to do when you have a manageable group" (OS2). She contrasted this to her country with its two levels of government which adds complication and leads to lengthy delays.

Importance placed on relationships

The importance placed on interpersonal relationships, stakeholder input and dialogic processes which were shared across all three orientations in the Q study were of no surprise to some experts, with one expert describing these features of evaluation practice as reflecting cultural norms: "No one would dream of marching into a community and saying 'we're going to do this evaluation and this is how we are going to do it' . . . yet that's how it's done in some other places (countries)" (OS1). Responding to the Q findings about stakeholder input and dialogic processes, a second expert noted their importance for producing authentic evaluative findings: "We involve people, not just to say, 'Let's get this perspective heard', but it's because we won't get the evaluation right unless we do that" (NZ3).

A third expert described the importance of the relationship-building aspects of government-funded programmes, particularly those involving Māori, which are often overlooked by officials. According to this expert, some agencies want to measure programme outcomes prematurely and do not understanding the need for relationship-building to occur before progress can be made towards outcomes:

> We would have government agencies saying ‘Yes, but what are the outcomes?’ We would always be trying to say there is some foundational relationship stuff here that sits underneath those outcomes and has to be done before the outcomes you are looking for, which are actually down the line (NZ1).

The importance placed on relationships in evaluation also has a cultural dimension, with Māori and Pasifika evaluators having ongoing relationships and accountabilities to the Māori and Pasifika communities involved in the initiatives being evaluated (Moewaka Barnes, 2003). This responsibility is reflected in an expert’s comments about evaluations involving not-for-profit providers of government-funded services:

> (You) want the voices of the people to shine through the (evaluation) report and to be privileged because you are entering their world . . . they haven’t got time to be empowered by any sort of great participatory method because they are too busy delivering programmes that are under-funded by government agencies who expect outcomes that are too far down the track (NZ1).

The relational aspect of New Zealand life is also evident in two experts’ descriptions of the local evaluation community. An overseas-based expert described New Zealand evaluators as being collaborative and supportive of each other. Another expert referred to the local evaluation community as being “cohesive” (NZ2) noting the high attendance numbers at the ANZEA annual conference (around 180 people) relative to our population size. (The number of attendees at the AES annual conference held in Australia ranges from 255 to 530, depending on the location and health of the government sector) (AES, 12 April 2016, personal communication).

Scrutiny and challenge

According to two experts (a New Zealander and a non-New Zealander), there is a lack of scrutiny and challenge to the way evaluation is practised in Aotearoa New Zealand. As noted above, one expert described the local evaluation community as cohesive. This expert noted a potential disadvantage of such cohesiveness: “I think as with any group there is always a risk that you end up with group-think . . . we have that really informal style of doing evaluation, (and) there’s no one really coming along and saying ‘hey guys, you have to sort yourselves out’” (NZ2). An overseas expert reflected that New Zealand evaluators “sometimes think and talk in shorthand . . . Sometimes I feel that because they’re not being challenged, they’re not actually having to make it explicit and maybe if they were, it might sharpen up a bit” (OS2).

Professor Michael Scriven’s time in New Zealand in the early 2000s as Professor of Evaluation at the University of Auckland, and the return of Dr Jane Davidson were described by two New Zealand-based experts as being very beneficial because they challenged evaluators’ practice. Another expert observed that challenge from stakeholders and evaluation participants is as important as external scrutiny: “If I think back to some of my experiences in New Zealand . . . I remember working on the (name of policy). The challenge came from kaumātua, from the Māori elders because the evaluation was being done in their community” (OS2). (Kaumātua refers to an elder, man or woman, who is held in high esteem).

Other reflections about the findings

Two New Zealand experts expressed surprise about some Q study findings. One expert wondered whether there may be a disjuncture between people’s perspectives as expressed in a Q sort, and what they actually do in practice: “I wonder whether there is some social desirability stuff going on in the responses (in the Q study). They may say this when they do Q but you might see different things going on when you look at their evaluation reports” (NZ3). The five shared themes across the three orientations was a surprise to another expert as she anticipated there would be greater diversity of evaluation practice expressed.

Both of these comments reflect the recruitment strategy for the Q study discussed in chapter 6, namely, the Q participants were recruited from the approximately 300 subscribers to the Wellington Evaluation Group’s email list who can be deemed to constitute the local evaluation community. It was not until I conducted the meta-evaluation that I realised that a wider group of professionals, including those who may not self-identify as evaluators, are being commissioned to undertake public sector evaluation. This is discussed further in chapter 9 (section 9.3.3).

8.7.3 An expert’s critique of research approach and methods

Rather than examining the findings of the Q study and meta-evaluation, an international expert critiqued the approach used to examine the research topic.

The expert's first comment concerns the fact that the study is based in western epistemology, thereby excluding indigenous epistemologies. The expert expressed surprise at this approach given the scholarship about indigenous evaluation coming out of Aotearoa New Zealand. The expert noted that there may be other equally coherent logics of evaluative reasoning in indigenous science and other epistemologies. Secondly, the expert observed that evaluative reasoning is mostly inferred from other things, rather than being observable in an evaluation report. The expert noted that the content and format of an evaluation report may be dictated by the commissioner and/or needs of the evaluation audience, making the criteria used in the meta-evaluation redundant. The expert also noted that if the elements of evaluative reasoning examined in the meta-evaluation are not visible in a report, they may still have been present in the evaluation. The expert suggested that ethnographic interviews with evaluation practitioners about their practice would have been more appropriate for answering the research questions.

The expert's critique served as a useful reminder of the need to clearly articulate the bounded nature of this study, its limitations, and the justification for choosing Q methodology and meta-evaluation to examine the topic (as discussed in section 3.2). Further, the expert's comment about evaluative reasoning being mostly inferred from other things (rather than being observable in an evaluation report) challenged my thinking. My response to this is provided in chapter 9 (section 9.4.7).

8.8 Macro-level influences on local evaluation

The responses of five experts to the Q and meta-evaluation summary findings identified influences associated with Aotearoa New Zealand's historical, social, cultural and geographical context that may shape how evaluation is conceptualised and practised. This section summarises these influences and identifies their potential impacts on evaluative reasoning.

In summary, New Zealand's pragmatic heritage may contribute to evaluation being conceptualised as a practical, action-orientated undertaking. The pragmatic tradition may also contribute to greater value and importance being placed on the practice of evaluation rather than the theory on which such practice is based. This leads to evaluation practice that is described as being informal, less systematic and lacking rigour compared with countries where evaluation theory is given greater emphasis. Such deficiencies may be compounded by a lack of external scrutiny of practice described by two experts.

New Zealand's intimate society and relational way of being contribute to evaluation being conceptualised as a relational practice. This is reflected in the importance placed on stakeholder knowledge and input, and dialogic processes evident in the Q study. In a relational-based society the giving of negative feedback may not be socially acceptable, creating potential challenges for the way evaluators present evaluative conclusions/judgments. The Treaty of Waitangi, together with the influence of Māori and Pasifika cultures on the New Zealand way of being are described as creating greater awareness about the role of values in evaluation than may be the case in other countries in which the experts are based. Lastly, Māori evaluators are described as leading the international evaluation community in culturally responsive evaluation, and through their contribution to the cultural validity discourse as discussed in chapter two.

These conceptualisations of local evaluation practice raise questions about their potential impact on evaluative reasoning practice in the Aotearoa New Zealand public sector. While the following questions focus only on evaluative reasoning, many are equally relevant for other aspects of local evaluation practice.

i. What is required to engage the pragmatic evaluator with evaluation reasoning theory, and to understand the implications of such theory for practice?

ii. What is required to reconcile the pragmatic evaluator's focus on the mechanics of doing evaluation with the cognitive deliberation and argumentation involved in evaluative reasoning?

iii. To what extent, and in what ways, does being an intimate, relational-based society impact on the way evaluators arrive at and present evaluative conclusions/judgments?

iv. To what extent, and in what ways, does the New Zealand propensity to avoid giving criticism and being unreceptive to criticism impact on the way evaluators arrive at and present evaluative conclusions/judgments?

v. To what extent, and in what ways, do the relational and dialogic features of local evaluation practice strengthen, or weaken evaluative reasoning?

vi. How do evaluators manage the tension between producing evaluative conclusions/judgments for accountability purposes and maintaining the mana of those being evaluated, particularly when evaluating Māori and Pasifika providers of government services?

vii. If external challenge and scrutiny of local evaluation practice are lacking, how does the evaluation community ensure the quality of public sector evaluative reasoning?

These questions suggest a range of potential impacts on evaluative reasoning practice. The purpose of identifying such questions is not to answer them, but rather to demonstrate how the application of evaluative reasoning theory may be mediated by context-related influences on evaluation practice. Responses to these questions present challenges to evaluation researchers in future.

8.9 Conclusion

This chapter has discussed the responses of three New Zealand-based and three international evaluation experts to a summary of findings from the Q study and meta-evaluation. These findings were endorsed by five experts who work or have worked in New Zealand as reflecting their experience of local evaluation practice. Although the experts responded to different aspects of the findings, underlying themes about the macro-level influences on evaluation practice were able to be identified. Such themes included New Zealand’s pragmatic propensities, the influence of Māori and Pasifika ways of being, our intimate society, and the importance placed on relationships. Such influences contribute to evaluation being conceptualised as both an action-orientated undertaking and a relational practice. Evaluation theory is described as being under-emphasised, leading to evaluation practice that is described as informal and lacking rigour. Such features of evaluation practice have potential impacts on evaluative reasoning practice in the Aotearoa New Zealand public sector. This is explored further in chapter 9.

PART D

BRINGING THE PERSPECTIVES TOGETHER

Chapter 9 brings together the perspectives on evaluative reasoning practice revealed from the Q study findings (chapter 6), meta-evaluation (chapter 7), and expert interviews (chapter 8) to address the three research questions: (i) how is evaluative reasoning understood and practised by evaluators working in, or commissioned by, the Aotearoa New Zealand public sector? (ii) how do contextual factors influence how evaluative reasoning is practised in the Aotearoa New Zealand public sector? and (iii) how can evaluative reasoning practice be strengthened in the public sector context? Chapter 10 (the conclusion) describes the study's contribution to knowledge and identifies areas for further research.

CHAPTER 9

EVALUATIVE REASONING IN THE AOTEAROA NEW ZEALAND PUBLIC SECTOR

9.1 Introduction

As described in chapter one, reading the literature about evaluation quality led me to a hunch: evaluative reasoning contributes to quality evaluation. This study set out to explore the conceptualisation and practice of evaluative reasoning in the context of public sector evaluation in Aotearoa New Zealand, and to identify how evaluative reasoning practice could be strengthened. A multiple method research design was used to generate diverse understandings on the topic aimed at creating potential opportunities for abduction. The outputs of abductive inquiry are " . . . plausible hypotheses, thereby opening up the space for others (or yourself) to find something else of interest that raises a new curiosity" (Wolf, Peace & Brown, 2015, n.p.). In section 9.4 of this chapter I use the concept of conjecture to present abductively-derived ideas arising from the study findings. I chose this word after reading Timmermans and Tavory's (2012) reference to abduction as "the most conjectural" (p.171) of the three logics of deduction, induction and abduction. Used as a verb, conjecture means "to infer or arrive at (an opinion, conclusion) from incomplete evidence" (Collins, 2014), or "to conclude, infer, or judge from appearance or probability" (Oxford, 1991). While none of the thinkers in the field of abduction use the concept of conjecture in the way I propose, it seems to me there needs to be a term, perhaps even a sensitising concept (Patton, 2002a), through which such abductive findings can be framed.

The chapter begins with an explanation of how the key findings from the Q study, meta-evaluation, and expert interviews were examined together. Using inductive analysis, sections 9.2 and 9.3 compare the findings from the Q study, meta-evaluation and expert interviews to answer the two descriptive research questions: (i) how is evaluative reasoning understood and practised by evaluators working in, or commissioned by the Aotearoa New Zealand public sector? and (ii) how do contextual factors influence how evaluative reasoning is practised in the Aotearoa New Zealand public sector? The concept of the evaluation imaginary (Dahler-Larsen, 2012; Schwandt, 2009b) is used to address this second question. Using abductive analysis, section 9.4 addresses the third research question: how can evaluative reasoning practice be strengthened in the public sector context? The final section (section 9.5) draws on the study findings to explicate the connection between evaluative reasoning and evaluation quality.

9.2 Integrating the findings: making sense across the data

As described in the research design chapter (chapter 3), the design allows for some cross-data comparison through examining the findings from the Q study, meta-evaluation, and in-depth interviews. The key findings were distilled from these inquiries based on themes of interest (a top-down approach) and themes identified in the findings (a bottom-up approach) as summarised in Table 9.1. The key findings were then examined to identify confirmatory data, non-confirmatory/refutational data, patterns, and anomalies (Miles & Huberman, 1994; Patton, 2002a; Tavory & Timmermans, 2014). Data that were non-confirmatory and/or anomalous were then examined to identify potential explanations for such differences. The next stage of the analysis involved standing back from the key findings from the individual studies to understand the data as a whole. Table 9.1 acted as a visual artefact which supported this sense making across the three data collections. During this process I tried to become “defamiliarised” (Tavory & Timmermans, 2014, p.55) with the data to enable me to consider it with fresh eyes. This making sense of the whole took many weeks, during which time I went back and forth between the data from the individual studies and the higher-level picture that was developing. I also searched for additional literature to support (or negate) the emerging sense of the whole.

Table 9.1 Thematic comparison of key findings from individual studies

Themes	Q Methodology	Meta-evaluation	Expert interviews
Purpose of evaluation	All three orientations: The purpose of evaluation is to produce evaluative conclusions/ judgments.	The reports demonstrate authors' endorsement of the purpose of evaluation as being to make an evaluative conclusion/judgment (24 of the 28 reports contain one or more evaluative judgments/ conclusions).	Nil.
Evaluation as a values-based inquiry	All three orientations express evaluation as a values-based endeavour. The need for values to be made explicit is emphasised, such as cultural values, values associated with the evaluand, and values underpinning evaluation theory. All three orientations express how evaluators need to be aware of how their personal values influence their perceptions of the evaluand, its context, and stakeholder views.	13 of the 24 reports have an evaluative judgment/conclusion based on one or more value terms, defined explicitly as criteria or less explicitly in the form of another comparator. The remaining 11 reports use value terms: 6 reports have an evaluative judgment/conclusion based on one or more value terms that are not defined in the report; 5 reports have an evaluative judgment/conclusion about values terms that are not referred to elsewhere in the report.	Evaluators' awareness of values is attributed to the Treaty of Waitangi and its part in public policy.

Themes	**Q Methodology**	**Meta-evaluation**	**Expert interviews**
Evaluation as social practice	All three orientations portray evaluation as a social practice. The importance of relationships and dialogue with evaluation participants is stressed, as is the valuing of the expertise and knowledge that participants bring with them to the evaluation.	Stakeholders and/or the evaluation commissioner are described as being involved in the development of criteria in five of the nine reports that explain how criteria were arrived at.	The relational aspects of evaluation practice were identified as reflecting cultural norms.
Situated nature of evaluation; the role of context in evaluative reasoning	All three orientations: Express the notion of evaluation as being a situated activity. The context dependent nature of quality/merit/value is emphasised.	Generally, authors provide sufficient contextual information to enable understanding of the findings. However about half of the reports contain no information about limitations associated with the evaluation, such as limitations associated with the context.	Nil.
Additional features of evaluation practice	Orientation one: An eclectic approach is used to respond to contextual factors.	Nil.	Pragmatic and eclectic approach to evaluation. Generalist skills. Valuing of practical knowledge rather than theoretical knowledge.

Themes	Q Methodology	Meta-evaluation	Expert interviews
Role of evaluator	All three orientations: • The evaluator does not regard themselves as an expert about the evaluand. Rather, they value the knowledge and experience of stakeholders and programme recipients. • While stakeholders may be involved in the evaluation process, the evaluator (not stakeholders) produces the evaluative conclusion/ judgment. Orientations 1 and 3: The evaluator's responsibility to audiences with asymmetric power dynamics is endorsed. **Orientation 1:** The evaluator role involves being flexible and responsive to the different contexts in which they work. **Orientation 2:** The evaluator role is as an analyst. **Orientation 3:** Emphasises the judgment-making role of the evaluator more than the other two orientations.	While the evaluator role is not explicitly expressed in the evaluation reports, a few reports give the impression of one of the following: • The evaluator is the expert and expects the report audience to trust their evaluative conclusions/ judgments (despite a lack of clarity about how the evaluator arrived at them). • The evaluator role involves summarising and communicating stakeholder views (but not analysing or interpreting them).	The evaluator does not regard themselves as an expert who acts as "judge and jury". Rather the evaluator values the knowledge and expertise that stakeholders and participants bring to an evaluation. The evaluator has to be versatile in their role due to a lack of specialist opportunities. For Māori and Pasifika evaluators in particular, their role involves ongoing relationships and accountabilities to Māori and Pasifika communities. The public sector evaluator may act as a conduit between front-line providers or the recipients of public services and the government agency.

Themes	Q Methodology	Meta-evaluation	Expert interviews
Contextual influences on Aotearoa New Zealand and/or public sector evaluation practice	Orientation 1 and 3: The evaluator feels some constraint in their work due to the political context of public sector evaluation. Orientation 2: Accountability is an important purpose of public sector evaluation.	Nil.	The following influences were identified: New Zealand as a small, intimate society; its pragmatic heritage; influences of Māori ways of being on evaluation practices, for example, relationality emphasis; evaluator proximity to government clients and policy makers.
Influence of Māori epistem-ologies	Among the reasons identified by Q participants for the relationality of evaluation practice is that such relationality is congruent with a Māori way of being.	The authors of three reports stated that Kaupapa Māori approaches were used in the evaluation.	Māori evaluators are identified as leading the international evaluation community in culturally responsive evaluation, and through their contribution to the cultural validity discourse.

Themes	Q Methodology	Meta-evaluation	Expert interviews
Defensibility of evaluative claims; the role of argument in evaluative reasoning	Orientations 2 and 3: Emphasise the importance of defensible evaluation conclusions/ judgments. For the orientation 3 evaluator, defensibility comes from inclusive processes involving stakeholders in the evaluation process. For the orientation 2 evaluator, such defensibility comes from a compelling argument which engages the evaluation audience.	Seventeen of the 28 reports contain an argument (as defined by Booth et al., 2008). Of these, 13 use one or more warrants in an explicit or implicit way. 11 of the 28 reports either do not contain an argument or have text where it is not clear whether the text refers to evidence or argument.	Nil.
Evaluation quality issues	Nil.	Nine of the 28 reports lack three or more elements of evaluative reasoning. Five of these reports contain evaluative judgments despite having descriptive evaluation objectives or no evaluation objectives or questions. Half of the 28 reports contain no information about the limitations associated with the evaluation.	Evaluation practice is described as informal and lacking rigour. External scrutiny and challenge is described as lacking.

9.3 How evaluative reasoning is understood and practised

9.3.1 Introduction

Section 9.3 brings together and interprets the findings from the individual studies to answer two of the three research questions. It provides an inductively-derived, descriptive account of how evaluative reasoning is understood and practised by some evaluators working in or for the public sector (research question one), and the contextual factors that influence this practice (research question two).

9.3.2 Overview

This section provides an overview of the findings from the Q study, meta-evaluation and expert interviews. Firstly, all three orientations in the Q study are attuned to evaluative reasoning concepts. The differences expressed in the three orientations can be described as nuanced heterogeneity - they are differences in focus and emphasis, rather than dissimilarities of a more fundamental nature. As proposed in chapter 6, some of this difference may be explained by whether the evaluator is working inside a government agency, or outside as a contractor. In contrast, the meta-evaluation findings provide a significantly more variable account of evaluative reasoning practice. On the one hand there are reports which demonstrate a systematic chain of reasoning from evaluation objective to evaluative conclusion/judgment. In contrast, there are reports which end with an evaluative conclusion/judgment despite having descriptive evaluation objectives or questions, and other reports in which the values used to assess the evaluand are stated but not defined.

Despite this difference between the Q orientations and meta-evaluation findings, five of the six evaluation experts endorsed the findings as being generally congruent with their experience of local evaluation practice.

9.3.3 Comparing Q and meta-evaluation findings

Comparing the findings of the Q study and meta-evaluation reveals a significant puzzle - why do the three Q orientations present a relatively coherent portrayal of evaluative reasoning concepts (albeit with differences in focus and emphasis), while the meta-evaluation findings provide a more divergent picture of evaluative reasoning practice? This section speculates on two possible explanations for this difference.

The first speculation derives from Argyris' (1976) constructs of espoused theory and theory-in-use. Argyris argues that people's espoused position or perspective on a particular issue may differ from their actual behaviour in relation to that issue. Therefore the Q participants could be thought of as expressing their espoused perspectives about evaluative reasoning in the Q sort. In contrast, the reports in the meta-evaluation could be regarded as being the manifestation of the authors' theory-in-use. This offers a fruitful topic for further research, for example, comparing individual evaluators' perspectives on evaluative reasoning via a qualitative interview, with how such reasoning is manifest in the evaluation reports they have authored.

While the espoused theory/theory-in-use explanation is plausible, an alternative speculation is offered arising from considering the Q study participants and report authors in the meta-evaluation. I turn first to the Q participants. As explained in chapter 6, while participant demographics are not a primary interpretative tool in Q as in R, participant information may be used to provide additional insight into the orientations once they have been abductively identified and described. Using purposive sampling as is appropriate in Q (see chapter 6), the Q participants were recruited via the Wellington Evaluation Group (15 participants) and from my personal evaluation network (15 participants, all of whom are members of one of the professional evaluation associations or are affiliated to the Wellington Evaluation Group). Consequently, all of the Q participants are part of either a formal or informal evaluation network. In contrast, of the principal authors of the 23 named reports in the meta-evaluation, 14 appear to be neither affiliated to an evaluation association such as ANZEA, AES and AEA, nor subscribe to the Wellington Evaluation Group. While the report sample used in the meta-evaluation is not representative, it suggests that practitioners undertaking public sector evaluation may be a diverse group with varied professional and theoretical backgrounds. Further, the meta-evaluation findings suggest that some of these practitioners may be working outside of the umbrella of a professional evaluation network. It is also surmised that some of these professionals may be less likely to have been exposed to evaluative reasoning theory and understand its implications for evaluation practice. There may be alternative theorising occurring within this cluster of professionals which is not aligned to evaluative reasoning theory as presented in chapter 5. An example of alternative theorising is evaluation undertaken as a research exercise in which evaluative elements are treated as secondary to the research elements or are non-existent (E. J. Davidson, personal communication, 19 May 2015).

This cluster of evaluation practitioners is contrasted to those professionals working inside and outside government agencies who are members of one or more of ANZEA, AES and AEA, and/or subscribe to the Wellington Evaluation Group. Such professionals may be part of the cohesive and collaborative evaluation community described by two of the evaluation experts as reported in section 8.7.2. It is assumed that such membership means these professionals are more likely to self-identify as evaluators, evaluation commissioners or have some other professional interest in evaluation, attend networking events and professional development activities (for example, evaluation seminars, workshops and conferences run by ANZEA, AES and WEG). Such professionals are assumed to be more likely to have been exposed to theorists' ideas, particularly those of Dr Jane Davidson whose work has contributed significantly to the evaluative reasoning discourse in Aotearoa New Zealand. (Dr Davidson has worked for a range of government agencies as well as speaking at ANZEA and AES conferences and running workshops and seminars).

This leads to the speculation that there may be different clusters of professionals undertaking public sector evaluation which, in turn, may help to explain the difference between the Q orientations' expression of evaluative reasoning concepts, and the less coherent picture about evaluative reasoning practice in the meta-evaluation. One cluster may comprise a network of professionals working inside and outside government who self-identify with the evaluation profession (as evaluation practitioners or commissioners), and are affiliated to a formal or informal professional evaluation network. Another cluster may comprise a more disparate group of practitioners working in a range of professional areas (for example, management, economic research, engineering, and academics working in health, education, social work or other social science fields) who are not affiliated to one of the evaluation associations. This speculation requires systematic examination to determine whether such clusters of professionals do in fact exist, and to identify potential impacts (if any) of such clusters on how evaluative reasoning is practised.

It is acknowledged that professional affiliation may not be the only dimension of difference among professionals undertaking public sector evaluation. There may be additional clusters based on other dimensions, for example, subject area (an economic cluster, a health assessment cluster), location (professionals working in/outside of Wellington, the centre of government), a Māori evaluation cluster, and a Pacific Island evaluation cluster. Such differences are explored further in section 9.3.5 using the concept of the evaluation imaginary (Dahler-Larsen, 2012; Schwandt, 2009b).

9.3.4 Description of evaluative reasoning understanding and practice

This section offers a description of how evaluative reasoning is understood and practised by some professionals undertaking public sector evaluation, based on inductive analysis of the findings of the Q study, meta-evaluation and expert interviews. The description provided below is limited to professionals who self-identify with the evaluation profession and are affiliated to it. The divergent picture of evaluative reasoning practice that emerged from the meta-evaluation means that the following description may not be relevant to all professionals undertaking public sector evaluation, particularly those working outside of a professional evaluation umbrella.

Evaluation as a values-based inquiry

All three Q orientations describe evaluation as a values-based inquiry. The orientations express the need for values to be made explicit, whether they are the values against which the evaluand is being assessed, the values (explicit and implicit) in the context of the evaluand, or the values inherent in evaluation theory. The three orientations also express (to a greater or lesser extent) the need for evaluators to be aware of how their personal values influence their perceptions of the evaluand, its context and stakeholder perspectives. The majority of report authors in the meta-evaluation also endorse evaluation as a values-based inquiry - 17 of the 28 reports in the meta-evaluation have one or more value terms in their evaluation objectives/questions, while a greater number of reports (24) end with one or more evaluative conclusions/judgments containing value terms. As noted in chapter 7, some of the reports ending with one or more evaluative conclusions/judgments have non-evaluative evaluation objectives/questions - an impossibility according to informal logic.

Relational approaches underpin evaluative reasoning

The Q orientations portray evaluation in Aotearoa New Zealand as a social practice, and evaluative reasoning practice as being based on a relational approach. Dialogic processes are used for the development of criteria (although such processes do not extend to the making of evaluative judgments/conclusions, as discussed below). Dialogic processes are portrayed in the Q study not as an optional extra, but as being fundamental to what it means to do good evaluation, for example, understanding what merit/worth/significance mean in a particular context, and accessing stakeholders' knowledge and expertise about an evaluand and its context. The relational nature of evaluation practice in Aotearoa New Zealand was supported by five of the experts who described it as reflecting cultural norms. Turning to the meta-evaluation, it is important to note that relational practices used by an evaluator may not be evident in an evaluation report. However some report authors do describe the relational practices they used in the evaluation, such as using a participatory method and specifically, involving stakeholders in developing criteria.

Centrality of judgment-making to evaluation

The three Q orientations endorse the purpose of evaluation and the responsibility of the evaluator as being to produce evaluative conclusions/judgments, rather than descriptive accounts of the evaluand for others to assess. Given that 24 of the 28 reports examined in the meta-evaluation contain one or more evaluative conclusions/judgments, one can assume these authors also endorse the purpose of evaluation as being to produce evaluative conclusions/judgments. The dialogic processes described above do not extend to the making of evaluative conclusions/judgments - this is portrayed in the three Q orientations as the role of the evaluator, not stakeholders. Similarly, there was no evidence in the meta-evaluation of stakeholders having had a role in judgment-making.

Variability of evaluative reasoning practice

As noted in chapter 8, four of the experts made observations about evaluation practice in general being less systematic and robust in Aotearoa New Zealand (compared to other countries which place greater emphasis on evaluation theory), as well as lacking scrutiny and challenge. Such issues were demonstrated in some of the meta-evaluation reports, particularly the nine reports with two or fewer elements of evaluative reasoning. Such shortcomings not only reflect the evaluative reasoning capability of the professionals who conducted the evaluation (whether government employee or contractor) but also that of the evaluation commissioner in the government agency that contracted the evaluation. This leads to the next topic - the area of evaluative reasoning practice in need of attention.

Area of evaluative reasoning practice requiring attention

As demonstrated in the meta-evaluation, strengthening the quality of probative inference through argument appears to be an area of evaluative reasoning practice in need of attention. The approaches used by some report authors in the meta-evaluation to support their evaluative claims - the "trust me, I am the evaluation expert" approach, the summary of stakeholder perspectives, and authors' assertions about the strength of their evidence - are not sufficient. Given the emphasis on evidence in the public sector policy space over the last 15 years as described in chapter 7, it is perhaps unsurprising that some evaluators may regard evidence in itself as adequate for drawing conclusions and making judgments about the nature of change.

This study has explicated the critical connection between evaluative claim and argument - because evaluative claims are based on probative inference, such inference needs to be underpinned by convincing argument. As explained in chapter 5, argument relates evidence to criteria. Supporting the argument through the inclusion of appropriate warrants gives greater weight to the claim. Given the negative connotations of the term argument for two of the Q participants, an alternative conceptualisation of an argument is that of building up a case to support an evaluative claim in the way a lawyer builds a case to convince a jury. For the purposes of this study, the term argument is used. As will be argued more fully in section 9.4.4, strengthening the quality of probative inference through warranted argument is not an option for public sector evaluators. The consequential nature of evaluative judgments/conclusions in the public sector context means evaluative claims need to be robust and defensible (Greene, 2011). Lastly, it is unsurprising that argument is identified as an element of evaluative reasoning practice requiring attention given that it may be regarded as the most intellectually challenging aspect of an evaluation (House, 1977).

9.3.5 Contextual factors influencing how evaluative reasoning is practised

Drawing on the findings from the expert interviews and relevant literature about Aotearoa New Zealand presented in chapter 2, this section addresses the second research question: how do contextual factors influence how evaluative reasoning is practised in the Aotearoa New Zealand public sector? To answer this question, I draw on the construct of the evaluation imaginary, defined as: "The views and assumptions undergirding evaluation . . . (that are) themselves undergirded by broader views, norms and values in society" (Dahler-Larsen, 2012, p. 27). Schwandt's (2009b) definition emphasises the socially constructed nature of the evaluation imaginary: " . . . that common, intersubjective or social understanding that makes possible common (evaluation) practices and a widely shared sense of legitimacy for contemporary evaluation practice" (p.22). The evaluation imaginary in a particular place and time is not static but is "a dynamic, continually produced and reproduced narrative" (Dahler-Larsen & Schwandt, 2012, p.81). Far from simply responding to the evaluation imaginary, evaluators contribute to its construction (Dahler-Larsen & Schwandt, 2012): "Evaluators and evaluations do not simply identify and respond to contextual factors, but by virtue of their action are always constructing, relating to, engaging in, and taking part in some reconstruction of the context in which they operate" (p.84).

Therefore conceptualisation and practice of evaluative reasoning is located within the evaluation imaginary, and as such is influenced by it. This relationship can be portrayed as three concentric circles, with evaluative reasoning conceptualisation and practice located at the centre (Figure 7). Given the premise of this study, that evaluative reasoning forms the core of what it means to do evaluation, evaluative reasoning is located in the centre circle. How evaluation reasoning is conceptualised and practised is shaped by the evaluation imaginary for a particular place and time (middle circle), which itself reflects the wider societal norms and values to which Dahler-Larsen refers (2012) (outer circle).

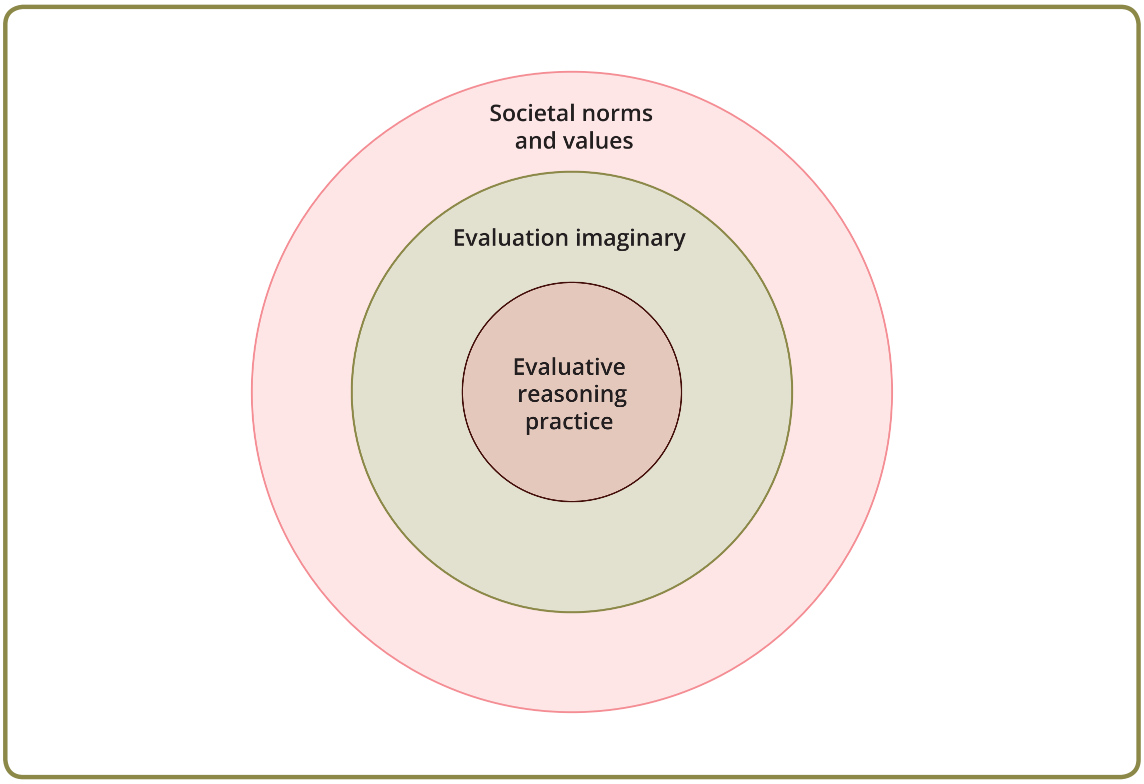


Figure 7 Evaluative reasoning within the evaluation imaginary

(Source: Diagram constructed from exposition in Dahler-Larsen, 2012; Dahler-Larsen & Schwandt, 2012; Schwandt, 2009b)

The notion of the evaluation imaginary articulated by Dahler-Larsen and Schwandt provides a useful construct for understanding how contextual factors and evaluator practice interact, and as a consequence, how professional evaluation may evolve over time in a particular society. However these authors' discussion does not address a number of questions that emerge from considering the construct of the evaluation imaginary with respect to a particular society, such as Aotearoa New Zealand. For example, the notion of an evaluation imaginary assumes a hegemonic society dominated by a single epistemology. Such an imaginary is less reliable when evaluation is shaped by diverse epistemic and cultural traditions. This is illustrated in the section below about the influence of Māori epistemologies on public sector evaluation practice. Further, there is the potential for alternative evaluation imaginaries to exist in a particular society based on the evaluation setting, for example, the evaluation imaginary in a public sector setting may differ from the evaluation imaginary in a community setting. Different evaluation imaginaries may also exist within a setting, for example, different parties within a particular setting may have differing views on what the evaluation imaginary is, or should be.

Four contextual factors are now identified that may be regarded as contributing to the shaping of evaluation in Aotearoa New Zealand in general, and public sector evaluative reasoning practice in particular. The four factors are: history, place and people; Māori epistemology; ANZEA; and the Chief Science Advisor. Four potential evaluation imaginaries are proposed as part of this discussion.

History, place, and people

As discussed in chapter 8, five of the evaluation experts identified factors relating to history, geography, and demography as contributing to how evaluation is practised in Aotearoa New Zealand (and therefore how evaluative reasoning is understood and practised in the public sector). The geographic isolation and small population of Aotearoa New Zealand leads to an intimate, relationally-based society (Salmond, 2012). This is reflected in the way public sector evaluators are more likely to have relationships with evaluation commissioners and policy-makers than in larger countries where larger-scale evaluation contracts are more common (Williams, 2003). It is also reflected in relationships with evaluation stakeholders as noted by White and Boulton (2011): “Evaluation in Aotearoa New Zealand sometimes feels like paddling in a shallow lagoon, one in which all evaluation stakeholders (past, future and present) are highly visible to us - and us to them” (p.73). Further, the experts noted that the size of the public sector requires public sector evaluators to be generalists who are eclectic in their approaches as there is little scope for specialisation. Experts’ comments about evaluation experience being valued more than theoretical knowledge may be seen as a reflection of the pragmatic traditions of Aotearoa New Zealand described in chapter 8. This in turn may explain why some experts refer to local evaluation practice as informal, less systematic and lacking robustness.

Māori epistemologies

While Māori epistemologies are outside the scope of this study, their presence and influence are evident in the findings. They are reflected in the comments of Q participants who are Māori emphasising the importance of relationships and dialogue kanohi ki kanohi (face to face), and the valuing of the practical experience and knowledge of the people who are affected by a programme. The use of Māori epistemologies is also illustrated in three of the reports examined in the meta-evaluation. The authors of these reports describe the Kaupapa Māori approach used to conduct their evaluation as summarised in chapter 7. The influence of Māori ways of being on evaluation practice was also noted by five of the six evaluation experts, particularly the emphasis on values in evaluation and the importance of culture in ensuring evaluation validity.

We now move from these specific findings to consider how Māori epistemology may influence evaluation practice generally and evaluative reasoning practice in particular. Three potential evaluation imaginaries for Aotearoa New Zealand are now described. The contested nature of an evaluation imaginary means that these suggestions are offered tentatively as a way of stimulating debate.

The first evaluation imaginary, referred to here as the Māori evaluation imaginary, is based on a distinction between evaluation according to Māori epistemologies as expressed in Kaupapa Māori evaluation described in section 2.5 and that based on western epistemology. A recent paper about wairuatanga (spirituality) in evaluation practice (Kennedy, Cram, Paipa, Pipi & Baker, 2015) illustrates an aspect of the Māori evaluation imaginary. For Māori, wairuatanga (spirituality) " . . . is threaded through beliefs, values and practices . . . and is an essential component of Māori wellness" (p.88, 89). The paper describes aspects of wairuatanga underpinning the practice of a group of Māori evaluators, and implications for how the evaluators go about their day to day work, for example, as expressed in "rituals of encounter" (p.95) and the building and maintaining of relationships with evaluation participants.

A second evaluation imaginary is proposed which does not involve a strict division between that which is Māori and that which is Pākehā. In order to understand this evaluation imaginary, we must first focus on the influence of Māori epistemology and tikanga Māori (Māori custom) on Pākehā culture. As an insider within Aotearoa New Zealand society, it is difficult to recognise those things in Pākehā culture that can be attributed to the influence of Māori epistemology and tikanga. It was therefore interesting to hear the outsider perspective of an evaluation expert (residing overseas) reported in chapter 8 who talked about " . . . a New Zealand way of being which is hugely influenced by a Māori way of being". This influence is also described in the literature, for example, M. King (2003):

> Pākehā culture continues to borrow and learn from Māori . . . it has taken words and concepts (mana, tapu, whānau, taonga, haka, tūrangawaewae),[18] attitudes (the tradition of hospitality which, in the early nineteenth century, was so much more visible from the Māori side of the frontier than the Pākehā), ways of doing business (an increasing willingness to talk issues through to consensus in preference to dividing groups 'for' and 'against' a given motion), and rites of passage (a loosening up of formerly formal and highly structured funeral services) (p.519).[19]

In the first Sir Paul Reeves Memorial Lecture in August 2012 titled *Beyond the Binary*, historian Dame Anne Salmond describes the breaking down of the strict division between that which is Māori and that which is Pākehā: "Rather than seeing Māori and Pākehā . . . as bi-polar opposites with some kind of Berlin Wall between them, (they are) increasingly regarded as complementary pairs joined together by a fertile middle ground" (p.9).

A personal experience confirms the observation that visiting another country can often bring increased awareness of one's own culture - in this case, insights about the influence of Māori culture on Pākehā New Zealand when I attended my first AEA conference in 2012. During the first session of the conference I was surprised by what I perceived as the abrupt, down to business manner in which the conference started. My surprise was such that after the session I mentioned to a New Zealand colleague that I had missed the conference welcome (presuming it had occurred before I arrived). My colleague reassured me that this was not the case. By way of explanation, in Māori culture the pōwhiri (welcome) of visitors onto a marae (the complex of buildings around the wharenui, the main building of a marae where formal greetings and discussions take place) is an important ceremony involving a number of protocols (Mikaere, 2013). My experience of the AEA conference opening made me reflect on how the intent of the pōwhiri - the welcoming of visitors - has become part of Pākehā culture. The welcoming of

[18] Mana: prestige, authority, control, power, influence, status, spiritual power, charisma; Tapu: that which is sacred, prohibited, restricted, set apart; Whānau: extended family, family group; Taonga: treasure, anything prized. Applied to anything considered to be of value including socially or culturally valuable objects, resources, phenomenon, ideas and techniques; Haka: vigorous dances with actions and rhythmically shouted words; Tūrangawaewae: place where one has the right to stand. Place where one has rights of residence and belonging through kinship and whakapapa.

[19] Macrons were not included in the original text.

visitors, particularly overseas visitors, in formal (and some informal) gatherings transcends the very business-like approach that was my experience of the AEA conference.

The observations of M. King (2003) and Salmond (2012), together with the expert's comments lead me to suggest that a second potential conceptualisation of an evaluation imaginary (at a generalised level) may not necessarily be portrayed as a division between that which is Māori and that which is Pākehā. Rather, this second evaluation imaginary is one that may be described as Aotearoa New Zealand, namely, evaluation that is relational, values-based, practice-orientated and judgment-focussed. This evaluation imaginary reflects the societal trend described by M. King (2003) and Salmond (2012) whereby Māori epistemology and culture are influencing (albeit, subtlety) the way evaluation (and therefore evaluative reasoning) is understood and practised. Drawing on the words of the evaluation expert, this evaluation imaginary describes an **Aotearoa New Zealand way of doing evaluation**.

These are not the only potential evaluation imaginaries for Aotearoa New Zealand. A third potential conceptualisation is an evaluation imaginary based on Pasifika epistemologies (also outside the scope of this study). This is a developing discourse as the number of Pasifika evaluators in Aotearoa New Zealand increases. Some of this discourse has been sparked by studies undertaken by Pasifika doctoral students (for example, Vaioleti (2006)). This imaginary is illustrated in an education context by Fotuali'i McGeady (2015) who articulates the importance of Va relationships in evaluation. Va refers to the space between. This space is not empty or void but is relational and sacred (Fotuali'i McGeady, 2015, n.p.). Fotuali'i McGeady stresses the importance for evaluators to respect and maintain Va in their practice.

ANZEA: Influencing how evaluation is understood and practised

Since its inception in 2006, ANZEA has contributed significantly to the construction of an evaluation imaginary which reflects the Aotearoa New Zealand context. This contribution is reflected in the principles underpinning *ANZEA Evaluator Competencies* (2011) and the *Evaluation Standards for Aotearoa New Zealand* (Social Policy Evaluation and Research Unit & ANZEA, 2015), as discussed below. The Treaty of Waitangi is described as providing the founding principles for evaluator engagement (*Evaluator Competencies*, principle one). Values are identified as an integral part of evaluation (*Evaluator Competencies*, principle two). Cultural values, and consequently cultural competence, are described as being central to evaluation (*Evaluator Competencies*, principle three). The relationality and participatory nature of evaluation practice is emphasised (*Evaluation Standards*, principle one):

> Honest, sincere, respectful, reciprocal and meaningful relationships (individual and collective) are built, nurtured and maintained with the people, organisations and communities involved in and affected (or likely to be affected) by evaluation. The people, organisations and communities (or their representatives) involved in and affected (or likely to be affected) by evaluation are appropriately informed and involved (p.19).

The dynamic nature of the evaluation imaginary referred to by Dahler-Larsen and Schwandt (2012) in which evaluators contribute to its ongoing evolution is also reflected in ANZEA's annual conference, and specifically in workshops and seminars where evaluators share about their practice. The recent publication of the inaugural edition of the journal *Evaluation Matters - He Take Tō Te Aromatawai* will also contribute to this shaping.

Influence of the Chief Science Advisor

In the Aotearoa New Zealand public sector context, "the legitimacy . . . for evaluation practice" (Dahler-Larsen & Schwandt, 2012, p.22) is provided by the evidence-based policy discourse as discussed in chapter 2. The role of evaluation in the evidence-based policy discourse was emphasised by visiting speakers at the opening and closing events held to mark the 2015 International Year of Evaluation. At the opening event, the Honourable Bill English, Minister of Finance spoke about the Government's need for evidence to support its decision-making, referring at various points to the need for facts. This did not go unnoticed by a later speaker, Associate Professor Robin Peace (an evaluation academic) who gently reminded the audience that evaluators work with values.

As discussed in section 2.3, the Government's Chief Science Adviser's role involves improving the quality of evidence used in public policy-making. Professor Gluckman promotes value-free social science for public policy purposes, equating values with bias and subjectivity. This stance has the effect of “neutralizing values” (Eisner, 1996, p.x), thereby concealing the values-based nature of political activity (House, 2004b). Moreover, his stance undermines the legitimacy and contribution of evaluation as a tool for the assessment and development of public policy.

Consequently, it is suggested there are at least two potential evaluation imaginaries for the Aotearoa New Zealand public sector which are predicated on opposing discourses - firstly, an evaluation imaginary based on the value-free discourse as articulated by the Chief Science Adviser where value-free evaluative evidence and evaluative conclusions are sought. The second evaluation imaginary is based on a “values-imbued” (House, 2004b, p.7) approach to evaluation as articulated by ANZEA and described by some of the evaluation experts as the Aotearoa New Zealand way of doing evaluation. The disjuncture between the two evaluation imaginaries has yet to be confronted, however challenging it may be to do so. If this disjuncture remains unresolved, it has the potential to inhibit the development and contribution of evaluation to the New Zealand public sector.

9.4 How evaluative reasoning practice can be strengthened

9.4.1 Introduction

This section responds to the third research question *how can evaluative reasoning practice be strengthened?* by presenting four abductively-derived conjectures. Three conjectures concern evaluative reasoning practice in a general sense, and one relates to evaluative reasoning practice in the public sector context. The section begins with an overview of abductive analysis as articulated by Tavory and Timmermans (2014).

9.4.2 Abductive analysis

Drawing on Peirce’s logic of abduction, Tavory and Timmermans (2014) have articulated how theorisation and observation are connected through abductive inference, in a process they refer to as abductive analysis. The primary aim of abductive analysis is theory generation. Tavory and Timmermans (2014) define abductive analysis as a “ . . . systematic process of meaning-making aimed at theoretical generalizations” (p.123). Given the use of the term analysis, one might assume that abductive analysis is something that happens at the latter end of the inquiry process after the data have been collected. While acknowledging that all research involves planning, Tavory and Timmermans (2014) emphasise that abductive insights do not occur serendipitously at the tail-end of an inquiry but must be deliberately planned for from the outset. Such planning has a number of aspects. Firstly and most importantly, theory plays a critical role in abductive analysis because unexpected findings are only a surprise if they are unable to be located into an existing theoretical frame. New insights are drawn out from existing theory: “Developing new theories depends on the researcher’s inability to frame findings in existing theoretical frameworks, as well as on the ability to modify and extend existing theories in novel ways” (p.41). Therefore abductive analysis requires the researcher to have in-depth knowledge of a broad range of relevant theory and “extensive theoretical preparation” (p.49).

Secondly, an inquiry’s research design should aim to create potential opportunities for unexpected findings, anomalies and puzzles which are the necessary ingredients for abductive inference. Such opportunities are realised through the research design: “Researchers design research to cultivate opportunities for abduction . . . (they) foster empirical surprises” (p.123). The researcher should “ . . . aim for variation” (p.126) and a “comparative agenda” (p.125) in the research design through the use of diverse methods and the collection of different data. Thirdly, Tavory and Timmermans (2014) stress the important role played by the researcher’s peers and the wider research community in the development of abductive theories. They use a term derived from Peirce, the “community of inquiry” (p.103) to describe this community. They encourage researchers to share their thinking over the course of their research with the community of inquiry for testing and refinement: “Allowing others access to the relationship between data and theory early on is crucial, both as a way to avoid future mistakes and to push the research in a different direction” (p.111). Having reached the later stages of their inquiry, the researcher exposes their abductive claims to the community of inquiry for examination and critique which may result in further enhancements (or the need to rework their thinking). As noted in chapter 3, Tavory and Timmermans (2014) identify three criteria for testing abductive claims: fit, plausibility, and relevance. The authors outline strategies for researchers to strengthen the fit, plausibility and relevance of their abductive claims, as summarised in Table 9.2.

Table 9.2 Strategies to strengthen abductive claims

Criterion	Key questions to address	Strategies
Fit	Does the evidence support what is claimed? (p.105). How do the observations connect to the theorization? (p.110). Is it a convincing fit? (p.106).	Transparency of the research is essential. The researcher should aim for "increased transparency in claims making" (p.107) through presenting some of the evidence to the community of inquiry and research audience (p.106). This allows people to judge for themselves the extent to which the evidence supports the claim(s) (p.106).
Plausibility	What are the alternative theoretical explanations and do they make sense? (p.111). What makes the researcher's abductive inference compelling over alternative explanations? (p.112).	The researcher should examine other plausible theoretical possibilities to explain the observed data (p.113). The suppositions and assumptions that such possibilities involve should be made explicit (p.114). The researcher's claims are exposed to questioning by the community of inquiry. This may result in other plausible explanations for consideration (p.113). New data is gathered if necessary (p.113).
Relevance	So what does this mean? (p.115). What makes the research worth the effort? (p.115).	Abductive inquiry should be useful: "(It) needs to be evaluated for its potential practical effects . . . its ability to lead to practical commitments and actions" (p. 115). Tavory and Timmermans note that "there are no methodological guidelines" (p.115) for researchers to follow to ensure relevance.

Source: Table complied by the author from material presented in Tavory and Timmermans (2014, p.105-115).

It should be noted that the strategies identified by Tavory and Timmermans (2014) to strengthen the fit and plausibility of abductive claims are not dissimilar to those identified by Booth et al., (2008) to support inductively-derived research claims (as described in chapter 7), namely, reasons supporting the claim, evidence supporting the reasons, and acknowledgment of and response to alternative explanations.

9.4.3 Use of abductive analysis

A study examining the use of abduction in the research process (Dobson, Gengatharen, Fulford, Barratt-Pugh, Bahn & Larsen, 2012) describes “a eureka moment” (p.8) experienced by researchers involved in four separate case studies. This revelation resulted in a significant understanding or insight for the researcher. My use of abductive analysis to make sense of the findings facilitated meaning making that was more incremental and emergent in nature. It involved an ongoing process of moving between the findings and the literature over a period of about three months, both in a deliberate fashion (in the sense of focusing on the question “’what does this mean?”) and in a less conscious manner (via the type of thinking that occurs in the course of daily activity and during times of relaxation). This resulted in new understandings emerging gradually over time which I discussed and tested with my supervisors to ensure they met the fit, plausibility and relevance criteria identified by Tavory and Timmermans (2014). Four conjectures are now presented.

9.4.4 Conjecture 1: Evaluative reasoning is a key element of the craft of evaluation

Evaluation is often conceptualised as a technical or management practice (Stern, 2006) shaped in large part by scientific and technical notions (Schwandt, 2002b). A few evaluation theorists (notably Weiss, 1998) have challenged this instrumental framing of evaluation, asserting that evaluation is a craft. This leads to the first conjecture, namely, **if evaluation is a craft, evaluative reasoning is a key element of the craft.**

To introduce this conjecture, we first examine the notion of craft. Writing about craftsmanship, Sennett (2008) a sociologist, describes craft as “ . . . involving dimensions of skill, commitment and judgment” (p.9) and craftsmanship as focusing “ . . . on the intimate connection between hand and head. Every good craftsman conducts a dialogue between concrete practice and thinking” (ibid). The notion of practical knowledge underpins craft and craftsmanship (Sennett, 2008) defined as “ . . . the things a person knows in relation to his or her own behaviour or situation but cannot necessarily express” (Oxford Dictionary of Sociology, 2015, n.p.). Practical knowledge is embedded knowledge in that it has become “routinized” (Sennett, 2008, p.50). Schwandt (2002b, 2008b) and Stake and Schwandt (2006) have developed a case for the role of practical knowledge in evaluation. In the evaluation context, practical knowledge encompasses "perception, discernment, insight, practical wisdom" (Schwandt, 2008b, p.30,34,35). Practical knowledge is not limited to the cerebral (Stake & Schwandt, 2006): " . . . this practical embodied knowledge - that is at once both cognitive and emotional - is a source both of our ability to discern quality and our efforts to ascribe meaning to the quality we see in an evaluand" (p.408).

Having proposed that craft may be considered through the lens of practical knowledge, I now use the study findings to support the assertion of evaluation as a craft and to argue that evaluative reasoning is a key element of this craft.

My study has demonstrated the "value-laden" (House, 2004b, p.7) nature of professional evaluation. As discussed in chapters 4 and 5, the purpose of evaluation is to make value claims about the merit or worth of an evaluand. This purpose was endorsed by the majority of the report authors in the meta-evaluation who provided evaluative conclusions/judgments (despite some of these conclusions/judgments lacking the elements that make them defensible), and in the Q study's orientation two (the analytic evaluator) and orientation three (the judgment-centred evaluator).

Further, values are inherent in the policies and programmes that are the focus of evaluation (Schwandt, 1997), the context surrounding the evaluand (Greene, 2005), and the evaluation methods used (N. L. Smith, 2010). The evaluator must be cognisant and take account of the potentially conflicting social, cultural and political values associated with a particular evaluand and its context (Greene, 2005). This requires the evaluator to surface embedded and taken-for-granted values as is described by the analytic evaluator (orientation two). Competing values may require the evaluator to make trade-offs, as described by the context responsive evaluator (orientation one) who feels constrained by the political context in which they work.

The evaluator must also be aware of how their personal values influence their professional practice as emphasised (to a greater or lesser extent) in all three Q orientations. The value-laden nature of evaluation requires the professional evaluator to consider how they will address issues of value in their practice. Referring to Schwandt's (2002b) "ideal types" (p.145) described in chapter 5, the evaluator may choose to act in a "value neutral" (p.145) manner, describing and analysing stakeholders' perspectives but not coming to an evaluative conclusion about them. Another evaluator may choose an advocacy and change agent role implicit in the "value-committed" (p.148) type. Yet another evaluator may choose to be a critical friend associated with the "value-critical" (p.151) type.

Paraphrasing Schwandt (2002b), the evaluator must consider the question: In whose interests should I be acting and for what purpose? Two further questions emerge from this question, namely, what are my responsibilities and accountabilities? What is the nature of my obligations and to whom? The personal response required from the professional evaluator to these questions takes the conceptualisation of evaluation practice beyond the notion of a technical or managerial activity. Rather, evaluation practice involves personal values and principles, and self-acuity. It is for this reason that Schwandt (2002b) describes evaluation as a "moral-political undertaking" (p.23). Similarly, House (1977, 2004a, 2004b) emphasises the moral responsibility of evaluators to recognise and address the political contexts in which they work.

Conceptualising evaluation as a moral-political undertaking has implications for the way in which the practice of evaluative reasoning is understood. If evaluation is a moral-political undertaking, then the practise of evaluative reasoning is more than a procedure or technique that can be easily acquired through attending a short course about how to develop evaluation rubrics. Similarly, it is more than something that is captured in a diagram or heuristic as used in the meta-evaluation. Rather, the practise of evaluative reasoning is a key component of the craft that is evaluation (Weiss, 1998).

Turner's (1994) writing on professional practice provides insights into the craft of evaluative reasoning by professional evaluators. Turner uses the legal profession to illustrate the expertise involved in professional practice.

> The problem . . . is (that) of getting judicial decisions out of books of laws. The act is performed with a higher degree of consistency among the trained and experienced than the untrained and inexperienced. It requires knowledge, or something like knowledge, that is itself not in any books or sets of explicit rules. Rudolph von Ihering appealed to the notion of a 'judicial sense' arising from experience in order to account for judges' abilities to do so. Today, (such) legal knowledge is a practical

> problem for artificial intelligence practitioners attempting to model legal reasoning (p.101).

Paraphrasing Turner's (1994) words, the problem of arriving at defensible evaluative conclusions/judgments about value/merit/worth of a particular evaluand in a particular context is the problem of understanding what value/merit/worth mean for the evaluand in its context, appraising relevant evidence against criteria or other comparator, building an evaluative case through deliberation and argument, and providing a transparent chain of reasoning from criteria to evaluative conclusion/judgment. All of these require knowledge that is not contained in any evaluation checklist, rubric development guide, or evaluation text. Rather, such knowledge can be thought of as evaluative sense - this is the craft of evaluative reasoning practised by evaluation professionals.

The question arises - what distinguishes the craft of evaluative reasoning of evaluation practitioners from the judgment-making of judges in determining the guilt of an accused, or that of doctors assessing a patient's symptoms? Based on the study findings and the review of literature, the following elements are identified as constituting the craft of evaluative reasoning.

- **Relational skills** to identify and understand perceptions of quality/value/merit of diverse stakeholders in relation to a particular evaluand and its context.
- **Context sensitivity** to define quality/value/merit (in the form of criteria or other comparator) in relation to a particular evaluand and its context.
- **Political acumen** to recognise and manage the political nature of evaluation practice.
- **Deliberative skills** to appraise evidence in light of the criteria or other comparator.
- **Argumentation skills** to present a defensible case that links claims and evidence to criteria or other comparator.
- **Astuteness** to ensure all relevant positions and perspectives have been considered in the assessment of quality/value/merit.
- **Reflexivity** about the evaluator's personal values and their impact on the deliberative process.
- **Discernment** to synthesise findings into evaluative conclusions/judgments.

- **Open-ness and humility** to reflect the contingent nature of evaluative conclusions/judgments.

For me, as for other professional evaluators, improving our evaluative reasoning craft is a matter of practice and experience. As a consequence, it is a journey rather than a destination.

9.4.5 Conjecture 2: Expert intuition offers a way of knowing for evaluative reasoning practice

Conjectures two and three concern the nature of discernment in professional evaluation, expressed colloquially as, how evaluation practitioners know what they know about an evaluand. To discern is to "recognise or perceive clearly" (Collins, n.d.). As described in section 9.4.4, for Stake and Schwandt discerning in professional evaluation encompasses intuition (Stake, 2004, 2013), perception and insight (Schwandt, 2008b), and practical and experiential knowledge (Stake & Schwandt, 2006). These authors encourage evaluators " . . . not to by-pass this kind of knowledge as a source for understanding quality, but to describe and respect it for its discriminative and operational power" (ibid, p.409).

Stake and Schwandt get to the core of what it means to reason evaluatively in a professional capacity - as evaluation practitioners how do we discern or know about an evaluand? Evaluation practitioners will relate to the notion of data not feeling right, or of having a gut feeling that something is happening that is not immediately evident. But the knowing to which Stake and Schwandt refer is something more substantive than is illustrated in these examples, particularly given Stake's (2013) claim that " . . . evaluating is partly an intuitive act" (p.108).

The intuition, perception and insight to which Stake and Schwandt refer are aligned to Kahneman's (2011) construct of "expert intuition" (p.11). Kahneman, a cognitive psychologist, was awarded the Nobel Prize in 2002 for his work on behavioural economics. According to Kahneman, experts possess expert intuition defined as: "Valid intuitions (which) develop when experts have learned to recognise familiar elements in a new situation and to act in a manner that is appropriate to it" (p.12). Expert intuition takes a long time and a great deal of practice to develop: "The acquisition of skill in complex tasks such as high-level chess . . . or firefighting is intricate and slow because expertise in a domain is not a single skill but a collection of skills" (p.238).

Expert intuition is also implicit in Eisner's (2004) view of evaluation as connoisseurship and expert criticism, described in chapter 5. According to Eisner (2004), the connoisseur or expert critic " . . . has learned what to look for, and can recognise quality when they see it. In addition, they can give reasons for their judgment . . . they can notice" (p.197). This "noticing ability" (p.198) can be thought of as expert intuition.

The claims made by Stake (2004, 2013) and Stake and Schwandt (2006) about discernment in evaluation being (in part) an intuitive act may be regarded as controversial to many in the evaluation profession. Their claims could be regarded as challenging the prescriptive and explicit nature of Scriven's logic of evaluation (1980a). I contend that the approach proposed by Stake and Schwandt to professional evaluative knowing should not be dismissed outright. The role of expert intuition in the discerning of evaluation practitioners is worthy of further examination (although outside the scope of this study). A range of questions arise about how we discern as evaluation professionals in relation to an evaluand. For example, is our professional discerning a complex fusion of explicit knowledge and expert intuition? Are we cognisant of whether, or how, expert intuition may influence our professional discerning? As evaluators, in what situations and under what circumstances should we actively exercise our expert intuition as part of our professional discerning, and when should we disregard it? At what point (if any) in our evaluation careers can we trust what we (think we) know instinctively about a particular evaluand?

In a recent publication Julnes and Bustelo (2016) also promote the acceptance of diverse ways of evaluative knowing. They describe Stake's approach as providing "holistic valuing" (p. 102), and Scriven's approach as "analytic valuing" (ibid). These authors suggest that rather than viewing the two approaches as contradictory and conflicting, they can be regarded as being complementary in that they offer opportunities to deepen and extend our evaluative knowing.

Lastly, if one subscribes to the notion of evaluation as craft described in the first conjecture, then the notion of the evaluator using expert intuition is unsurprising. Craftsmanship according to Sennett (2008) involves the development of skills whereby " . . . information and practices (are converted) into tacit knowledge" (p.50) which is the expert intuition described by Kahneman (2011).

9.4.6 Conjecture 3: Abductive inference offers a way of knowing for evaluative reasoning practice

The third conjecture asserts an important role for abduction in evaluative reasoning. This study has described abductive logic, explained its relationship to induction and deduction (Stephenson, 1961), and demonstrated its place in and contribution to the logic of systematic inquiry (Reichertz, 2014). Strategies to assess and strengthen abductive claims have also been presented (Tavory & Timmermans, 2014).

As described in conjectures one and two, Stake and Schwandt have focused attention on the nature of knowing that enables the evaluator to reason towards a robust evaluative conclusion/judgment. They encourage evaluators to use diverse ways of knowing, beyond those derived by deductive and inductive logics, variously referred to as practical knowledge (Schwandt, 2008b) and expert intuition (Kahneman, 2011; Stake, 2004). This opens the door for abductive thinking to become a legitimate and accepted aspect of evaluative reasoning practice. Abduction provides another means to “reason towards meaning” (Shank, 2008, p.2), offering opportunities for “ . . . imaginative . . . and intuitive interpretations” (Charmaz, 2008, p.158), and “creative meaning-making” (Tavory & Timmermans, 2014, p.121). Abductive inference enables evaluators to generate “ . . . alternative perspectives . . . (and) . . . feasible explanations for evaluative data” (DePoy & Gilson, 2008, p.28, 29).

The opportunity for generative thinking offered by abduction is significant for evaluation practitioners. Policy-makers are working in complex problem areas such as family violence and child poverty, sometimes referred to as “wicked problems” (Eppel, Turner & Wolf, 2011, p.193, 203). Experimentation and learning underpin policy design and implementation (Eppel, et al. 2011) as policy-makers seek new insights and knowledge about the problem and how it can be addressed. Abductive thinking opens up opportunities for new understandings to emerge about an evaluand and the problem it aims to solve, beyond those offered by inductive and deductive approaches.

The identification of conjectures one, two and three late in my research journey provided a new insight about Figure 3 (presented in chapter 5) which portrays evaluative reasoning as it is described in the literature. This insight is now described and the figure is modified to incorporate it (Figure 8).

The large box shape to the right of Figure 8 can be thought of as *the meaning-making box.* This is where the evaluator makes meaning of the evidence against the criteria through analysis and argumentation. As described in conjectures two and three, this meaning-making may occur via diverse ways of knowing, which may include practical knowledge (Stake & Schwandt, 2006), expert intuition (Kahneman, 2011), probative (inductive) inference (House, 1977; Scriven, 1991) and abductive inference (DePoy & Gilson, 2008).

Returning to the premise of this study, it is important to emphasise that the evaluation practitioner's use of diverse ways of knowing such as expert intuition (conjecture 2) or abduction (conjecture 3) is not an excuse for evaluative reasoning that lacks robustness and transparency. Regardless of how the evaluator comes to their evaluative conclusion/judgment, the conclusion/judgment must be supported by a transparent chain of reasoning linking criteria (or other comparators) with evidence and claim, and include a warranted argument appropriate for the claim. Evaluators' discernment or knowing (whether it is based on explicit knowledge, or a fusion of explicit and other knowing) has to be translated into sound reasoning that is able to be scrutinised.

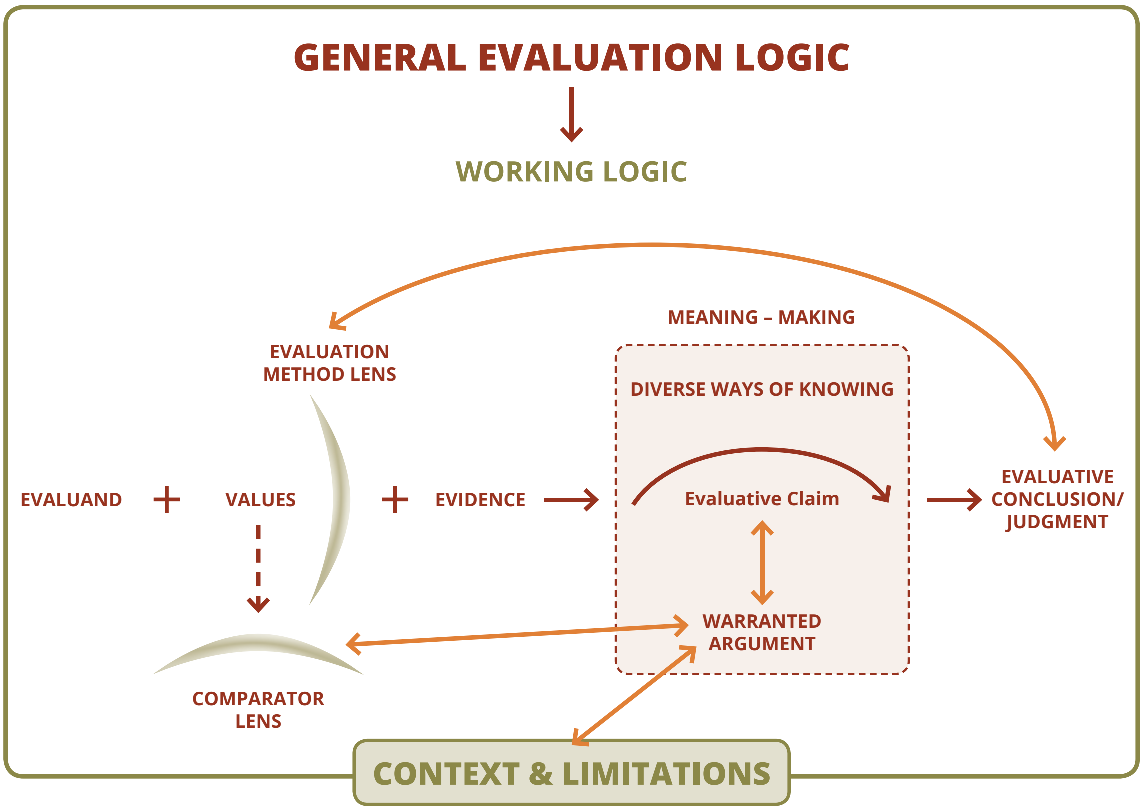


Figure 8 Evaluative reasoning

(Source: Author, 2016)

9.4.7 Conjecture 4: Visible and transparent evaluative reasoning is a fundamental aspect of the evaluator’s obligation to work for the public good

As noted in chapter 8, one of the six evaluation experts critiqued the use of meta-evaluation to examine evaluative reasoning practice. The expert posed two questions: Is evaluative reasoning visible or is it mostly inferred from other things? If evaluative reasoning is visible, where is it most visible? These questions require further consideration as they are highly relevant to the premise of this study, namely, that sound evaluative reasoning is an essential contributor to quality evaluation and is therefore the responsibility of professional evaluators. This fourth conjecture - that visible and transparent evaluative reasoning is part of the evaluator’s obligation to work for the public good - arises from the following response to the expert’s questions.

Reasoning evaluatively is a cognitive activity involving the intellectual tasks of comparison, critical analysis, interpretation, deliberation, and discernment. This study argues that the professional evaluator’s responsibility is to transform the cognitive activity that is evaluative reasoning into something that is comprehensible and transparent to those who have an interest in the evaluand. This transformation from cognitive activity into something that is both explicit and accessible to the evaluation audience is usually in the form of reporting, either written and/or verbal. Without this transformation from cognitive activity into a tangible form, the professional evaluator is simply issuing evaluative conclusions/judgments that are unable to be scrutinised and are therefore potentially unwarranted. Such transparency is emphasised in three of the Program Evaluation Standards (Joint Committee, 2011), namely, the need for explicitness and clarity of evaluative reasoning (Accuracy standards A1 and A7, Utility standard U4) as detailed in chapter 5.

The need for explicit and transparent evaluative reasoning is more pronounced in the public sector context. Writing about the role of evaluation in contemporary society, Dahler-Larsen (2012) notes “Evaluation and the modern idea of democracy are closely linked” (p.9). Evaluation undertaken by public administrations in democratic societies is concerned with public interest issues (Chelimsky, 2014) as highlighted in the AEA *Guiding Principles for Evaluators:* “Evaluators have obligations that encompass the public interest and good” (2014, n.p.). This principle belies the ambiguity surrounding the notion of public interest due to the plurality of values underpinning democratic societies (Chelimsky, 2014). While not underestimating the impact of such ambiguity on public sector evaluation practice, there is one aspect of public interest that is less fraught for evaluators to attain, namely, producing evaluations of public policies and initiatives that are transparent and independent. Such evaluations serve the public interest, whether the evaluation findings are in the public domain or the evaluation report remains in the policy-makers' files, as is sometimes the case. Such transparency encompasses evaluative reasoning, specifically, the evaluator is explicit about the values on which their evaluation is based, establishes clear links between evidence, inference and claim, and provides evaluative conclusions/judgments that are able to be "subject to rational analysis" (House, 2004b, p.8). Further, it is argued that evaluative reasoning as articulated in Scriven's logic of evaluation and explicated by other theorists described in chapter 6 (and summarised in the conceptual framework in chapter 5) provides a framework of the elements and process of reasoning that will facilitate such transparency. Specifically, it provides a means whereby the source and nature of criteria and standards on which an evaluation is based are explicit, encourages considered deliberation in applying criteria and standards to evidence, as well as the setting out of evaluative claims and arguing their relevance.

Therefore responding to the expert's questions, reasoning evaluatively is essentially a cognitive act which needs to be made visible and transparent in evaluation reporting (written or verbal). Such visibility and transparency is required in order that the outputs of reasoning evaluatively - evaluative claims - can be scrutinised for their robustness and relevance. This is not an option for the public sector evaluator, but is part of their obligations to work in the public interest and for the public good. Finally, evaluative reasoning is visible in other ways, for example, in the way an evaluator describes their approach to undertaking evaluation. However, these are secondary to the evaluator's responsibility to provide evaluative reasoning that is visible and transparent to evaluation audiences.

9.5 Coming full circle

. . . the end of all our exploring
will be to arrive where we started
and know the place for the first time.

Little Gidden, Four Quartets, T. S. Elliot, 1942.

As has been noted, this research started with a hunch that evaluative reasoning contributes to evaluation quality. Exploration of this hunch led me down various paths - learning about the logic-based roots of evaluative reasoning, understanding differing conceptualisations of how evaluative reasoning should occur, thinking about how discernment occurs in an evaluative context, and considering dissimilar evaluation imaginaries for Aotearoa New Zealand in general and its public sector in particular. It is therefore appropriate to come full circle by explicating the connections between evaluative reasoning, evaluation quality, and the notion of evaluation imaginaries.

Professional evaluators are concerned to ensure the quality of the evaluations they undertake. They focus, among other things, on the suitability of the evaluation design and sample, the choice of an appropriate evaluation method and its correct application, and the quality of evidence collected. This thesis has argued that sound evaluative reasoning is an essential contributor to evaluation quality. Consequently, other dimensions of evaluation quality such as design, method and evidence are necessary but not sufficient. Weak evaluative reasoning will compromise any other attempts to create a quality evaluation. As has been demonstrated in this study, evaluative reasoning provides the means, or in the words of Scriven (2012b) “the logical infrastructure” (p.18), by which one can reason from an evaluative claim about an evaluand to an evaluative conclusion/judgment in a manner that demonstrates a systematic “train of reasoning” (Toulmin et al., 1979, p.13). Put simply, evaluative reasoning constitutes the essence of what we do as evaluation professionals that differentiates our work from research and other forms of systematic enquiry. Specifically, sound evaluative reasoning facilitates the following which contribute to evaluation quality.

- Transparency about our positionality in relation to a particular evaluand, the evaluation participants and stakeholders.

- Explicitness about the values that underpin the assessment of an evaluand and the consequent evaluative conclusions/judgments about that evaluand.
- A transparent chain of reasoning from values, evidence and claim to evaluative conclusion/judgment.
- Probative inferences that are relevant and plausible for the evaluation audience.
- Claims that are linked to evaluative conclusions/judgments by warranted argument.
- An evaluative conclusion/judgment that will withstand scrutiny.

My attention now turns to the connection between the notion of evaluation imaginaries, and evaluative reasoning and evaluation quality. The study has shown that a particular evaluation imaginary may influence how evaluative reasoning is conceptualised and practised, and how evaluation quality is understood. For example, an evaluation imaginary in which equity values are prominent may lead to evaluative reasoning practice where the evaluator role is as facilitator and there is a high level of involvement of evaluation stakeholders and participants, for example, in determining criteria and in shaping evaluative conclusions/judgments. A criterion for evaluation quality according to this evaluation imaginary may be the extent of involvement of stakeholders and participants in the evaluation process. In contrast, an evaluation imaginary in which scientific values are prominent may lead to evaluative reasoning practice where the evaluator role is as an expert, remaining detached from evaluation stakeholders and participants. A criterion for evaluation quality according to this evaluation imaginary may be the extent to which the evaluative conclusion/ judgment is deemed to be independent and objective.

9.6 Conclusion

Chapter 9 has brought together the findings of the Q study, meta-evaluation and expert interviews to describe how evaluative reasoning is understood and practised by some evaluators working in or for the Aotearoa New Zealand public sector, and the contextual factors that have influenced such understanding and practice. Drawing on the concept of the evaluation imaginary proposed by Dahler-Larsen (2012) and Schwandt (2009b), the study findings have led to four evaluation imaginaries being proposed for Aotearoa New Zealand. The chapter proposes that specific evaluation imaginaries may influence how evaluative reasoning is conceptualised and practised, and evaluation quality is understood.

CHAPTER 10

CONCLUSION

10.1 Introduction

This study has argued that sound evaluative reasoning is an essential element of evaluation quality. As such, evaluative reasoning is a lens through which to consider how to improve the quality of evaluations undertaken or commissioned by the Aotearoa New Zealand public sector. My argument is grounded in the theory of evaluation derived from western philosophy, specifically, informal logic. This theory forms the conceptualisation and design of this multiple method inquiry into how evaluative reasoning is understood and practised by professionals who undertake public sector evaluation in Aotearoa New Zealand.

The study findings suggest that the understanding and practice of evaluative reasoning by professionals undertaking public sector evaluation is variable - from that which reflects evaluative reasoning theory, to alternative theorising which is not aligned to evaluative reasoning theory. It is suggested that a potential explanation for such variability can be derived from understanding whether the person undertaking the evaluation identifies as a professional evaluator and belongs to a formal or informal evaluation network. Professionals who do not identify as an evaluator and do not belong to an evaluation network may be less likely to have been exposed to evaluative reasoning theory and understand its implications for evaluation practice. While this is a plausible hypothesis or proposition, it needs to be tested more systematically in further research.

Further, the study identifies inductively-derived features about the way in which evaluation and consequently evaluative reasoning is conceptualised and practised by some evaluation practitioners in Aotearoa New Zealand, namely, as being values-based, practice-orientated, judgment-focussed and relational. These features are attributed to history, cultural norms, and the influence of Māori epistemology.

Turning attention to how evaluative reasoning practice can be strengthened, the study offers some abductively-derived conjectures. Having demonstrated that there are conceptual and practical reasons for professionals undertaking evaluation to have in-depth understanding of evaluative reasoning theory and its application, the study also argues that there is an ethical dimension associated with evaluative reasoning in the public sector context. Visible and transparent evaluative reasoning is an ethical imperative for evaluators in respect of their obligations to work in the public interest and for the public good (AEA Guiding Principles for Evaluators, 2014).

10.2 Contribution to knowledge

10.2.1 Evaluative reasoning theory

This study provides a modest contribution to the theory of evaluative reasoning. It anchors Scriven's (1967, 1980a, 1993, 1995) logic of evaluation into informal logic, thereby justifying evaluation in terms of western philosophic theory. I have been unable to locate any other evaluation literature which describes this link between informal logic and evaluation. The study has also attempted to provide a theoretical overview of all of the elements that constitute evaluative reasoning from beginning (understanding and defining value) to end (the evaluative conclusion/judgment). Theorists have tended to focus on specific elements of evaluative reasoning, for example, Davidson (2005) has focussed on evaluative rubrics, Eisner (2004) on valuing, House (1977, 1980, 1995) on valuing, deliberation and argumentation, Schwandt on valuing and ethics (2002b, 2008b), and Stake on valuing (1997, 2004, 2013). The study has also attempted to summarise the differing theoretical perspectives about evaluative reasoning, particularly in relation to how values are understood and discerned, as evidenced by the ongoing debate between Scriven and Stake (2013).

10.2.2 Evaluative reasoning practice

As noted in chapter five, the literature about evaluative reasoning is much smaller than the literature on other evaluation topics such as methods and use. Examining the evaluative reasoning literature, a reader may be struck by the number of books and papers explicating how evaluative reasoning should be done. Authors have paid considerably less attention to how evaluative reasoning is practised by evaluators working in real-life situations with a particular evaluand, and faced with the pragmatics of politics, conflicting stakeholder perspectives, budget and other constraints. Such is the paucity of practice-based literature about evaluative reasoning that Patton (2012) has referred to “the black hole of valuing” (p.97). The notable exceptions are Arens’ doctoral thesis about evaluative reasoning as demonstrated in five evaluation studies judged outstanding by the AEA (2005), the examination by Hurteau, Lachapelle and Houle (2006) of evaluative reasoning elements in papers about evaluations published in evaluation journals, and a meta-analysis of evaluation reports by Hurteau, Houle and Mongiat (2009). This study adds to this body of empirical research. Drawing on the concept of the evaluation imaginary (Dahler-Larsen, 2012; Schwandt, 2009b), the study has demonstrated how evaluative reasoning is understood and practised according to the values and norms underpinning a particular evaluation imaginary (discussed further in section 10.3.4).

The study has portrayed evaluative reasoning as an essential aspect of the craft that is evaluation, dismissing instrumental notions of evaluation as a technical activity and evaluative reasoning as something that can be easily learnt by attending a training workshop. Despite this stance, it is proposed that the heuristic used in the meta-evaluation (refer Figure 6, chapter 7) is a high-level summary of evaluative reasoning theory, thereby providing an accessible conceptual framework or visual checklist for novice evaluation practitioners. The heuristic may make it easier for novice evaluators to keep the principles of evaluative reasoning in mind while they go about their work. This suggested use of the heuristic comes with the disclaimer that the quality, rather than the procedure of evaluative reasoning is important.

10.2.3 Evaluation practice in Aotearoa New Zealand

The study has particular relevance for audiences in Aotearoa New Zealand because it adds to the growing knowledge about local evaluation practice, as evidenced in the publication of the first edition of *Evaluation Matters - He Take Tō Te Aromatawai* (2015), and doctoral theses on evaluation topics such as process use (Blewden, 2014) and evaluation influence in the health sector (Appleton-Dyer, 2012). Local authors Davidson (2005) and J. King, McKegg, Oakden & Wehipeihana (2013) have written about evaluative reasoning, focusing on the how to aspect as evidenced in the subtitle of Davidson’s book *The nuts and bolts of sound evaluation.* Similarly,

J. King et al., provide guidance on the use of evaluative rubrics. This study has a different focus, emphasising the quality of evaluative reasoning and its theoretical underpinnings. If evaluation commissioners and practitioners were to have a more in-depth theoretical understanding, this could lead to more critical attention on evaluative reasoning practice which might then improve the quality of public sector evaluation over time.

10.3 Implications for theory and practice: further research

10.3.1 Fundamental issues in evaluation

The International Year of Evaluation was celebrated in 2015. It was facilitated by EvalPartners, a group consisting of sixty-three national evaluation bodies, evaluation organisations and other parties. The purpose of this year-long event was to " . . . bring together diverse stakeholders into a movement designed to mobilise the energies and enhance the synergy of existing and new monitoring and evaluation initiatives at international and national levels" (EvalPartners, n.d). The range of activities held around the world as recorded on the EvalPartners' website and the resources produced suggest that evaluation is currently undergoing a period of revitalisation as noted by Picciotto (22 April, 2016) on an EVALKTALK posting. Despite such optimism, a number of fundamental issues (N. L. Smith, 2009) face the evaluation profession, defined as "those essential, underlying concerns that shape the future and nature of the evaluation enterprise" (p.48). According to N. L. Smith (2009), one fundamental issue concerns the validity of evaluators' understanding of quality: "How do we arrive at the most valid understandings of quality? Controlled experiments? Moral deliberation? Phenomenological renderings?" (ibid). In his paper Smith (2009) appears to conceptualise the discernment of quality as being related solely to method. This study has set out to demonstrate that, within the context of a western epistemology, sound evaluative reasoning is an essential aspect of evaluation. I have argued that regardless of the evaluator's efforts to select and apply the most appropriate method for the evaluation purpose or complexity of valuing required, such method-focused efforts are compromised without sound evaluative reasoning. Returning to N. L. Smith's (2009) question above, yes, valid understandings of quality require appropriate methods, but together with sound evaluative reasoning.

10.3.2 Further research

The gulf between evaluation theorists and their theories on the one hand, and practitioners and their practice on the other has been identified by Chelimsky (2013). This is despite their "interdependent" (p.91) relationship. Reflecting on Chelimsky's observations, Rog (2015) emphasises the importance of "infusing our evaluation practice with different types of theory . . . and infusing practice into theory (p.224). Given the less developed state of the literature about the practice of evaluative reasoning, practice-based research about evaluative reasoning would foster both theory-informed practice and practice-informed theory.

My study findings have identified a number of potential topics for further research about evaluative reasoning practice in general, and specifically about evaluation reasoning practice in the public sector context, as follows.

i. Citizens in western democracies are seeking greater participation in public policy decision-making resulting in new terms of engagement between the state and the public (Ryan, 2011). Ryan notes that this participation is significantly more substantive than the type of consultation commonly used by governments' to involve citizens in public affairs. What are the implications of participatory policy making for the conceptualisation, design, conduct and reporting of public sector evaluations? Will such participation require the development of new approaches to valuing?

ii. What can the evaluation profession learn about discernment and judgment-making from cognitive psychology, legal reasoning and other disciplines?

iii. Julnes and Bustelo (2016) promote the " . . . balancing and mixing of (Scriven's) analytic and (Stake's) holistic valuing" (p.102). How could this be done in practice? What would this look like?

iv. How is expert intuition (Kahneman, 2011) able to be fostered in evaluation practitioners?

v. How do evaluators approach and manage evaluations where there are competing value positions? Is it theoretically possible and/or practically feasible for the evaluator to suggest alternative evaluative conclusions/judgments argued from different value positions?

vi. How do internal evaluators manage potential tensions between political values and those of stakeholders in respect of a particular evaluand?

vii. Evaluation rubrics are promoted in the literature as the best way to articulate criteria in respect of a specific evaluand (Davidson, 2005). As noted in chapters 5 and 7 theorists such as Stake and Schwandt (2006) criticise the use of criteria and criterial thinking because of their focus on certain characteristics of an evaluand thereby excluding a wider knowledge of it. In light of these concerns, what other approaches are possible for articulating criteria?

viii. If evaluation is a craft and evaluative reasoning an essential element of the craft, what are the implications for the way in which evaluation is taught in academic courses and professional development learning situations? What value does on-the-job, apprentice-type learning have?

10.3.3 Professionalisation of evaluation

A second fundamental issue identified by N. L. Smith (2009) concerns how the evaluation profession can ensure the quality of evaluation practice. N. L. Smith offers a number of suggestions, namely " . . . accreditation and licensing? Consensual professional standards? Mandatory meta-evaluation?" (p.48). Such mechanisms are indicators of the professionalisation of evaluation, a topic that is being discussed in Aotearoa New Zealand (McKegg, 2014) and internationally (Altschuld & Engle, 2015). Given the centrality of evaluative reasoning to evaluation, improving the quality of evaluative reasoning practice is an important aspect of any efforts aimed at building professionalism. This study has indicated that relatively little is known about who is doing evaluation in Aotearoa New Zealand and the public sector in particular. The proposition identified by this study that there may be professionals in Aotearoa New Zealand working outside of the umbrella of formal or informal evaluation networks is a consideration for any professionalisation efforts by ANZEA and AES. Further research examining the primary professional identifications of those who undertake public sector evaluation and how this influences their evaluation practice would be both interesting and helpful in developing evaluation practice. Research could also investigate whether there are differences in practice between evaluators located in Wellington (where the majority of government agency head offices are located, and where most commissioning of public sector evaluation occurs) and those who work outside of Wellington, for example, in the not-for-profit sector or in metropolitan Auckland. Further, this study has assumed there are benefits for individuals from participating in a formal evaluation network such as ANZEA. A question arises whether the socialisation that may occur through such membership may foster conformity and discourage diversity of practice, providing an additional topic for further research.

10.3.4 Evaluation imaginaries for Aotearoa New Zealand

The study findings have suggested a number of dissimilar evaluation imaginaries for Aotearoa New Zealand and the public sector in particular, specifically, the values-based imaginary represented by ANZEA, the value-free imaginary as described by the Chief Science Adviser, and Māori and Pasifika imaginaries based on Māori and Pasifika epistemologies. These evaluation imaginaries are presented tentatively given they may not provide a complete picture of evaluation practice in Aotearoa New Zealand. For example, there may be an evaluation imaginary associated with the not-for-profit sector. Despite the tentative nature of the evaluation imaginaries presented here, they point to diversity in the way evaluation, and evaluative reasoning is understood and practised locally.

Lastly, the notion of the evaluation imaginary may be helpful for ANZEA's current consideration about the professionalisation of evaluation. Any professionalism efforts will involve debate and discussion about who we are as evaluation practitioners and commissioners, and what we want to become in the future. For this reason, debate and further research is required to test the evaluation imaginaries presented here.

10.4 Concluding comment

Exciting things are currently happening in the evaluation space - technology is providing tools to make our work easier in the field, innovative methods are emerging, and new digital approaches are making findings more accessible to evaluation audiences. It is hoped that in the midst of such stimulating times, we do not become distracted from evaluative reasoning which is central to our craft as evaluation professionals and an essential component of evaluation quality.

REFERENCES

Abma, T. A. (2006). The social relations of evaluation. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation* (pp. 185-200). London: Sage Publications.

Abma, T. A., & Widdershoven, G. A. M. (2008). Evaluation and/as social relation. *Evaluation, 14*(2), 209-225.

Abma, T. A., & Widdershoven, G. A. M. (2011). Evaluation as a relationally responsible practice. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage book of qualitative research* (pp. 669-680). Thousand Oaks, CA: Sage Publications.

Alkin, M. C., & Christie, C. A. (2004). An evaluation theory tree. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences*. (pp. 12-65). Thousand Oaks, CA: Sage Publications.

Alkin, M. C., Vo, A. T., & Christie, C. A. (2012). The evaluator's role in valuing: Who and with whom? *New Directions for Evaluation, 133*, 29-41.

Altschuld, J. W., & Engle, M. (2015). Accreditation, certification and credentialing: Relevant concerns for U.S. evaluators. *New Directions for Evaluation, 145* (Spring), 5-19.

Alvesson, M., & Skoldberg, K. (2009). *Reflexive methodology: New vistas for qualitative research* (2nd ed.). London: Sage Publications.

American Evaluation Association (2014). *Guiding principles for evaluators.* Retrieved from http://www.eval.org/p/cm/ld/fid=51

Aotearoa New Zealand Evaluation Association. (n.d.). *About us.* Retrieved from http://www.anzea.org.nz/about-us/

Aotearoa New Zealand Evaluation Association. (2011). *Evaluator competencies*. Retrieved from http://www.anzea.org.nz/

Appleton-Dyer, S. (2012). Understanding the mechanisms and outcomes of evaluation influence within population health partnerships. (Doctoral dissertation, The University of Auckland, Auckland, New Zealand). Retrieved from https://researchspace.auckland.ac.nz/handle/2292/19851

Arens, S. A. (2005). A study of evaluative reasoning in evaluation studies judged "outstanding" (Unpublished doctoral thesis), Indiana University, Bloomington, United States.

Argyris, C. (1976). Theories of action that inhibit individual learning. *American Psychologist, 31*(9), 638-654.

Australasian Evaluation Society. (2013). *Guidelines for the ethical conduct of evaluations.* Australasian Evaluation Society. Retrieved from http://www.aes.asn.au/images/stories/files/About/Documents%20-%20ongoing/AES%20Guidlines10.pdf

Bahler, K. (2003). Evaluation and the policy cycle. In N. Lunt, C. Davidson, & K. McKegg (Eds.), *Evaluating policy and practice: A New Zealand reader* (pp. 27-39). Auckland: Pearson Education.

Bamberger, M., Rugh, J., & Mabry, L. (2006). *Real world evaluation: Working under budget, time, data and political constraints.* Thousand Oaks, CA: Sage Publications.

Barrett, M., & Connolly-Stone, K. (1998). The Treaty of Waitangi and social policy. *Social Policy Journal of New Zealand, 11*. Retrieved from http://www.msd.govt.nz/about-msd-and-our-work/publications-resources/journals-and-magazines/social-policy-journal/spj11/treaty-of-waitangi-and-social-policy.html

Blair, J. A. (1995). Informal logic and reasoning in evaluation. *New Directions for Evaluation, 68*, 71-80.

Blewden, M. B. (2014). Why do evaluators intentionally seek process use? Exploring meaning and reason as explanation. (Doctoral dissertation, Massey University Wellington, New Zealand). Retrieved from http://mro.massey.ac.nz/handle/10179/5199

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2006). *Cost-benefit analysis: Concepts and practice* (3rd ed.). New Jersey: Pearson Education.

Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed.). Chicago: The University of Chicago Press.

Boston, J., Martin, J., Pallot, J., & Walsh, P. (1996). *Public management: The New Zealand model*. Auckland: Oxford University Press.

Brown, S. R. (1980). *Political subjectivity: Applications of Q methodology in political science*. New Haven: Yale University Press.

Brown, S. R. (1991). Q methodology. *Qualitative research for the human sciences*. Retrieved from https://facstaff.uww.edu/cottlec/QArchive/Primer1.hmtl

Bryman, A. (2008). *Social research methods* (3rd ed.). Oxford: Oxford University Press.

Byrnes, G. (2010). “Relic of 1840” or founding document? The Treaty, the Tribunal and concepts of time. *Kōtuitui: New Zealand Journal of Social Sciences Online, 1:1*, 1-12. Retrieved from http://www.royalsociety.org.nz/publications/journals/nzjs/ doi:10.1080/1177083X.2006.9522407

Campbell, D. T. (1982). Experiments as arguments. *Science Communication, 3*, 327-337.

Charmaz, C. (2008). Grounded theory as an emergent method. In S. Nagy Hesse-Biber & P. Leavy (Eds.), *Handbook of emergent methods* (pp. 155-170). New York: The Guildford Press.

Chelimsky, E. (1998). The role of experience in formulating theories in evaluation practice. *American Journal of Evaluation, 19*(1), 35-55.

Chelimsky, E. (2006). The purposes of evaluation in a democratic society. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.), *The Sage handbook of evaluation* (pp. 33-55). London: Sage Publications.

Chelimsky, E. (2012). Valuing, evaluation methods, and the politicization of the evaluation process. *New Directions for Evaluation, 133*, 77-83.

Chelimsky, E. (2013). Balancing evaluation theory and practice in the real world. *American Journal of Evaluation, 34*(1), 91-98.

Chelimsky, E. (2014). Public-interest values and program sustainability: Some implications for evaluation practice. *American Journal of Evaluation, 35*(4), 527-542.

Cook, T. D. (2006). Describing what is special about the role of experiments in contemporary educational research: Putting the "gold standard" rhetoric into perspective. *Journal of MultiDisciplinary Evaluation, 6*, 1-7. Retrieved from http://evaluation.wmich.edu/jmde/

Cooksy, I. J., & Caracelli, V. J. (2009). Meta-evaluation in practice: Selection and application of criteria. *Journal of MultiDisciplinary Evaluation, 6*(11), 1-15. Retrieved from http://evaluation.wmich.edu/jmde/

Cram, F. (1997). Developing partnerships in research: Pākehā researchers and Māori research. *SITES: Journal of social anthropology and cultural studies, 35*, 44-63.

Cram, F. (2001). Rangahau Māori: Tona tika, tona pono - The validity and integrity of Māori research. In M. Tolich (Ed.), *Research ethics in Aotearoa New Zealand* (pp. 35-52). Auckland: Pearson Education.

Cram, F. (2009). Maintaining indigenous voices. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics* (pp. 308-322). Thousand Oaks, CA: Sage Publications.

Cunningham, C. W., & Durie, M. H. (1998, July). *A taxonomy and a framework for outcomes and strategic research goals for Māori research and development*. Paper presented at Te Oru Rangahau: Māori Research and Development Conference, School of Māori Studies, Massey University.

Cunningham, R., Signal, L., & Bowers, S. (2010). *Evaluating Health Impact Assessments in New Zealand.* Retrieved from http://www.health.govt.nz/publication/evaluating-health-impact-assessments-new-zealand

Dahler-Larsen, P. (2012). *The evaluation society*. Stanford, CA: Stanford University Press.

Dahler-Larsen, P., & Schwandt, T. A. (2012). Political culture as context for evaluation. *New Directions for Evaluation, 135*, 75-87.

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

Davidson, E. J. (2006). The RCTs-only doctrine: Brakes in the acquisition of knowledge? *Journal of MultiDisciplinary Evaluation, 6*, ii-v. Retrieved from http://evaluation.wmich.edu/jmde/

Denzin, N. K. (1978). *The Research Act: A theoretical introduction to sociological methods* (2nd ed.). New York: McGraw-Hill.

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

Denzin, N. K., & Lincoln, Y. S. (2011b). The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (Vol. 4th, pp. 1-19). Thousand Oaks, CA: Sage Publications.

DePoy, E. & Gilson, S. (2008). *Evaluation practice: How to do good evaluation research in work settings*. New York: Routledge.

Dobson, P., Gengatharen, D., Fulford, R., Barratt-Pugh, L., Bahn S., & Larsen, A. (2012, December). *Eureka moments in research: Exploring abductive processes using four case examples.* Paper presented at the 23rd Australian Conference on Information Systems, Geelong, Australia.

Donaldson, S. I., & Christie, C. A. (2005). The 2004 Claremont debate: Lipsey vs. Scriven *Journal of MultiDisciplinary Evaluation, 3*, 60-77. Retrieved from http://evaluation.wmich.edu/jmde/

Eisner, E. W. (1996). Foreword. In L. Heshusius & K. Ballard (Eds.), *From positivism to interpretivism and beyond: Tales of transformation in educational and social research* (pp. ix-xi). New York: Teachers College Press.

Eisner, E. W. (2004). The roots of connoisseurship and criticism: A personal journey. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 196-202). Thousand Oaks, CA: Sage Publications.

Eppel, E., Turner, D., & Wolf, A. (2011). Complex policy implementation: The role of experimentation and learning. In B. Ryan & D. Gill (Eds.), *Future state: Directions for public management in New Zealand* (pp. 182-212). Wellington: Victoria University Press.

European Centre for Health Policy. (1999). Definitions from European Centre for Health Policy. Retrieved from http://apps.who.int/disasters/repo/ 13849_files/n/definitions_EURO_ECHP.pdf

EvalPartners (2015). *International Year of Evaluation*. Retrieved from http://www.evalpartners.org/evalyear/about

Fetterman, D. M. (2004). Branching to our roots for insight. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 304-318). Thousand Oaks, CA: Sage Publications.

Fitzpatrick, J. L. (2012). An introduction to context and its role in evaluation practice. *New Directions for Evaluation, 135*, 7-24.

Fotuali'i McGeady, D. (2015, July). *Effectively navigating for Pacific success: The impact of Va relationships on evaluation practice*. Paper presented at the Aotearoa New Zealand Evaluation Association Annual Conference Auckland, New Zealand.

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

Fournier, D. M. (2005). Logic of evaluation: Working logic. In S. Matheson (Ed.), *Encyclopaedia of evaluation* (pp. 238-242). Thousand Oaks, CA: Sage Publications.

Fournier, D. M., & Smith, N. L. (1993). Clarifying the merits of argument in evaluation practice. *Evaluation and Program Planning, 16*, 315-323.

Frondizi, R. (1971). *What is value? An introduction to axiology*. La Salle: Open Court Publishing Company.

Gluckman, P. (2013). *The role of evidence in policy formation and implementation.* Office of the Prime Minister's Science Advisory Committee. Retrieved from www.pmcsa.org.nz

Greene, J. C. (1990). Technical quality versus user responsiveness in evaluation practice. *Evaluation and Program Planning, 13*, 267-274.

Greene, J. C. (2005). Context. In S. Matheson (Ed.), *Encyclopaedia of evaluation* (pp. 82-84). Thousand Oaks, CA: Sage Publications.

Greene, J. C. (2011). The construct(ion) of validity as argument. *New Directions for Evaluation, 130*, 81-91.

Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. Newbury Park: Sage Publications.

Hare, R. M. (1967). What is a value judgment? In P. W. Taylor (Ed.), *Problems of moral philosophy - An introduction to ethics* (3rd ed., pp. 388-405). Belmont, CA: Wadsworth.

Hawke, G., Bedford, R., Kukutai, T., McKinnon, M., Olssen, E., & Spoonley, P. (2014). *Our Futures Te Pae Tāwhiti: The 2013 census and New Zealand's changing population*. Wellington: The Royal Society of New Zealand. Retrieved from http://www.royalsociety.org.nz/expert-advice/papers/yr2014/our-futures/

Hawkins, P. (2015, July). *Sailing into the winds of change to new evaluation zones.* Keynote address at the annual conference of the Aotearoa New Zealand Evaluation Association, Auckland, New Zealand. Retrieved from http://www.anzea.org.nz/anzea-conferences/anzea-conference-2015/conference-2015-presentations-archive/

Health Research Council. (2010). *Te Ara Tika guidelines for Māori research ethics*. Wellington: Author. Retrieved from http://www.hrc.govt.nz/sites/default/files/Te%20Ara%20Tika%20Guidelines%20for%20Maori%20Research%20Ethics.pdf

Heisenberg, W. (1990). *Physics and philosophy: The revolution in modern science*. London: Penguin.

Henry, E., & Pene, H. (2001). Kaupapa Māori: Locating indigenous ontology, epistemology and methodology in the Academy. *Organisation, 8*(2), 234-242.

Henry, G. T. (2002). Choosing criteria to judge program success: A values inquiry. *Evaluation, 8*(2), 182-204.

Heshusius, L., & Ballard, K. (1996). How do we count the ways we know? In L. Heshusius & K. Ballard (Eds.), *From positivism to interpretivism and beyond: Tales of transformation in educational and social research* (pp. 1-16). New York: Teachers College Press.

Hollinger, R. (1994). *Postmodernism and the social sciences: A thematic approach* (Vol. 4). Thousand Oaks, CA: Sage Publications.

Honderich, T. (Ed.). (2005). The Oxford companion to philosophy (2nd ed.). Oxford: Oxford University Press.

House, E. R. (1977). *The logic of evaluative argument.* Los Angeles: Centre for the Study of Evaluation, UCLA Graduate School of Education.

House, E. R. (1980). *Evaluating with validity*. Beverley Hills, CA: Sage Publications.

House, E. R. (1995). Putting things together coherently: Logic and justice. *New Directions for Evaluation, 68*, 33-48.

House, E. R. (1996). The problem of values in evaluation. *Evaluation Journal of Australasia, 8*(1), 3-14.

House, E. R. (2004a). Intellectual history in evaluation. In M. C. Alkin (Ed.), *The roots of fourth generation evaluation: Theoretical and methodological origins* (pp. 218-224). Thousand Oaks, CA: Sage Publications.

House, E. R. (2004b). The role of the evaluator in a political world. *Canadian Journal of Program Evaluation, 19*(2), 1-16.

House, E. R. (2014). Origins of the ideas in "Evaluating with Validity". *New Directions for Evaluation, 142*, 9-15.

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

Howe, K. R. (1992). Getting over the quantitative-qualitative debate. *American Journal of Education, 100*(2), 236-256.

Hughes, J. A., & Sharrock, W. W. (1990). *The philosophy of social science* (3rd ed.). London: Longman.

Hurteau, M., Houle, S., & Mongiat, S. (2009). How legitimate and justified are judgments in program evaluation? *Evaluation, 15*(3), 307-319.

Hurteau, M., Lachapelle, G., & Houle, S. (2006). Understanding evaluative practices in order to improve them: The specific modelling process to programme evaluation. *Measurement and Evaluation in Education, 29*(3), 27-44.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Josephson, J. R., & Josephson, S. G. (1999). *Abductive inference: Computation, philosophy, technology.* New York: Cambridge University Press.

Julnes, G. (2012a). Managing valuation. *New Directions for Evaluation, 133*, 3-15.

Julnes, G. (2012b). Developing policies to support valuing in the public interest. *New Directions for Evaluation, 133*, 109-129.

Julnes, G., & Bustelo, M. (2016). Providing appropriate evaluative support for valuing in the public interest. *American Journal of Evaluation, 37*(1), 100-103.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kallemeyn, L. M., Hall, J., Friche, N., & McReynolds, C. (2015). Cross-continental reflections on evaluation practice: Methods, use and valuing. *American Journal of Evaluation, 36*(3), 339-357.

Kampen, J. K., & Tamas, P. (2014). Overly ambitious: Contributions and current status of Q methodology. *Qual and quant, 48*, 3109-3126.

Kennedy, V., Cram, F., Paipa, K., Pipi, K., & Baker, M. (2015). Wairua and cultural values in evaluation. *Evaluation Matters He Take Tō Te Aromatawai, 1*(1), 83-105.

Kerr, S. (2012). Kaupapa Māori theory-based evaluation. *Evaluation Journal of Australasia, 12*(1), 6-25.

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of MultiDisciplinary Evaluation, 9*(21), 11-20.

King, J. A. (1998). Making sense of participatory evaluation practice. *New Directions for Evaluation, 80*, 57-67.

King, M. (1999). *Being Pākehā now*. Auckland: Penguin Books.

King, M. (2003). *The Penguin history of New Zealand.* Auckland: Penguin Books.

Kirkhart, K. E. (2010). Eyes on the prize: Multicultural validity and evaluation theory. *American Journal of Evaluation, 31*(3), 400-413.

Klemke, E. D., Hollinger, R., & Kline, A. D. (Eds.). (1980). *Introductory readings in the philosophy of science*. New York: Prometheus Books.

Kolakowski, L. (1966). *Positivist philosophy: From Hume to the Vienna Circle*. London: Pelican Books.

Krathwohl, D. R. (1980). The myth of value-free evaluation. *Educational Evaluation and Policy Analysis, 2*(1), 37-45.

LaFrance, J., Nichols, R., & Kirkhart, K. E. (2012). Culture writes the script: On the centrality of context in indigenous evaluation. *New Directions for Evaluation, 135*, 59-74.

Leeuw, F. L. (2003). Reconstructing program theories: Methods available and problems to be solved. *American Journal of Evaluation, 24*(5), 5-20.

Leeuw, F. L. (2008, September). *Evaluating interventions and underlying behavioural mechanisms.* Paper presented at the annual conference of the Australasian Evaluation Society, Perth, Australia.

Letherby, G., Scott, J., & Williams, M. (2013). *Objectivity and subjectivity in social research.* Los Angeles: Sage Publications.

Lowe, E. J. (2005). Ontology. In the *Oxford companion to philosophy* (2nd ed., p.670): Oxford: Oxford University Press.

Lunt, N. (2003). Knowledge for policy: The emergence of evaluation research within New Zealand. In N. Lunt, C. Davidson, & K. McKegg (Eds.), *Evaluating policy and practice: A New Zealand reader* (pp. 4-15). Auckland: Pearson Education.

Lunt, N., Davidson, C., & McKegg, K. (Eds.). (2003). *Evaluating policy and practice: A New Zealand reader*. Auckland: Pearson Education.

Lunt, N., & Trotman, I. G. (2005). A stagecraft of New Zealand evaluation. *Evaluation Journal of Australasia, 5*(1), 3-10.

MacDonald, B., & Kushner, S. (2005). Democratic evaluation. In S. Mathison (Ed.), *Encyclopaedia of evaluation* (pp. 109-113). Thousand Oaks, CA: Sage Publications.

Magendanz, D. (2003). Conflict and complexity in value theory. *The Journal of Value Inquiry, 37*, 443-453.

Mathison, S. (2005). *Encyclopaedia of evaluation*. Thousand Oaks, CA: Sage Publications.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62*(3), 279-300.

McKegg, K. (2014, September). *Evaluation today: International and local perspectives*. Paper presented at a workshop of the Aotearoa New Zealand Evaluation Association, Wellington, New Zealand.

McKeown, B., & Thomas, D. (2013). *Q methodology* (2nd ed. Vol. 66). Newbury Park: Sage Publications.

Mertens, D. M. (2007). Transformative paradigm: Mixed methods and social justice. *Journal of Mixed Methods Research, 1*(3), 212-225.

Mikaere, M. (2013). *Māori in Aotearoa New Zealand: Understanding the culture, protocols and customs*. Auckland: New Holland Publishers.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ministry of Business, Innovation and Employment (2016). All of Government Consultancy Services. Wellington: Author. Retrieved from https://www.gets.govt.nz/MBIE/ExternalTenderDetails.htm?id=5759284

Ministry of Māori Development. (1999). *Guidelines for evaluation with Māori.* Wellington: Author.

Ministry of Social Development. (2004). *Nga Ara Tohutohu Rangahau Māori: Guidelines for Research and Evaluation with Māori*. Wellington: Author.

Mirza, N. A., Akhtra-Danesh, N., Noesgaard, C., Martin, L., & Staples, E. (2014). A concept analysis of abductive reasoning. *Journal of Advanced Nursing, 70*(9), 1980-1994.

Moewaka Barnes, H. (2003). Māori and evaluation: Some issues to consider. In N. Lunt, C. Davidson, & K. McKegg (Eds.), *Evaluating policy and practice: A New Zealand reader* (pp. 146-150). Auckland: Pearson Education.

Moewaka Barnes, H. (2009). *The evaluation hīkoi: A Māori overview of programme evaluation.* Retrieved from http://www.shore.ac.nz/massey/fms/Colleges/College%20of%20Humanities%20and%20Social%20Sciences/Shore/reports/HMB_Māori-Evaluation-Manual-2009.pdf?EDB8EFDD55E24388A0D3EFBB42B89D2A

Moore, A. D. (2004). Values, objectivity and relationalism. *The Journal of Value Inquiry (38)*, 74-90.

Moorfield, J. C. (2005). *Te Whanake Te Aka Māori-English, English-Māori Dictionary and Index*. Auckland: Pearson Longman.

Morrison, A. (May 2014). Picking up the pace in the public services. *Policy Quarterly, 10*(2), 43-48.

Morse, J. M. (2010). Procedure and practices of mixed method design: Maintaining control, rigor and complexity. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social and behavioural research* (2nd ed., pp. 339-378). Thousand Oaks, CA: Sage Publications.

Newman, I., & Ramlo, S. E. (2010). Using Q methodology and Q factor analysis in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in mixed methods research* (2nd ed., pp. 505-530). Thousand Oaks, CA: Sage Publications.

Norris, N. (2015). Democratic evaluation: The work and ideas of Barry MacDonald. *Evaluation 21*(2), 135-142.

Nutley, S., Davies, H., & Walter, I. (2003). Evidence-based policy and practice: Lessons from the United Kingdom. *Social Policy Journal of New Zealand, 20*, 29-48.

Office of Treaty Settlements. (2015). *Healing the past, building a future: A Guide to Treaty of Waitangi claims and negotiations with the Crown*. Wellington: Author.

O'Hear, A. (1989). *An introduction to the philosophy of science*. Oxford: Clarendon Press.

Oldroyd, D. (1986). *The arch of knowledge: An introductory study of the history of the philosophy and methodology of science.* New York: Methuen.

Parliamentary Counsel Office (n.d.). The principles of the Treaty as expressed by the Courts and the Waitangi Tribunal. Retrieved from http://www.waitangitribunal.govt.nz/assets/Documents/Publications/WT-Principles-of-the-Treaty-of-Waitangi-as-expressed-by-the-Courts-and-the-Waitangi-Tribunal.pdf

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (2002a). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (2002b). A vision of evaluation that strengthens democracy. *Evaluation, 8*(1), 125-139.

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York: Guildford Press.

Patton, M. Q. (2012). Contextual pragmatics of valuing. *New Directions for Evaluation,133,* 97-108.

Picciotto, R. (April 22, 2016). Revitalisation of evaluation [Online forum comment]. Retrieved from American Evaluation Association Discussion List EVALTALK@ LISTSERV.UA.EDU

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.

Ramlo, S. E. (2016). Mixed method lessons learned from 80 years of Q methodology. *Journal of Mixed Methods Research, 10*(1), 28-45.

Ramlo, S. E., & Newman, I. (2010). Classifying individuals using Q methodology and Q factor analysis: Applications of two mixed methodologies for program evaluation. *Journal of Research in Education, 20*(2), 20-31.

Ramlo, S. E., & Newman, I. (2011). Q methodology and its position in the mixed-methods continuum. *Operant Subjectivity: The International Journal of Q Methodology, 34*(3), 172-191.

Reichardt, C. S., & Rallis, S. F. (1994). The qualitative-quantitative debate: New perspectives. *New Directions for Program Evaluation, 61,* 1-3.

Reichertz, J. (2014). Induction, deduction and abduction. In U. Flick (Ed.), *The Sage handbook of qualitative data analysis* (pp. 123-135). London: Sage Publications.

Rescher, N. (1969). *Introduction to value theory*. Englewood Cliffs: Prentice-Hall.

Rog, D. J. (2015). Infusing theory and practice, practice into theory: Small wins and big gains for evaluation. *American Journal of Evaluation, 36*(2), 223-238.

Rogers, P. J., & Davidson, E. J. (2013). Australian and New Zealand evaluation theorists. In M. C. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 371-385). Thousand Oaks, CA: Sage Publications.

Ryan, B. (2003). Death by evaluation: Reflections on monitoring and evaluation in Australia and New Zealand. *Evaluation Journal of Australasia, 3*(1), 6-16.

Ryan, B. (2011). The signs are everywhere: 'Community' approaches to public management. In B. Ryan & D. Gill (Eds.), *Future state: Directions for public management in New Zealand* (pp. 85-122). Wellington: Victoria University Press.

Ryan, B., & Gill, D. (2011). Past, present and the promise: Rekindling the spirit of reform. In B. Ryan & D. Gill (Eds.), *Future state: Directions for public management in New Zealand* (pp. 306-318). Wellington: Victoria University Press.

Salmond, A. (2008). *Two worlds: Tangled histories*. Waitangi Rua Rautau Lecture. Retrieved from http://www.radionz.co.nz/national/programmes/ waitangiruarautaulectures/

Salmond, A. (2012). *Beyond the binary - Shifting New Zealand's mindset*. Paper presented at the First Sir Paul Reeves Memorial Lecture. Retrieved from http://www.radionz.co.nz/national/programmes/reeves/audio/2535591/2012-beyond-the-binary-shifting-new-zealand's-mindset

Schwandt, T. A. (1997). The landscape of values in evaluation: Charted terrain and unexplored territory. *New Directions for Evaluation, 76*(Winter), 25-38.

Schwandt, T. A. (2001a). Responsiveness and everyday life. *New Directions for Evaluation, 92*, 73-86.

Schwandt, T. A. (2001b). A postscript on thinking about dialogue. *Evaluation, 7*, 264-276.

Schwandt, T. A. (Ed.). (2002a). *Traversing the terrain of role, identity and self.* Charlotte, NC: Information Age Publishing.

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

Schwandt, T. A. (2007a). *The Sage dictionary of qualitative inquiry* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Schwandt, T. A. (2007b). Judging interpretations. *New Directions for Evaluation, 114*, 11-14.

Schwandt, T. A. (2008a). Educating for intelligent belief in evaluation. *American Journal of Education, 29*, 138-150.

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

Schwandt, T.A. (2009a). Globalizing influence on the western imaginary. In K. E. Ryan & B. Cousins (Eds.), *The Sage International Handbook of Educational Evaluation* (pp.19-36. Thousand Oaks, CA: Sage Publications.

Schwandt, T. A. (2009b). Toward a practical theory of evidence for evaluation. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 197-211). Thousand Oaks, CA: Sage Publications.

Schwandt, T. A. (2010, September). *Evaluative reasoning, evidence and use*. Paper presented at the annual conference of the Australasian Evaluation Society, Wellington, New Zealand.

Scobie, G. (2009, August). *Evidence-based policy: Reflections from New Zealand*. Paper presented at the Strengthening Evidence-based Policy in the Australian Federation Conference, Canberra, Australia. Retrieved from http://www.pc.gov.au/research/completed/strengthening-evidence

Scott, J. (2015). Practical knowledge. In Online Oxford Dictionary of Sociology. Oxford University Press. Retrieved from http://www.oxfordreference.com/view/10.1093/acref/9780199683581.001.0001/acref-9780199683581?btog=chap&hide=true&page=156&pageSize=10&skipEditions=true&sort=titlesort&source=%2F10.1093%2Facref%2F9780199683581.001.0001%-2Facref-9780199683581

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

Scriven, M. (1972). Objectivity and subjectivity in educational research. In L. G. Thomas (Ed.), *Philosophical redirection of educational research. The seventy-first yearbook of the National Society for the Study of Education* (pp.94-142). Chicago: The National Society for the Study of Education.

Scriven, M. (1976). *Reasoning.* New York: McGraw-Hill.

Scriven, M. (1980a). *The logic of evaluation*. Thousand Oaks, CA: Edgepress.

Scriven, M. (1980b). The exact role of value judgments in science. In E. D. Klemke, R. Hollinger, & A. D. Kline (Eds.), *Introductory readings in the philosophy of science* (pp. 269-292). New York: Prometheus Books.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park: Sage Publications.

Scriven, M. (1993). The nature of evaluation. *New Directions for Evaluation, 58*, 5-48.

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

Scriven, M. (1994b). The fine line between evaluation and explanation. *Evaluation Practice, 15*(1), 75-77.

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

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

Scriven, M. (2004). Reflections. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 183-195). Thousand Oaks, CA: Sage Publications .

Scriven, M. (2007a). The logic of evaluation. In H. V. Hansen (Ed.), *Dissensus and the search for the common ground* (pp.1-16). Proceedings of the International Conference: Dissensus & the Search for Common Ground. Ontario Society for the Study of Argumentation.

Scriven, M. (2007b). *The logic and methodology of checklists*. Retrieved from http://michaelscriven.info/images/

Scriven, M. (2009). Meta evaluation revisited. *Journal of MultiDisciplinary Evaluation, 6*(11), i-vii. Retrieved from http://evaluation.wmich.edu/jmde/

Scriven, M. (2011a). *Conceptual revolutions in evaluation: Past, present and future*. Paper presented at the Symposium in honour of Michael Scriven, Claremont Graduate University.

Scriven, M. (2011b, November). *The rest of the iceberg: The logic of evaluation*. Paper presented at the 25th Annual Conference of the American Evaluation Association, Anaheim, CA.

Scriven, M. (2012b). The logic of valuing. *New Directions for Evaluation,133,* 17-28.

Scriven, M. (2013a). The foundations and future of evaluation. In S. I. Donaldson (Ed.), *The future of evaluation in society: A tribute to Michael Scriven* (pp. 11-44): Charlotte, NC: Information Age Publishing.

Scriven, M. (2013b, September). *Reconstructing the foundations of evaluation: Practical philosophy of science vs positivist philosophy of science.* Paper presented at the annual conference of the Australasian Evaluation Society, Brisbane, Australia.

SenGupta, S., Hopson, R., & Thompson-Robinson, M. (2004). Cultural competence in evaluation. *New Directions for Evaluation,102*, 5-18.

Sennett, R. (2008). *The Craftsman*. New Haven: Yale University Press.

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

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

Shadish, W. R., & Leviton, L. C. (2001). Descriptive values and social justice. In A. P. Benson, D. M. Hinn, & C. Lloyd (Eds.), *Visions of quality: How evaluators define, understand and represent program quality* (pp. 181-200). Oxford Elsevier Science.

Shank, G. (2008). Abduction. In L. M. Given (Ed.), The S*age encyclopaedia of qualitative research methods* (pp. 2-3). Thousand Oaks, CA: Sage Publications.

Simons, H. (2015). Democratic evaluation: Its power and relevance in today's world. *The Evaluator* (Spring), 6-9.

Smith, L. T. (1999). *Decolonizing methodologies: Research and indigenous peoples.* Dunedin: University of Otago Press.

Smith, L. T. (2005). On tricky ground: Researching the native in the age of uncertainty. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 85-108). Thousand Oaks, CA: Sage Publications.

Smith, M. J. (1998). *Social science in question.* London: Sage Publications.

Smith, N. L. (1981). The certainty of judgments in health evaluations. *Evaluation and Program Planning, 4*, 273-278.

Smith, N. L. (1987). Toward the justification of claims in evaluation research. *Evaluation and Program Planning, 10*, 309-314.

Smith, N. L. (1995). The influence of societal games on the methodology of evaluative inquiry. *New Directions for Evaluation, 68*, 5-14.

Smith, N. L. (2009). Fundamental issues in evaluation. In K. E. Ryan & B. Cousins (Eds.), *The Sage International Handbook of Educational Evaluation* (pp.37-50). Thousand Oaks, CA: Sage Publications.

Smith, N. L. (2010). Characterizing the evaluand in evaluating theory. *American Journal of Evaluation, 31*(3), 383-389.

Snell, M. (2011). *Cost-benefit analysis: A practical guide* (2nd ed.). London: Thomas Telford.

Social Policy Evaluation and Research Committee. (2008). SPEaR good practice guidelines. Wellington: Author.

Social Policy Evaluation and Research Unit, & Aotearoa New Zealand Evaluation Association. (2015). *Evaluation standards for Aotearoa New Zealand.* Retrieved from superu.govt.nz

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Stainton Rogers, R. (2005). Q methodology. In J. A. Smith, R. Haree, & L. V. Langenhove (Eds.), *Rethinking methods in psychology* (pp.178-207). London: Sage Publications.

Stainton Rogers, R., Stenner, P., Gleeson, K., & Stainton Rogers, W. (1995). *Social Psychology: A critical agenda*. Cambridge: Polity Press.

Stake, R. E. (2001). Representing quality in evaluation. In A. Benson, D.M. Hinn & C. Lloyd (Eds.), *Visions of quality: How evaluators define, understand and represent program quality* (pp. 3-12). Oxford: Elsevier Science.

Stake, R. E. (2004). *Standards-based and responsive evaluation*. Thousand Oaks, CA: Sage Publications.

Stake, R. E. (2013). The people and the profession. In S. I. Donaldson (Ed.), *The future of evaluation in society: A tribute to Michael Scriven* (pp. 107-114). Charlotte, NC: Information Age Publishing.

Stake, R. E., Migotsky, C., Davis, C., Cisneros, E. J., Depaul, G., Dunbar, G., . . . Chaves, I. (1997). The evolving syntheses of program value. *Evaluation Practice, 18*(2), 89-104.

Stake, R. E., & Schwandt, T. A. (2006). On discerning quality in evaluation. In I. F. Shaw & J. C. Greene (Eds.), *The Sage handbook of evaluation*. London: Sage Publications.

State Services Commission. (2011). *Better Public Services draft issues paper: Results*. Retrieved from http://www.ssc.govt.nz/sites/all/files/bps-2256063.pdf

State Services Commission. (2015). Annual report for the year ending 30 June 2015. Retrieved from http://www.ssc.govt.nz/ar2015.

State Services Commission. (n.d.). *Better Public Services*. Retrieved from https://www.ssc.govt.nz/better-public-services

Statistics New Zealand. (n.d.). *Snapshots of New Zealand*. Retrieved from http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz.aspx

Steering Group for the Managing for Outcomes Roll-out 2004/05. (2002). *Managing for outcomes: Guidance for departments*. Wellington, New Zealand. Retrieved from http://www.ssc.govt.nz/upload/downloadable_files/Doing_the_Right_Things...imrpoving_evaluative_activity.pdf

Steering Group for the Managing for Outcomes Roll-out 2004/05. (2003). *Learning from evaluative activity: Enhancing performance through outcome-focused management*. Wellington, New Zealand. Retrieved from http://www.ssc.govt/upload/downloadable _files/Learning _from_Evaluative_Activity.pdf

Stenner, P. (2009). Between method and ology: Introduction to special issue. *Operant Subjectivity: The International Journal of Q Methodology, 32*, 1-5.

Stenner, P. (2011). Q methodology as qualiquantology: Comment on Susan Ramlo and Isadore Newman's "Q Methodology and its position in the mixed methods continuum" *Operant Subjectivity: The International Journal of Q Methodology, 34*(3), 192-203.

Stenner, P., & Stainton Rogers, R. (2004). Q methodology and qualiquantology: The example of discriminating between emotions. In Z. Todd, B. Nerlich, B. S. McKeown, & D. D. Clarke (Eds.), *Mixing methods in psychology: The integration of qualitative and quantitative methods in theory and practice* (pp. 99-118). New York: Psychology Press.

Stephenson, W. (1953). *The study of behaviour: Q-Technique and its methodology*. Chicago: The University of Chicago Press.

Stephenson, W. (1961). Scientific creed 1961: Abductory principles. *Psychological Record, 11*, 9-17.

Stephenson, W. (1993/1994). Introduction to Q methodology. *Operant Subjectivity: The International Journal of Q Methodology, 17*(1/2), 1-13.

Stern, E. (2006). Contextual challenges for evaluation practice. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage Publications.

Stevenson, L. F. (2005). Epistemology. In *The Oxford companion to philosophy* (2nd ed., p.258). Oxford: Oxford University Press.

Stufflebeam, D. L. (1999). *Program evaluations metaevaluation checklist*. Retrieved from www.wmich.edu/evalctr/checklists

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

Stufflebeam, D. L. (2001b). The meta-evaluation imperative. *American Journal of Evaluation, 22*(2), 183-209.

Suaalii-Sauni, T. (2015, July). *Englishing Pasifika/Pacific-ness in programme evaluation*. Paper presented at the annual conference of the Aotearoa New Zealand Evaluation Association, Auckland, New Zealand.

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. Chicago: The University of Chicago Press.

Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory, 30*(3), 167-186.

Taylor, P. W. (1961). *Normative discourse*. Englewood Cliffs: Prentice-Hall.

Te Ara Encyclopedia of New Zealand. The number eight wire tradition. Retrieved 6 January 2016 from http://www.teara.govt.nz/en/inventions-patents-and-trademarks/page-1

Te Puni Kōkiri. (2001). *He Tirohanga ō Kawa ki te Tiriti o Waitangi: A guide to the principles of the Treaty of Waitangi as expressed by the Courts and the Waitangi Tribunal*. Wellington: Author.

Toulmin, S., Rieke, R., & Janik, A. (1979). *An introduction to reasoning*. New York: Macmillan Publishing.

Trotman, I. (2003). Evaluation in New Zealand: A founder’s perspective. In N. Lunt, C. Davidson, & K. McKegg (Eds.), *Evaluating policy and practice: A New Zealand reader* (pp. 21-39). Auckland: Pearson Education.

Turner, S. (1994). *The social theory of practices: Tradition, tacit knowledge and presuppositions*. Cambridge: Polity Press.

United Kingdom Government (1999). *Modernising Government*. Retrieved from https://www.wbginvestmentclimate.org/uploads/modgov.pdf

Vaioleti, T. (2006). Talanoa research methodology: A developing position on Pacific research. *Waikato Journal of Education,12*, 21-34. Retrieved from http://whanauoraresearch.co.nz/files/formidable/Vaioleti-Talanoa.pdf

Valovirta, V. (2002). Evaluation utilisation as argumentation. *Evaluation, 8*(1), 60-80.

van Exel, J., & de Graaf, G. (2005). *Q methodology: A sneak preview*. Retrieved from www.jobvanexel.nl/

Watts, S. (2011). Subjectivity as operant: A conceptual exploration and discussion. *Operant Subjectivity: The International Journal of Q Methodology, 35*(1), 37-47.

Watts, S., & Stenner, P. (2012). *Doing Q methodological research: Theory, method and interpretation*. London: Sage Publications.

Wehipeihana, N. (2008). Indigenous evaluation: A strategic objective of the Australasian Evaluation Society. *Evaluation Journal of Australasia, 8*(1), 40-44.

Wehipeihana, N. (2013, October). *A vision for indigenous evaluation.* Paper presented at the annual conference of the Australasian Evaluation Society, Brisbane, Australia. Retrieved from https://nanwehipeihana.wordpress.com/2013/11/01/a-vision-for-indigenous-evaluation/

Wehipeihana, N., Bailey, R., Davidson, E. J., & McKegg, K. (2014). Evaluator competencies: The Aotearoa New Zealand experience. *The Canadian Journal of Program Evaluation, 28*(3), 49-69.

Weiss, C. (1998). *Evaluation: Methods for studying programs and policies.* Englewood Cliffs: Prentice-Hall.

White, P., & Boulton, A. (2011). Sailing through relationships? On discovering the compass for navigating the 21st-century evaluation in the Pacific. *New Directions for Evaluation, 131*, 71-76.

Wihongi, H. A. (2010). *Tino Rangatiratanga in health policies and practices.* (Doctoral dissertation, University of Waikato, Hamilton, New Zealand). Retrieved from http://hdl.handle.net/10289/4819

Williams, B. (2003). Getting the stuff used. In N. Lunt, C. Davidson, & K. McKegg (Eds.), *Evaluating policy and practice: A New Zealand reader* (pp. 196-213). Auckland: Pearson Education.

Wolf, A. (2008/2009). Subjectivity, the researcher, and the researched. *Operant Subjectivity: The International Journal of Q Methodology, 32*, 6-28.

Wolf, A. (2012, February). *Q methodology*. Workshop presented at the New Zealand Social Statistics Network. Victoria University of Wellington, Wellington, New Zealand.

Wolf, A., Peace, R., & Brown, S. R. (2015, September). *Q methodology and the varieties of abductive thinking*. Paper presented at the 31st annual conference of the International Society for the Scientific Study of Subjectivity, Universita Politecnica delle Marche, Ancona, Italy.

APPENDIX A:

Q STUDY - INFORMATION SHEET FOR PARTICIPANTS

Evaluative reasoning in public sector evaluation in Aotearoa New Zealand

INFORMATION SHEET: Q Sort

I am undertaking a PhD study about evaluative reasoning in public sector evaluation in Aotearoa New Zealand. The study uses a multi-method design to investigate how evaluative reasoning is understood and practiced by evaluators working in or for the public sector. The methods include Q methodology, document review, key informant interviews and a literature review.

Invitation to participate

You are invited to be part of this study by participating in a Q Sort. This involves sorting 35 statements according to how they align with your point of view as a professional evaluator. On completion of the sort, you will be asked four short questions about your reasons for the way you sorted the statements. The Q Sort will take no longer than one hour. It will occur at a time and place convenient to you.

Participant identification and recruitment

It is hoped that thirty evaluators will agree to participate in a Q sort. A purposive approach is being used to identify potential participants. You were identified as a potential participant through informal evaluation networks. The participant group will include internal and external evaluators who are working in different areas of evaluative practice across the public sector. Māori, Pasifika and Pākehā evaluators will be included.

Data confidentiality

Before the Q sort you will be asked to identify yourself by providing a 3 digit code which will become your unique identifier. This identifier will be used on all research documentation (your name or identifying information will not be used). This identifier will enable you to identify your results when the findings from the 30 Q sorts are sent to you on completion of this stage of the research.

The data will only be used for the purposes of this study. It will be stored on a password protected computer in the researcher's home (with a burglar alarm) and on a back-up hard drive (stored in a locked filing cabinet in the researcher's home). The electronic files will be deleted following the study's completion.

Participant rights

You are under no obligation to accept this invitation. If you decide to participate, you have the right to:

- withdraw from the study at any time
- ask any questions about the study at any time during participation
- provide information on the understanding that your name will not be used unless you give permission to the researcher
- be given access to a summary of the project findings when it is concluded.

Project Contacts

My contact details and my supervisor's contact details are as follows. Please contact me or my supervisor if you have any questions about the project.

Researcher: **Heather Nunns** 17 Ronald Woolf Place Wellington 6037 heather@analyticmatters.co.nz 027 3329 785, (04) 478 2248.	**Supervisor:** **Dr Robin Peace** Associate Professor School of People, Environment and Planning Massey University Box 756, Wellington 6140 R.Peace@massey.ac.nz (04) 801 5799 ext 62172.

Ethical conduct

This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of Massey University's Human Ethics Committees. The researcher named above is responsible for the ethical conduct of this research.

If you have any concerns about the conduct of this research that you wish to raise with someone other than the researcher or supervisor, please contact Professor John O'Neill, Director, Research Ethics, telephone (06) 350 5249, email: humanethics@massey.ac.nz.

APPENDIX B:

Q STUDY - PARTICIPANT CONSENT FORM

Evaluative reasoning in public sector evaluation in Aotearoa New Zealand

Q SORT PARTICIPANT CONSENT FORM

I have read the Information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I agree to participate in the Q Sort under the conditions set out in the Information Sheet.

Signature:.. Date:..

Full Name (printed):..

APPENDIX C:

Q STUDY - INSTRUCTIONS FOR Q PARTICIPANTS

Q methodology:

Instructions for Q participants

Here are 35 cards. Each card contains a statement made by an evaluator about what evaluation is, or how it should be conducted.

Please sort the statements to reflect your point of view as a professional evaluator, based on a +4 to -4 scale:

- +4 being the two statements that are *most similar* to your views
- -4 being the two statements that *most different* to your views.

A velcro board is provided for you to place your cards. Please attach the specified number of cards (shown by the velcro dots) based on how much you think the statements are the same or similar to your views.

As a first step you may want to read through all of the cards and place each card in one of three piles:

- **pile 1:** the statements that are the *same or very similar* to your views
- **pile 2:** the statements that you are *unsure* about
- **pile 3:** the statements that are *very different* to your views.

Pick out the TWO cards that are *most* similar to your own views and place them on the velcro dots below +4. Then, pick the TWO cards that are *most different* to your views, and place them on the velcro dots below -4.

From the remaining cards pick THREE cards which are most similar to your views, and place them on the three velcro dots below +3. Then pick THREE cards which are most different to your views, and place them on the three velcro dots below -3, and so on.

You should end up with seven cards - these are the statements which are neither similar to your views or different from your views, or about which you have no clear opinion. Please put them on the seven velcro dots under 0. (Don't worry if your views about similarity and difference do not exactly line up with the +4/-4 scale).

At the end, all of the cards should be stuck onto the velcro board. Once you have done this, please indicate that you are finished.

Finally, before you start there are two small things to note:

- Each card includes a number at the bottom of the card. These numbers do not mean anything. These numbers help the researcher to record your placement of individual cards.
- The word 'evaluand' is used in some of the statements. It means 'the thing that is being evaluated', for example, a policy, programme, strategy, organisation etc.

Any questions about what you are being asked to do?

APPENDIX D:

Q STUDY - INFORMATION ABOUT Q SENT TO INTERESTED PARTICIPANTS POST-SORT

Q methodology: a brief overview

Heather Nunns, October 2012

Thank you for participating in a Q Sort for my PhD study. You expressed an interest in finding out more about Q methodology (Q).

About Q: theory

Q methodology provides for the "systematic study of subjectivity" (Brown, 1991) where subjectivity is defined as an individual's point of view (McKeown & Thomas, 1988) or first person perspective (Watts & Stenner, 2012). Unlike qualitative research methods where the researcher is an interpretive intermediary of the research participant's point of view, Q enables an individual's viewpoint to be captured without distortion. Q methodology is based on the premise that while subjectivity is unable to be 'proved', it can be shown to have structure and form (McKeown & Thomas, 1988). Watts and Stenner (2012, p.30) describe Q as "making a science of the subjective".

According to William Stephenson (1902-1989) the creator of Q methodology, subjectivity is a behaviour or activity in relation to the immediate environment (as opposed to a mental concept): "Viewpoints have no existence in the absence of some behavioural engagement with their object, and being made up of activity, are subject change and transition" (Watts & Stenner, 2012, p.44).

Stephenson was a student of the high profile British psychologist Charles Spearman who developed the method of factor analysis. Factor analysis is a technique of data reduction which reveals patterns of association (factors) among a series of measured variables on tests or traits using a selected population (Watts & Stenner, 2012). Stephenson inverted Spearman's method to enable by-person factor analysis, that is, "a population or sample of tests (or other items) are measured or scaled relatively by a collection of individuals" (Watts & Stenner, 2012, p.15). Stephenson used the letter "Q" to distinguish this approach, hence the name Q methodology.

Conducting a Q study

Stephenson used the term "concourse" to describe "the flow of communicability surrounding any topic" (Stephenson 1978, cited in Brown 1991).[20] A concourse comprises words, pictures or objects about a topic, from the formal (e.g. a report) to the informal (e.g. a private conversation or cartoon). The researcher develops a group of statements (or pictures) about the research topic for the Q sample which is designed to provide "a representative miniature" of the larger concourse (Brown, 1991).

Participants are asked to sort or rank the statements (or pictures) according to a specific instruction (referred to as "the condition of instruction"). This sorting or ranking is referred to as a Q sort. The Q sort enables the participants to express his/her viewpoint on the topic (McKeown & Thomas, 1988). The participant's subjectivity is expressed in how the statements (or pictures) are understood and how they are ranked (Brown, 1991).

The Q sort results are subject to correlational and Q factor analysis (using Q software) to identify patterns of association, referred to as "orientations", and the extent of each participant's association with a particular orientation. Wolf (2012) identifies two broad stances researchers may use to interpret patterns from Q factor analysis. Firstly, in a person-centred Q study, the researcher enquires into the ways in which people view a matter from their perspective and the underlying predispositions that may influence a person's response to the items in the Q sort. In a discourse-centred Q study, the researcher is interested in the discourses to which people align.

Want to learn more?

Dr Amanda Wolf, School of Government, Victoria University of Wellington is an

[20] Van Excel and de Graaf (2005) note that concourse should not be confused with discourse. The concourse refers to all of the relevant aspects of all of the discourses on a topic.

expert in Q methodology (Amanda.wolf@vuw.ac.nz) and runs workshops on Q through the New Zealand Social Statistics Network - I highly recommend her workshop.

Below are five references I have found useful. I particularly recommend the Watts and Stenner book – it is a very readable text describing the theory underpinning Q, as well as providing a step by step account of how to conduct a Q study.

Brown, S. R. (1991). Q Methodology. Qualitative research for the human sciences. Retrieved http://facstaff.uww.edu/cottlec/QArchive/Primer1.html

McKeown, B., & Thomas, D. (1988). Q methodology (Vol. 66). Newbury Park: Sage.

van Exel, J., & de Graaf, G. (2005). Q methodology: A sneak preview. Retrieved www.jobvanexel.nl/

Watts, S., & Stenner, P. (2012). Doing Q Methodological Research: Theory, method and interpretation. London: Sage.

Wolf, A. (February 2012). Q Methodology. Workshop presented at the New Zealand Social Statistics Network, Victoria University of Wellington.

Below are examples of research studies that used Q methodology. There are numerous other Q studies available on the web.

- Q was used to understand the experiences of people who have had a transient ischemic attack (i.e. a mini stroke) http://www.hindawi.com/journals/srt/2012/486261/
- Researchers used Q to explore how people who hear voices construe their experience http://www.psychminded.co.uk/news/news2003/july03/qmethodological.pdf
- The purpose of this study was to better understand opposition and support for renewable energy options in Ireland *www.qub.ac.uk/research-centres/ . . . /Filetoupload,32040,en.ppt*

APPENDIX E:

Q STUDY - FACTOR ARRAYS

An asterisk represents a distinguishing statement.

	Item	Factor 1	Factor 2	Factor 3
1	Evaluators need to maintain a detached stance from an evaluand so they can provide a 'distanced view'. This requires minimum interaction with staff involved with the evaluand. It's the only way to ensure an independent and objective assessment of the evaluand	-4	-3	-4
2	Stakeholders should not have any input into the evaluation process. Assessing the performance or quality of an evaluand is the sole responsibility of the evaluator	-4	-4	-4
3	The evaluator should provide a 'thick description' of the evaluand so stakeholders can make up their own minds about the evaluand's performance and quality	-1	2*	-3*
4	It is the primary responsibility of stakeholders, not the evaluator, to make evaluative judgments. The evaluator should only describe and report the various perspectives about the evaluand and make descriptive statements such as 'if you value A, then B is the case'	-2	-3	-2
5	The evaluator should not provide any assessments of an evaluand's quality or performance. Instead she should give the information she has gathered about the evaluand to those who want to assess its operations or achievements	-3	-3	-3
6	As evaluators, we hope our work will make a difference by helping to create a better world. However, obligations to our political masters, clients, stakeholders and informants mean this desire often has to be moderated, and sometimes it's tough	2*	0	1

	Item	Factor 1	Factor 2	Factor 3
7	Evaluators have responsibilities to multiple audiences, including the all-powerful policy makers, the 'powerless' people who are often the recipients of the programmes and policies we evaluate, and the general public	4	1	4
8	The role of the evaluator is to facilitate a structure for stakeholders to engage in an evaluation, and to work with stakeholders to assess how an evaluand has performed	0	1	1
9	Every institution has values that are so embedded that they have become 'taken for granted' aspects of the institution's practices. Such values must not be taken for granted if evaluators are to provide a neutral, external perspective about the institution	1*	4*	0
10	As evaluators, we don't have free rein. We're constrained by the context in which the evaluand operates, the politics at work in and around the evaluand, and the politics of the government	2	-2*	2
11	One key thing about evaluators' work is that we're always in the midst of cultural norms, values, and ways of knowing	3	0	1
12	Methodology can be thought of as the best approach to obtain the data required for an evaluation. More importantly, methodology is about issues of power and control – that is, whose interests an evaluation will serve	0	-1	0
13	The 'value-free' debate misses the point – as evaluators, what we think, how we practice, the methods we use . . . they're all value-laden	1	0	2
14	Evaluation theories are not value neutral. They reflect implicit and explicit assumptions about how things work. We need to critically examine evaluation theories to identify culturally embedded perspectives	0	1	2
15	As evaluators, we need to be conscious of the implicit cultural values that have shaped the practice of professional evaluation	0	1	0

	Item	Factor 1	Factor 2	Factor 3
16	As evaluators, we need to understand how our 'value lens' influences our perceptions of what we're evaluating, its context, and how we understand stakeholders' perspectives	**2**	**3**	**3**
17	Public sector evaluation and emotion are closely linked. Beliefs about the standards and criteria that should be used to evaluate policies and programmes are closely tied to deeply held ideological positions about what 'the good society' or 'public good' looks like	**1**	**0**	**0**
18	Evaluations of public sector initiatives or other public services such as education should focus on effectiveness and efficiency. Appropriate evaluation methods include cost benefit analysis, and standardized measures	**-2**	**2***	**-1**
19	An evaluand should be measured against its objectives to avoid stakeholders debating what 'good performance' or 'quality' means in relation to the evaluand	**-2**	**-2**	**-1**
20	The process for identifying the standards to evaluate an evaluand is a critical aspect of evaluation. It involves thoughtful dialogue among diverse stakeholders. For what constitutes a 'good' or 'quality' evaluand in a particular context is often a matter of much debate	**3**	**3**	**3**
21	The evaluator needs to select the standards against which an evaluand will be assessed that are most relevant to the client, recipients and stakeholders	**1**	**-2***	**1**
22	Evaluators should not identify explicit standards of performance in advance of an evaluation. Instead, they should seek to understand what is the nature of 'quality' in the situation in which they are working. Such understanding is emergent and incremental	**0**	**3**	**0***
23	It is acceptable for the standards used in an evaluation to be assumed (for example, by use of programme goals and objectives) or for standards to remain implicit in the evaluation process	**-3**	**-1**	**-3**

	Item	Factor 1	Factor 2	Factor 3
24	The evaluation profession has given much of its attention to methods that will generate the required data. This has been at the expense of understanding what is involved in developing a defensible argument, that is, a clear chain of reasoning that connects the evidence to an evaluative conclusion	**0***	**-1**	**-1**
25	One of our core responsibilities as evaluators is to ensure our evaluative claims and conclusions are legitimate and justified. Because our claims and conclusions are based on 'all things considered' inferences, they must be defensible	**2**	**4**	**4**
26	Evaluators have to appeal to an audience's reason and understanding to persuade them that the findings of an evaluation are plausible and reasonable. We do this through argument	**-1**	**2***	**0**
27	The only way to produce legitimate and defensible evaluative conclusions is for the evaluator to use inclusive processes which capture the perspectives of multiple stakeholders	**0**	**0**	**3**
28	Getting to an evaluative conclusion requires the evaluator to be analytic and dispassionate. It involves the evaluator being an empiricist and logician	**-3**	**2***	**-2**
29	Evaluation does not involve the precise comparison of an evaluand's performance to standards. Instead, the evaluator's experience enables him to interpret data in an intuitive manner	**-2**	**-4**	**-1**
30	Evaluative conclusions are inextricable blends of fact and value claims. They involve the evaluator combining relevant multiple criteria and interests into 'all-things-considered' judgments	**1**	**0**	**2***
31	We should not assert that our evaluative judgments are clear-cut. We should avoid absolute statements about performance and instead use comparative statements.	**-1**	**-1**	**-1**

	Item	Factor 1	Factor 2	Factor 3
32	We should never give up the evaluator's responsibility to provide evaluative judgments, but provide them softly framed so as to encourage the reader's own interpretations	**-1**	**-2**	**-2**
33	There is no one way of portraying the role of the evaluator in valuing. There are a variety of evaluator approaches and each carries with it different implications for the way that evaluation is conducted	**3**	**1**	**0**
34	I argue that just as we have multiple evaluation methods in our toolkits which we fit with the context, so too there are a variety of approaches for valuing. We must consider which valuing approach works best with whom and under what circumstances	**4**	**-1***	**1**
35	To evaluate is less about judging, and more about describing, explaining and informing. These aspects must be emphasised if evaluation is to be a useful tool for social science research	**-1**	**0**	**-2***

APPENDIX F:

META-EVALUATION - EVALUATION REPORT SAMPLE

Report number[21]	Author type	Commissioning and/or funding agency	Type of evaluand	Orientation	Evaluation approach/ method
1	External	Sport New Zealand	Sport and recreation programme	Development	Developmental, Kaupapa Māori approach
2	Internal	Ministry of Business, Innovation and Employment	Housing programme	Management	Qualitative
5	Internal	Tertiary Education Commission	Training programme	Management	Formative, qualitative
6	Internal	Ministry of Social Development	Childhood parenting intervention	Accountability	Multi-method
7	Internal	Ministry of Social Development	Personal safety programme	Management	Qualitative
8	Internal	Department of Corrections	Rehabilitation programme	Management	Formative
10	External	Ministry of Justice	Criminal justice programme	Management	Implementation, outcome
11	Internal	Department of Conservation	Conservation policy	Accountability	Desk-based, secondary data
12	External	Ministry of Foreign Affairs and Trade	Development aid strategy	Accountability	Development evaluation
14	External	Land Information New Zealand	Economic policy framework	Management	Mixed method, including cost effectiveness

Report number[21]	Author type	Commissioning and/or funding agency	Type of evaluand	Orientation	Evaluation approach/ method
15	External	New Zealand Fire Service	Safety programme	Management	On-line survey and case studies
16	External	Health Promotion Agency	Media campaign	Management	Telephone survey only
17	External	Te Puni Kōkiri	Community development programme	Management	Kaupapa Māori approach, process and outcome
18	External	Alcohol Advisory Council of New Zealand	Therapeutic programme	Accountability	Summative, impact
19	External	Ministry of Foreign Affairs and Trade	Development aid programme	Accountability	Development evaluation
21	External	New Zealand Transport Agency	Roading infrastructure project	Accountability	Cost benefit analysis, desk-based
25	External	Ministry of Education	Educational programme	Management	Māori research methodology
26	Internal	Ministry of Primary Industries	Economic policy	Accountability	Desk-based, secondary data
27	External	Ministry of Education	Educational programme	Management	On-line survey only
30	External	Ministry of Health	Health-related intervention (not a programme)	Accountability	Value for money, desk-based, secondary data

Report number[21]	Author type	Commissioning and/or funding agency	Type of evaluand	Orientation	Evaluation approach/ method
34	External	New Zealand Transport Agency	Research use	Management	Qualitative and quantitative
35	Internal	Education Review Office	Educational programme	Accountability	On-line survey and qualitative
36	Internal	Tertiary Education Commission	Governance arrangements	Management	Implementation
37	Internal	Education Review Office	Educational programme	Management	Qualitative
38	External	Department of Labour	Workplace programme	Management	Implementation evaluation
39	Internal	Department of Labour	Employment policy	Management	Implementation evaluation
41	External	Ministry of Primary Industries	Economic development funding programme	Management	Multi method, including value for money
42	External	Ministry of Economic Development	Housing-related health intervention	Accountability	Cost benefit analysis, desk-based
43	External	Ministry of Health	Health intervention (not a programme)	Management	Health Impact Assessment
44	External	Ministry of Health	Health intervention (not a programme)	Management	Health Impact Assessment, developmental

[21] The numbers shown in the column below are not consecutive due to some reports being removed during the sample selection process

APPENDIX G:

META-EVALUATION - REPORT RECORDING SHEET

Report no. and title: ***Internal/external authors:***	
Information about evaluation purpose/evaluand/context/methods/audience:	
Evaluative evaluation objectives or questions:	
Comparator & standards	
Comparator: yes/no **If yes: criteria or other?** **Extent of definition?**	
Comparator identification (and definition) - how and who? **Justification of the comparator?**	
Standards of performance: If yes: how were they identified and by whom? **Justification of the standards?** **If no: what?**	

Warranted argument and judgment(s)	
Warranted argument connecting claim(s) and evidence: **Nature of warrant/ backing?** **Strength of warrant/ backing?**	
Evaluative judgment(s)? **If yes:** **(1) how info synthesised into a judgment?** **(2) judgment(s) relate to the purpose/objectives?** **If no evaluative judgment, what is provided?**	
Limitations section? **Contextual factors that have influenced:**	
Overall comments: clear chain of evaluative reasoning?	

APPENDIX H:

META-EVALUATION - EXAMPLES OF GENERIC STANDARDS

Generic standards ('traffic lights') for a value for money evaluation (report 30)

Red	**R**	This driver suggests value for money (VfM) is poor in this area. Significant opportunities for improvement exist.
Amber	**A**	This driver suggests VfM is fair in this area. Some opportunities for improvement exist.
Green	**G**	This driver suggests VfM is good in this area.
Grey	**U**	Data for this driver was insufficient to provide a VfM conclusion. If confidence in the data for this driver was assessed as Red we have not provided a VfM conclusion.

Generic standards used in a programme evaluation (report 17)

Dimensions of merit	**Descriptor**
Fully achieved	Major improvements or achievements of practical significance; no major issues or gaps.
Mostly achieved	Good progress or achievements for the time and money invested; if there are issues or gaps these are actively being addressed.
Partially achieved	Some progress; but less than expected for the time and money invested; issues or gaps may be receiving some attention but require increased or additional action.
Minimally achieved	Little progress; significantly held back by major issues or gaps that are not: recognised, acknowledged or actively being addressed.
Not achieved	No evidence of any progress.
Insufficient evidence	Insufficient evidence to reach a sound evaluative conclusion.

APPENDIX I:

META-EVALUATION - EXAMPLES OF TAILORED STANDARDS

Report 14 is an evaluation of a regulatory policy framework. The evaluation was based on three criteria - relevance, effectiveness and efficiency. The standards for the efficiency criterion are shown below.

EFFICIENCY

Excellent

- Achieved reduction in costs beyond what was reasonably anticipated.
- The system itself has a reduced cost and other stakeholders share/experience efficiencies.
- Runs with minimal need for intervention or additional inputs from government.
- Purposefully designed system with clear expectations on the roles and responsibilities and meets stakeholders' accountability requirements in a timely manner.
- The system supports minimal duplication of effort and stakeholders leverage existing processes or information for other purposes.
- The system has sufficient people with the 'right' capability and capacity to meet the needs (now and for the future).
- The objections and Tribunal process works in a timely, resource and cost-efficient manner.
- The (name of position) exercises his powers to oversee the Framework in a balanced and prudent manner.

Good

- Achieved and maintained a reasonable reduction in costs.
- The system itself has a reduced cost and other stakeholders have opportunities to share efficiencies.
- Runs with minimal need for intervention or additional inputs from government.
- Purposefully designed system with clear expectations on the roles and responsibilities and meets stakeholders' accountability requirements.
- The system supports minimal duplication of effort and stakeholders have the opportunity to leverage existing processes or information for other purposes.
- The system currently has sufficient people with the 'right' capability and capacity to meet the needs.
- The objections and Tribunal process generally works in a timely, resource and cost-efficient manner.
- The (name of position) exercises his powers to oversee the Framework in a balanced and prudent manner.

Adequate

- Achieved a reasonable reduction in costs.
- The system itself has a reduced cost.
- Runs with some/acceptable intervention or additional inputs from government.
- System has expectations on the roles and responsibilities and meets stakeholders' accountability requirements.
- The system has minimal duplication of effort.
- The system has people with capability and capacity to meet the needs.
- The objections and Tribunal process sometimes works in a timely, resource and cost-efficient manner.
- The (name of position) usually exercises his powers to oversee the Framework in a balanced and prudent manner.

Inadequate

- Provides (name of organisation) with the ability to apportion rates but the system present obstacles to successfully striking rates.
- Central government cannot rely on the apportioning rates system.
- The Framework provides Third Parties with inadequate guidelines and/or Third Parties face significant barriers to entry and development.
- Ratepayers don't know or understand how their rating valuation is determined and cannot access information.
- The objections and Tribunal does not provide an avenue for ratepayers to have their rating valuation reviewed, or to appeal the reviewer's decision.
- The (name of position) is insufficiently empowered to give effect to his/her statutory function.

APPENDIX J:

EXPERT INTERVIEWS - SUMMARY OF FINDINGS FOR EXPERTS TO READ PRIOR TO THE INTERVIEW

Evaluative reasoning:
Understanding and practice in
the New Zealand public sector

Heather Nunns
heather@analyticmatters.co.nz

Thank you for agreeing to be interviewed for my PhD study. This paper summarises the research and key findings, and provides two questions you are asked to respond to.

About the research

The purpose of the research is to understand how evaluative reasoning is understood and practised by New Zealand evaluators working for, or commissioned by, public sector agencies. Evaluative reasoning is defined as "the systematic means for arriving at evaluative conclusions, the principles that support inferences drawn by evaluators" (Fournier, 1995, p.1). The research methods used are (in sequential order): a review of literature, Q methodology (to explore how public sector evaluators understand evaluative reasoning), and meta-evaluation of public sector evaluation reports (to examine evaluative reasoning practice). My hunch is that the research findings (summarised on the following pages) reflect to some extent particular aspects of evaluation practice in New Zealand. I want to explore this hunch with New Zealand-based evaluation experts, and international evaluation experts with knowledge about New Zealand.

Questions

You are asked to consider two questions about the summary of research findings below:

1. *Are the findings surprising? If yes, why? If no, why not?*
2. *Is there anything about the findings that is unique to New Zealand? If yes, what? why?*

   If not, why?

Summary of Q methodology findings

Q methodology (which uses quantitative and qualitative methods) provides a means of capturing individuals' perspectives about a specific topic, in this case evaluative reasoning, and enables latent patterns across these perspectives to emerge that may not be revealed by non-statistical methods. These patterns (which are statistically significantly different), referred to as 'orientations', are interpreted by the researcher. The Q study revealed three orientations as follows (each orientation has been named according to its main themes):

orientation 1 - the idealist but pragmatic evaluator with an eclectic approach

orientation 2 - the analytic evaluator focused on building a convincing case

orientation 3 - the judgment-centred evaluator using inclusive practices.

There are five shared themes across the three orientations, as follows.

a. Evaluation is about and involves values. Each orientation gives emphasis to different values, such as institutional values, cultural values, values implicit in evaluation theory, and evaluator values.

b. The purpose of evaluation and the evaluator's role is to make defensible evaluative conclusions/ judgments.

c. An independent evaluative conclusion/judgment does **not** require the evaluator to be detached and distanced from the evaluand, or for stakeholders to be excluded from the evaluation process.

d. The evaluator is not the expert about the evaluand but values stakeholders' knowledge of the evaluand and its context. This means stakeholders should be involved in the evaluation process, particularly in developing evaluative criteria and standards.

e. Evaluation practice is based on relationships and dialogic approaches, for example involving stakeholders in developing criteria.

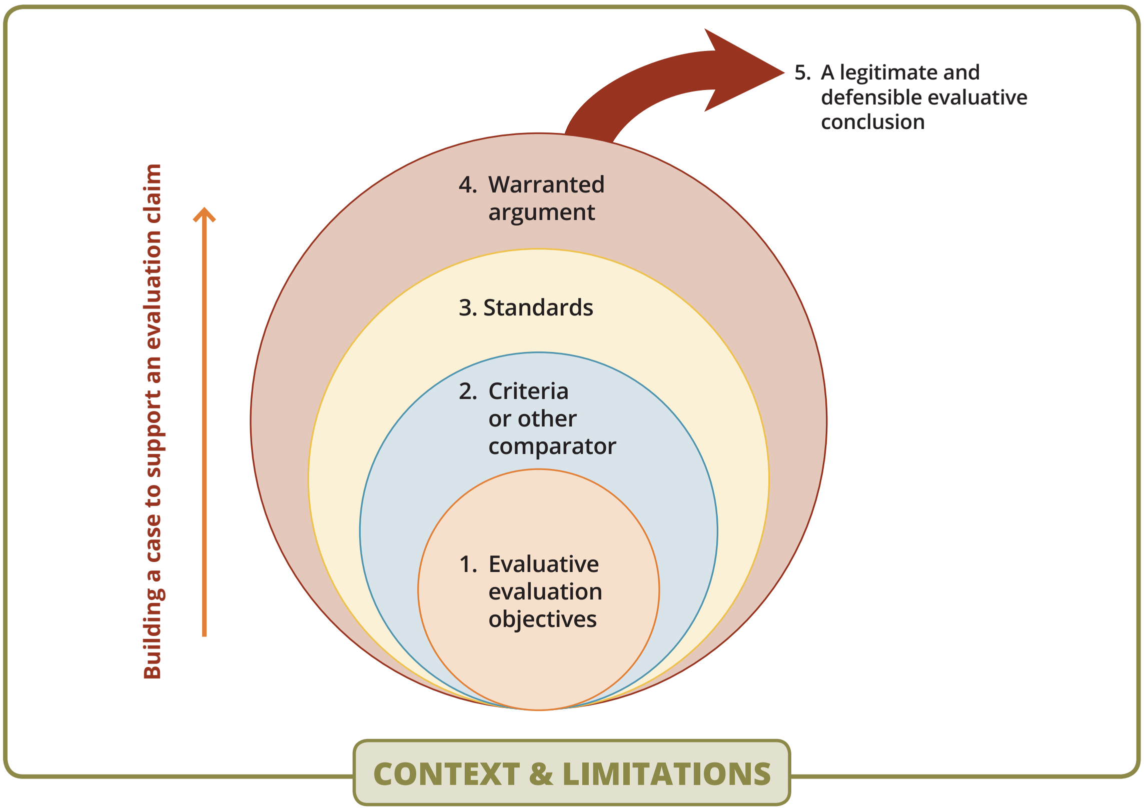
In summary, differences expressed in the three orientations are about nuance and emphasis, rather than dissimilarities of a more fundamental nature. This suggests homogeneity in evaluator perspectives and approaches.

Summary of meta-evaluation findings

A meta-evaluation was undertaken of a non-representative sample of 28 publically accessible evaluation reports written or commissioned by 20 New Zealand public sector agencies during 2010-2013. The reports were examined to find evidence of five key elements (listed below) of evaluative reasoning based on a conceptual framework drawn from the literature for building an argument or case to support an evaluative conclusion/judgment that is valid and defensible (see figure on page 3). The research examined whether each of the five elements is present, rather than assessing their quality.

1. evaluative evaluation objectives/questions
2. criteria or other comparator
3. defined standards
4. warranted argument[22]
5. an evaluative conclusion or judgment.

[22] Given the research aim was to find evidence of the five elements in the reports, the standard is whether there is evidence of the element in the report, or not. While this assessment was straightforward for elements 1, 2, 3 and 5, it was less straightforward for warranted argument (element 4). The definition of warranted argument used in this study is based on that of Booth, Colomb & Williams (2008, p.109) who describe a research argument as consisting of five components: (1) a claim (2) reasons that support the claim (3) evidence that supports the reasons (4) an acknowledgment of and a response to alternatives/complications/objections, and (5) a principle which makes the reasons relevant to the claim, referred to as *the warrant.* The argument in each report was examined to determine whether these components were addressed


5. A legitimate and defensible evaluative conclusion
4. Warranted argument
3. Standards
2. Criteria or other comparator
1. Evaluative evaluation objectives
Building a case to support an evaluation claim
CONTEXT & LIMITATIONS

Only eight of the 28 reports have evidence of all five elements. Eleven reports demonstrate three or four of the elements. The most common omission is that value terms referred to in the report are not defined, for example, by criteria, indicators or in a descriptive textual definition. The final group is made up of nine reports which lack three or more of the five elements. Three of these reports end with a conclusion/judgment that uses evaluative language despite an absence of most or all of the preceding elements.

Of the five elements of evaluative reasoning examined, **warranted argument** is the element which appears to be most neglected. Eleven of the 28 reports either do not contain an argument or contain text that is ambiguous, that is, it is not clear whether the text refers to evidence or is the author’s argument. A further seven reports combine evidence and the authors’ interpretation of the evidence (argument). At least half of these seven reports contain text where it is difficult to differentiate evidence from argument. Consequently, around half of the 28 reports lack an argument or contain text that is ambiguous.

Given that valuing is “contextually embedded and dependent” (Patton, 2012, p.98), the reports were examined to understand the context for the evaluation and its impact on the evaluation. The limitations section in a report provides useful information about the factors that have impacted on the design and conduct of an evaluation. Half of the 28 reports contained no information about the limitations associated with the evaluation. This limited readers' understanding of the contextual and other factors that may have influenced the design and conduct of the evaluation.

In summary, the meta-evaluation findings indicate that the majority of evaluation reports (20 out of 28) lack one or more elements of evaluative reasoning, thereby compromising the strength of the evaluative claim(s) and argument underpinning the evaluative conclusion/judgment.

Making sense of the findings from the two methods

The findings indicate differences between the understandings about evaluation reasoning suggested in the Q study, and how evaluative reasoning is occurring in practice as indicated in the meta-evaluation. Some of this difference may be explained by the composition of the participants in the Q study and the authors of reports in the meta-evaluation. I now realise that the Q study participants are evaluators who are working 'close to' the New Zealand evaluation community, whereas some (but not all) of the report authors are 'further away' from the evaluation community. While evaluators (or professionals who do evaluations but may not refer to themselves as evaluators) may understand their role as providing an evaluative conclusion/judgment, they may lack knowledge about what is required to build an argument that results in an evaluative conclusion/ judgment that is valid and defensible. The findings also signal the need for more research to be done on the application of evaluative reasoning theory to real-life evaluation scenarios.