Developing defensible criteria for public sector evaluations

Author: Mathea Bendino Shulamith Roorda

ORCID ID:	0000-0001-6453-5248
Degree:	Doctor of Philosophy
Date of submission:	May 2020
Department:	Centre for Program Evaluation, Melbourne Graduate School of Education, The University of Melbourne

Submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy

Abstract

Criteria convey dimensions of quality and goodness as relevant for a program and its context. They also provide the first of two value premises from which one can reason to an evaluative judgement (the other value premise being standards). Selecting and justifying relevant criteria is critical to defensible evaluative reasoning, especially so in evaluations of publicly funded programs. Yet to date, much of the theory on evaluative reasoning has been at a general level, with little focus on the individual elements of reasoning, including the development of defensible criteria.

The aim of this study was to identify characteristics of defensible criteria for program evaluations. It also sought to understand how criteria are currently managed in Australian and New Zealand program evaluations. The study exemplified research as an emergent process, with findings from an initial phase of the research informing the development of an evidence-informed tool for establishing defensible criteria.

The study contributes to closing a significant gap in research on evaluation, specifically as it concerns valuing. Three characteristics were identified as important for developing defensible criteria. Two of these - inclusion of all relevant dimensions of value and authoritative sources - are required to justify criteria. A third, full description, has a role in supporting the first two characteristics, as it is only when abstract value terms are explicitly defined or described that criteria can be assessed for comprehensiveness and authoritativeness.

The first phase of the study included an in-depth systematic examination of criteria development in Australian and New Zealand program evaluation. This occurred through a survey of 137 evaluators and a review of 47 published evaluation reports. It found that explicit criteria are not routinely included in evaluation reports. The survey research provided empirical evidence that a critical element of evaluative reasoning is weak in Australian and New Zealand program evaluations.

The findings provided an evidence-based platform from which to develop a theory-informed framework for developing defensible criteria. In the second phase of the study, a conceptual framework was developed that makes several novel and significant contributions to the field of evaluation. It provides a way for

practitioners to engage with value theory and specifically normative ethical principles which deal with conceptions of good and bad. The conceptual framework was developed into a criteria matrix tool, along with a handbook to support evaluation practitioners to engage with normative ethical perspectives. Initial field testing provided proof of concept that the tool could support evaluators to identify dimensions of value that might otherwise be ignored.

Declaration

This is to certify that:

- The thesis comprises only my original work towards the PhD except where indicated in the Preface;
- Due acknowledgement has been made in the text to all other material used;
- The thesis contains fewer than 100,000 words, exclusive of tables, references and appendices.

Mathea B. Roorda

23 May 2020

Preface

This thesis is comprised of my own original work. I have conceptualised, researched, and undertaken the data collection, analyses, discussion and conclusions. Work from Chapters 8 and 9 has been published in a co-authored manuscript as outlined below. The second author provided feedback, commentary and advice, and journal editors mandated various changes in accordance with their publication guidelines.

Publication integral to this thesis:

Roorda, M. and Gullickson, A.M. (2019). Developing evaluation criteria using an ethical lens. *Evaluation Journal of Australasia*, 19 (4), 179 – 194.

Author	*Planning %*	*Lit Review %*	*Design of Conceptual Framework %*	*Writing %*	*Graduate supervision %*	*Mean %*
Roorda	80	95	85	80	0	85
Gullickson	20	05	15	20	50	15

Acknowledgements

From the summit that this dissertation represents, I would like to express my gratitude to the many people who have played a vital part in this dissertation journey.

At the very start, Professor Janet Clinton was instrumental in helping me define the destination. Thank you, Janet, for your encouragement and ongoing faith in my ability to see the research through to completion. Dr. Amy Gullickson, my primary supervisor, encouraged me to explore, always with the endpoint in mind, and never let me forget that this was important work.

Through this PhD journey there have been some memorable experiences. Early on, Amy organised what turned out to be literally an earth-shattering skype conversation with Professor Michael Scriven. Michael joined the conversation from California, Amy from Melbourne and I, from a seventh-floor apartment in central Wellington. The conversation had barely begun when an earthquake measuring 6.5 on the Richter scale had me ducking for cover. Through the sound of breaking glass and sirens signalling chaos on the streets below, the conversation continued for the next hour. My family, friends and colleagues were thereafter split into one of two groups: those who, once they knew who was on the end of the skype, accepted, without further explanation, why I was keen for the conversation to continue, and those who thought I was completely bonkers to stay in the building when seemingly everyone else in the city was escaping as fast as they could.

A sincere thanks to Dr. Ghislain Arbour, my second supervisor, who always had great questions and insights, and to other members of my PhD Advisory Committee: Laureate Professor John Hattie (Chair), Professor Helen Sullivan and Brad Astbury.

Evaluation scholars from around the world have enriched this study. Thank you to Nicholas Saxton, whom I have never met but who shared a paper that led me to normative ethics, Professor Tom Schwandt and James Griffith for their interest and input into the study, and to Professor Michael Morris and Dr. David Turner who reviewed the normative ethical principles in the criteria matrix handbook. At the European Evaluation Society conference in 2016, Dr. Bianca Montrosse-Moorhead and Bob Williams were the only participants to turn up to my session. Yes, it was a

very long way to travel to speak to an audience of two but the discussion that ensued was stimulating, encouraging and memorable. Professor Donald Yarbrough, thank you for the discussion about criteria and evaluation standards. Joan Sykora, thank you for introducing me to the American Evaluation Association conferences and for being such a fabulous and generous roommate.

Closer to home, I want to express my gratitude to all who participated in the study as respondents in the online survey. A huge thank you also to Associate Professor Robin Peace, Alison Gray, Marnie Campbell and Fiona Scott-Melton for your friendship and support at critical times. My evaluation colleagues: Ruth Aston, Timoci O'Connor, Julian King, Alison Rogers and Kim Grey (post-graduate colleagues) and Heather Nunns, Robyn Bailey, Rae Torrie and Vicki Wilde (co-members of a professional development group that has continued to meet since we completed the first postgraduate diploma in evaluation at Massey University in 2008): thank you for listening, sharing, critical feedback and encouragement. Anne Markiewicz and Ian Patrick: thank you for giving me a home away from home in Melbourne, and for your friendship and occasional diversions from study.

And finally, a thank you to my daughters Jay and Alex, and a special thank you to my partner, Keran, for her unwavering love and support. Now this summit has been reached, let's get back into the mountains and climb them together.

Table of contents

ABSTRACT I

DECLARATION III

PREFACE IV

ACKNOWLEDGEMENTS V

TABLE OF CONTENTS 1

GLOSSARY 6

ABBREVIATIONS 6

PART ONE: INTRODUCTION TO THE STUDY 7

1 INTRODUCTION 8

1.1 A conundrum 8

1.2 A puzzle 9

1.3 Organisation of the thesis 10

1.4 Summary of findings 12

1.5 Chapter summary 14

2 BACKGROUND 15

2.1 Purpose 15

2.2 Introduction to evaluation 15

2.3 Values in evaluation 19

2.4 Evaluation criteria 27

2.5 Normative ethical perspectives 43

2.6 A gap and an opportunity 47

2.7 Aims of this study 48

2.8 Research questions 49

2.9 Bounding the study 49

2.10 Contributions of the study 50

2.11 Chapter summary 51

3 PROGRAM EVALUATION IN AUSTRALIA AND NEW ZEALAND 52

3.1 Purpose 52

3.2 Program evaluation in Australia 52

3.3 Program evaluation in New Zealand 55

3.4 Concluding comments 57

3.5 Chapter summary 58

4 METHODOLOGY 59

4.1 Purpose of this chapter 59

4.2 Researcher's context 59

4.3 Initial conceptualisation of the study 61
4.4 Mixed methods strategy for this study 62
4.5 Design of the study 63
4.6 Ethical considerations 68
4.7 Chapter summary 68

PART TWO: CRITERIA IN CURRENT EVALUATION PRACTICE 69

5 PRACTITIONER PERSPECTIVES ON DEVELOPING CRITERIA 70
5.1 Purpose of this chapter 70
5.2 Research objectives 70
5.3 Survey design 70
5.4 Findings 79
5.5 Theory-practice alignment 97
5.6 Chapter summary 101

6 CRITERIA IN EVALUATION REPORTS 102
6.1 Purpose of this chapter 102
6.2 Research objective 102
6.3 Findings 113
6.4 Theory-practice alignment 132
6.5 Chapter summary 136

7 SYNTHESIS OF FINDINGS FROM PHASE ONE 137
7.1 Purpose of this chapter 137
7.2 Approach to synthesis 137
7.3 Current approach to managing criteria 142
7.4 Theoretical framework of characteristics of defensible criteria 144
7.5 Chapter summary 144

PART THREE: TOWARDS SYSTEMATIC DEVELOPMENT OF CRITERIA 146

8 NORMATIVE ETHICS AND CRITERIA DEVELOPMENT 147
8.1 Purpose of this chapter 147
8.2 Rationale for a conceptual framework for identifying defensible criteria 147
8.3 Introducing the conceptual framework 148
8.4 Role of the framework in developing defensible criteria 152
8.5 Chapter summary 153

9 DEVELOPMENT OF A CRITERIA MATRIX HANDBOOK 155
9.1 Purpose of this chapter 155
9.2 Rationale for developing a criteria matrix handbook 155
9.3 Overview of the criteria matrix handbook 156
9.4 Process for developing and refining the handbook 157
9.5 Adoption of the Checklist Project process: a reflection 174
9.6 Chapter summary 175

10 PRACTITIONER ENGAGEMENT WITH NORMATIVE ETHICS 176
10.1 Purpose of this chapter 176
10.2 Research design 176
10.3 Field testers' case examples 178
10.4 Analysis of the case examples 194
10.5 Chapter summary 196
PART FOUR: DISCUSSION AND CONCLUSION 197
11 DISCUSSION 198
11.1 Research problem and overview of study 198
11.2 Phase one: approach, summary and discussion of findings 199
11.3 Phase two: approach, summary and discussion of findings 201
11.4 Contributions to knowledge 203
11.5 Limitations 205
11.6 Final reflections and opportunities for future research 207
11.7 Conclusion 210
REFERENCES 211
APPENDIX A: ONLINE SURVEY OF EVALUATORS 229
APPENDIX B: INFORMATION SHEET FOR ONLINE SURVEY 233
APPENDIX C: CODING OF CRITERIA 236
APPENDIX D: EXPERT REVIEW OF CRITERIA MATRIX 271
APPENDIX E: FEEDBACK FORM FOR CRITERIA MATRIX HANDBOOK REVIEWERS 272
APPENDIX F: CRITERIA MATRIX HANDBOOK 273

LIST OF TABLES AND FIGURES

Table 1: Comparison of two logics of valuing 23
Table 2: Sources for identifying criteria 37
Table 3: Examples of theories that align with different normative perspectives 44
Table 4: Overview of the study design 64
Table 5: Alignment of survey questions to the key research questions 72
Table 6: Sources for developing criteria as identified in the literature 73
Table 7: Rationale for the survey selection criteria 74
Table 8: Illustration of coding framework for Question 10 79
Table 9: What respondents make assessments against 80
Table 10: Additional sources identified by respondents 84
Table 11: Sources prioritised by two factors 89
Table 12: Alignment to a factor by sector/s where respondents work 91

Table 13: Convergence of themes identified in the data analysis 94
Table 14: Factors that inform which sources are used (n=129) 95
Table 15: Final sample frame for reports 103
Table 16: Sector of reports (n) 105
Table 17: Portfolio areas (n) 106
Table 18: Evaluation purpose 107
Table 19: Evaluator's workplace 107
Table 20: Example of coding approach for the document review 112
Table 21: Types of criteria evident in reports 114
Table 22: Sources that inform selection of criteria 121
Table 23: Criteria defined by two sources 123
Table 24: Criteria implicit in an evaluation of a youth program 131
Table 25: Summary of key findings from phase one 139
Table 26: Comparison of principles used in two normative frameworks 149
Table 27: Definition and focus on five normative principles 150
Table 28: Criteria for developing and reviewing the handbook 161
Table 29: Expert feedback on the normative perspectives and principles 164
Table 30: Reviewers' feedback on the handbook 169
Table 31: Criteria developed for a disability inclusion action plan 180
Table 32: Dimensions of value for a disability inclusion action plan 182
Table 33: Dimensions of value for a curriculum framework 185
Table 34: Criteria for a curriculum framework evaluation 187
Table 35: Dimensions of value for an AGYW program 191
Table 36: Criterion of relevance for the AGYW program 194

Figure 1: Structure of the thesis 11
Figure 2: Logic of evaluation 22
Figure 3: Overview of survey focus 70
Figure 4: Respondents' evaluation experience (%) 76
Figure 5: Respondents' workplace (%) 76
Figure 6: Extent to which evaluation is a focus of respondents' work (%) 77
Figure 7: Exploratory factor analysis 78
Figure 8: Extent to which respondents say they commonly use a source 82
Figure 9: Sources commonly used by respondents in two countries 83
Figure 10: Importance versus use of sources 92
Figure 11. Comparison of use and importance of cultural expertise 94

Figure 12. Hierarchical approach in an explicit criterion (NZ70) 115
Figure 13. Hierarchical approach in an implicit criterion (AU20) 116
Figure 14: Sources used to inform criteria 119
Figure 15: Sources that link to, or support criteria selection (n) 122
Figure 16: Most common criteria identified in the reports 126
Figure 17: Conceptual framework for identifying dimensions of value 151
Figure 18: Description of an argument to develop defensible criteria 153
Figure 19: Review of the criteria matrix handbook 168
Figure 20: Revised conceptual framework of defensible criteria 208

Glossary

Criterion / criteria	Definition/s or description/s of quality or success against which an evaluative judgement is made.
Evaluand	This refers to the program that is, or has been, evaluated
Evaluation commissioner	Person(s) whose primary responsibility it is to contract for, oversee and coordinate a program evaluation
Hapu	Kinship group of multiple whānau
Pākehā	New Zealander of European descent
Normative perspectives	These are ways of understanding concepts such as 'good' or 'bad'; 'right' or 'wrong' that broadly align with consequentialist, deontological or virtue-based theories.
Program evaluation	Evaluation of public sector policies and programs
Tikanga	Māori practices or processes
Whakapapa	Genealogy
Whānau	Family

Abbreviations

AEA	American Evaluation Association
AES	Australasian Evaluation Society
ANZEA	Aotearoa New Zealand Evaluation Association
EFA	Exploratory factor analysis
FaHCSIA	Family, Housing, Community Services and Indigenous Affairs, Australian precursor to the Department of Social Services (DSS)
MPI	Ministry for Primary Industries (New Zealand)
NPM	New public management

PART ONE: INTRODUCTION TO THE STUDY

> '…best practices can't be developed separately because evaluation theory and practice are interdependent: each one learns from the other and, in that learning process, both are inspired to stretch, to bend a little, to grow. Further, their mutual dependence endows them both with legitimacy: theory protects practice from singularity and anecdotalism, and practice protects theory from abstraction. Their relationship is a mediation between principle and context, which lends breadth, depth and realism to our work. (Chelimsky 2012, p.1)

Part One comprises the first four chapters of the thesis, which establish the context for the study. Chapter 1 introduces the research topic and provides an outline of the thesis. Chapter 2 discusses key spheres of theory and practice that are within the scope of the study and identifies gaps in knowledge that lead to the research aim and questions. Chapter 3 includes contextual information about program evaluation in Australia and New Zealand. The final chapter in Part One discusses the research methodology (Chapter 4).

1 Introduction

The researcher, like many other evaluators (Christie and Rose 2003; Stevahn, King, Ghere and Minnema 2005) fell into evaluation in 1998 with no theoretical knowledge of the field. Her approach, as an evaluation analyst in a New Zealand public sector agency, was eclectic and pragmatic. She drew on her background as a social science researcher and ideas gained from seminars and workshops held by visiting overseas evaluation experts and occasional attendance at annual evaluation conferences. Concepts that had a good fit with the public sector context were picked up, for example program logic, most significant change and utilisation-focused evaluation. Close collaboration with policy and program staff was expected by senior public sector management and the information requirements of key stakeholders within government were considered paramount.

After seven years of evaluation practice she became a student in the newly formed postgraduate diploma in social sector evaluation research at Massey University. It was here that she read her first text on evaluation theory. That initial study sensitised her to the difference between evaluation and research and in time she encountered a conundrum, and then a puzzle that ultimately led to the central question around which this study is focused.

1.1 A conundrum

On the heels of the postgraduate qualification in evaluation came an awareness of two discourses concerning evaluation quality swirling around New Zealand's public sector: one advocating for value-free evaluation, the other for value-engaged evaluation. These discourses played out as follows. At various public sector forums, the newly appointed Chief Science Advisor to the Prime Minister began advocating for the 'use of scientific approaches to monitoring and evaluating policy initiatives' (Gluckman 2011, p. 3), 'the … need for the collection and interpretation of data to be value free' (Gluckman 2012, p. 2) , and 'objective evaluation' (Gluckman 2013a, p. 7). Public sector colleagues reported that evaluations they had conducted were under greater scrutiny; fewer were being published and experimental designs were viewed as the preferred approach. Around the same time, evaluation scholar Jane Davidson returned from the US to work in New Zealand and urged evaluators to be explicit about values. How was the public sector evaluator to manage this

conundrum? Was it possible to explicitly engage with values and still produce objective, high quality, trustworthy program evaluations?

1.2 A puzzle

In the same year the researcher started her postgraduate studies, Davidson (2005) published a text that ultimately resulted in evaluation rubrics being pioneered in New Zealand and Australian program evaluation. Davidson introduced rubrics as an evaluation-specific method to support evaluators in 'the process of combining evidence with values to determine merit, worth, or significance' (Martens 2018, p.25). Evaluation rubrics generally comprise three components: criteria, performance standards and descriptors or 'examples of what performance looks like for each criterion at each performance level' (Martens 2018, p.26). Those who promote the use of evaluation rubrics argue they can assist in making evaluative reasoning explicit (Davidson 2013; Martens 2018).

Herein lay the puzzle. Evaluative reasoning is concerned with reaching well supported judgements about the value of something. To reach such a judgement involves defining criteria, which are the aspects of a program that evaluators look at to determine whether performance is good or not, as well as standards that define 'how good is good'. While this process can be made explicit through a tool such as rubrics, how can the evaluator reach a defensible evaluative judgement if they do not also justify their criteria? And if explicit criteria are not sufficient to reach defensible judgements, then what else might evaluation practitioners need to consider in support of evaluative reasoning?

In the public sector context, evaluation is concerned with identifying and assessing the quality of the programs or policies. Such judgements rest on particular conceptualisations of 'good', 'bad', 'success', 'non-success' and 'quality', which are often contested. How can evaluators working in the public sector space negotiate multiple stakeholders and interests to arrive at criteria that are defensible?

Many scholars including House (1980a); Nunns, Peace, and Witten (2015); Schwandt (2008) and Scriven (1996) note the centrality of criteria to quality evaluation. However, there is little practical guidance in the literature to support evaluation practitioners to engage with normative perspectives in ways that address Shadish et al's (1991) concerns, including that of promoting certain

conceptions of right or good over others. More generally, the development of criteria is a topic that has received little attention within the discipline.

In the interests of bringing theory and practice together, the discipline needs to examine and reflect on current practice, and develop and test approaches to identifying and selecting criteria that serve the interests and needs of public sector decision-makers while avoiding the risk of being co-opted into a given policy agenda (Greene 2002; Schwandt 1992). Evaluation findings can have far-reaching consequences for program recipients and others (including the environment) who may be affected by public sector policies and programs. Evaluations that do not include all relevant values are at risk of producing invalid judgements (Davidson 2005). Thus, the inclusion of explicit, comprehensive and defensible criteria is fundamental to sound evaluative reasoning and the delivery of quality, robust evaluations.

1.3 Organisation of the thesis

The thesis comprises 11 chapters with accompanying appendices. The chapters are organised as follows.

1.3.1 Part One: Introduction to the study

The introduction to the study comprises four chapters. The current chapter (Chapter 1) introduces the research topic and the organisation of the thesis. Chapter 2 introduces literature of relevance to the study. Gaps in the literature are identified, which lead to the research aims and research questions. The final section of this chapter defines the study scope. Chapter 3 outlines key historical, political events that inform how evaluation is conceptualised and practised in Australia and New Zealand. The methodology and the research design that have guided this study are included in Chapter 4. The study is conceptualised as a two-phase emergent mixed methods design, with the second phase responding to what was learned in the first phase of the study.

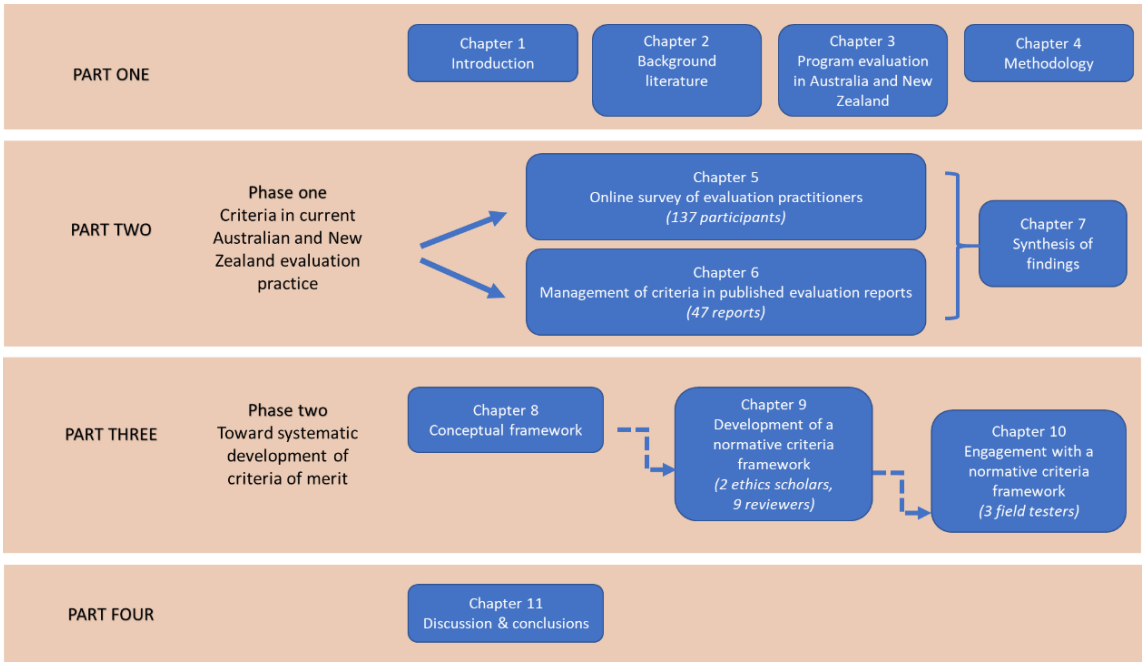


Figure 1: Structure of the thesis

1.3.2 Part Two: Criteria in current evaluation practice

Part Two introduces phase one of the study (Figure 1). This phase examines the management of criteria in Australian and New Zealand program evaluation practice from two perspectives. First, in Chapter 5, criteria development is examined from the perspective of 137 Australian and New Zealand evaluation practitioners who conduct or manage public sector evaluations. The objective is to explore how criteria are developed in practice as well as study practitioners' views on the characteristics of defensible evaluation criteria. Data was collected via an online survey.

A second perspective (the focus of Chapter 6) describes the management of criteria as evident in a sample of 47 published Australian and New Zealand program evaluation reports. The objective of this component of the study is to describe how criteria are established and justified in the context of specific program evaluations, and to identify factors that support the defensibility of criteria. The reports were submitted to a review that included both content and thematic analyses. Chapter 1 is a synthesis of findings from the online survey and document review that then establishes a rationale for Part Three.

1.3.3 Part Three: Towards systematic development of criteria

Part Three introduces phase two of the study. This phase explores how engagement with normative perspectives might support the development of a comprehensive set of criteria and thereby address some of the limitations and challenges identified with current management of criteria. Phase two proceeds as a stepped process (as illustrated in Figure 1). Chapter 8 begins with an introduction to normative ethical perspectives and a rationale for explicitly considering them when identifying dimensions of value for program evaluations. Next, a conceptual framework is presented that also draws on components of prescriptive and descriptive valuing (Shadish et al. 1991).

To explore the framework's potential for supporting evaluation practitioners to systematically identify dimensions of value and then develop criteria of merit, the researcher proposed to field test the framework. A criteria matrix handbook was developed to support practitioners' engagement with the framework, and Chapter 9 documents the systematic review of the information that is in this document. The framework was then field tested by four evaluation practitioners with an evaluation each was engaged with at the time. The results of the initial field testing are presented in Chapter 10.

1.3.4 Part Four: Discussion and conclusions

Chapter 11 includes the overall discussion, consideration of the contributions the study makes to knowledge and to evaluation practice. The methodological strengths and limitations of the research are then discussed, followed by opportunities for future research, and the research conclusions.

1.4 Summary of findings

The research contributes to the field of evaluation in a number of ways. The first is the study itself – it includes an empirical description and examination of how criteria are developed in Australian and New Zealand program evaluation practice. The findings are based on a survey of 137 evaluators and a review of published evaluation reports and provide an evidence-based platform from which to build a theory of evaluation and assist with advancing more systematic and contextually relevant evaluation practice.

Second, the study adds to existing evidence that evaluative reasoning is weak in program evaluation. Evaluative judgements exist in evaluation reports, but the criteria on which these judgements are made are not always explicit, comprehensive or justified. A failure to attend to the value premises against which judgements are made weakens the evaluative reasoning and thus the strength of the judgement.

Third, criteria in program evaluations do not always include the value perspectives of all legitimate stakeholders. An analysis of reports where indigenous people were a primary focus of the evaluand[1] suggested little evidence of indigenous values being incorporated into the way value was conceptualised.

Fourth, explicit criteria are necessary to support evaluative reasoning but not sufficient. Evaluative judgements must be based on all relevant dimensions of value, that is, criteria should be comprehensive. Defensible criteria are also based on authoritative sources, as relevant to the evaluand and evaluation context.

Fifth, a theory-informed conceptual framework is presented, bringing together both prescriptive and descriptive valuing traditions. Its aim is to support evaluation practitioners to systematically develop defensible criteria. The framework provides a way for practitioners to engage with prescriptive valuing – normative ethical perspectives – that deal with conceptions of 'good', 'bad' and so forth. On the descriptive side, the framework requires the evaluator to identify dimensions of value for all groups with a legitimate interest in a program. The framework seeks to identify a comprehensive set of dimensions of value through consideration of three normative perspectives – consequentialist, deontological and virtue-based. It is hoped that evaluators explore the applicability of the framework in different contexts and for different purposes, that they assess its strengths and weaknesses, and share their experiences so that the conceptual framework can be improved.

Sixth, the criteria matrix demonstrates it is possible to practically engage with normative ethics to develop criteria. It provides one approach for evaluators to systematically surface the value perspectives that inform criteria. The case examples produced by the field testers provide proof of concept and indicate the

[1] Evaluand is a generic term for whatever object is being evaluated. In this study, the object is a government-funded program.

framework can be used in different social program contexts as a prospective or retrospective tool. In two of the three case examples, criteria were identified that might otherwise have been ignored. Systematic consideration of the three normative perspectives is a notable improvement over what many evaluators do in their evaluations, as evidenced in the review of evaluation reports. Further testing and exploration of the criteria matrix tool will no doubt result in improvements and potential hybrid approaches for identifying criteria of merit.

Seventh, the development of a criteria matrix handbook, reviewed by ethics' scholars and experienced evaluators in Australia and New Zealand, provides a helpful resource to assist evaluators to engage in this challenging task.

Eighth, a strength of the framework is its flexibility. It can be used to inform the focus of an evaluation, with criteria identified at the outset that then define the evaluation questions. Alternatively, it can be used to check if existing criteria are comprehensive and justified. It can be undertaken as a desktop-based exercise, with the evaluator drawing on existing documents and literature, or it can support a participatory evaluation approach, with the evaluator facilitating stakeholders to identify dimensions of value.

1.5 Chapter summary

This chapter has introduced the research topic that is the focus of this thesis. The genesis for the research topic emerged as the researcher grappled with two diverse discourses concerning New Zealand public sector evaluation: one advocating for value-free and the other for values-explicit evaluation, and both with the end goal of producing quality information. Related to this conundrum, the researcher questioned the requirements of criteria to support defensible evaluative judgements. Evaluation rubrics were heralded as an explicit approach to support defensible evaluative reasoning, yet 'explicitness' seemed insufficient to justify criteria. In Chapter 2 these issues are explored through a review of relevant literature.

2 Background

2.1 Purpose

This chapter examines key spheres of theory and empirical research that are within the scope of this study. Section 2.2 introduces evaluation and its sub-field, program evaluation. Section 2.3 identifies that concepts of valuing and values are central to evaluation and yet have received relatively little attention in the literature. The historical reasons for this are considered. The role of values, and specifically criteria, in evaluative reasoning is then introduced. This leads to Section 2.4, which examines literature on evaluation criteria. Three characteristics of a criterion that are pertinent to the development of defensible criteria are identified. These characteristics provide a conceptual framework for both phase one and two of the study. Section 2.5 provides an overview of normative ethical perspectives, which provide a theoretical foundation for a conceptual framework introduced in phase two of the study. From the review of the literature, gaps in knowledge are identified, along with an opportunity to investigate the characteristics of defensible criteria. These are discussed in Section 2.6. In Sections 2.7 and 2.8 the research aims and research questions are outlined. The boundaries of the study are defined in Section 2.9. Next, in Section 2.10, the expected contributions of the study to knowledge and evaluation practice are set out. The chapter concludes with a summary (Section 2.11).

2.2 Introduction to evaluation

Evaluation has a long history, very likely linked to our ancestors' earliest attempts to survive and thrive on this planet. Learning what foods are safe to eat, which concoctions can cure or prevent different ailments, identifying the most effective ways to hunt animals and fish, all require evaluative skills, that is, performing activities, then assessing their results, and determining their effectiveness in responding to human needs and wants.

There is also historical documentation that suggests more formal evaluation, that is, explicit and deliberate evaluative inquiry, has been around for thousands of years. Shadish and Luellen (2005) note, for example, that personnel evaluation has been practised in China since about 2200 B.C. They also refer to a Bible chapter in

the Old Testament (Daniel, Chapter 1) that describes a quasi-experiment evaluating two types of diet[2].

Formal evaluation has grown exponentially since the late 1970s, with many scholars, e.g. Chouinard and Milley (2015); Crane (1988); Davidson (2014b, 2014a); Fournier (1995); Montrosse-Moorhead, Bellara, and Gambino (2017); Scriven (1991); Scriven (2007) now referring to it as a discipline or profession, with a range of theories, approaches and methods to support robust and valid investigations.

Evaluation has been defined by Scriven (1991) as:

> the process of determining the **merit**, **worth**, or **value** of something (p. 139, bold text as in original).

This definition of evaluation is purposefully broad, intended to encompass evaluation of any type of object. Scriven's understanding of evaluation also aligns with the dictionary definition, where the root word of evaluation (value) has both French and Latin origins and refers to the 'worth or quality of something as measured by a standard of equivalence'[3].

It is important to reflect on the origins of Scriven's conceptualisation of evaluation as this provides context to value terms such as 'merit'[4] and 'worth'[5] that many of us take for granted in evaluation as practised in 21st century. Evaluation scholar Dan Stufflebeam has described how Scriven was asked to address a Social Science Consortium in the mid-1960s (Stufflebeam 2013). The Consortium was seeking more appropriate ways to design and conduct evaluation. At the time, evaluation

[2] Over a 10-day period Daniel and three friends were said to have consumed only vegetables, while those around them ate their normal diet which probably consisted of meat, fish and wine. The results were then compared with a view to assessing the health benefits associated with an alternative diet.

[3] OED accessed online March 2016

[4] When an evaluation focuses on 'merit' (rather than program goals or objectives) it is concerned with the *intrinsic* quality of something (Lincoln and Guba 1979). A typical question about merit asks: 'Does the program do well what it is intended to do?' (Stufflebeam and Shinkfield 2007, p. 10). Take, for example, a new program introduced in primary schools to teach children protective behaviours so they can feel strong and stay safe in their community. The evaluation might include a focus on the quality of the teaching and the appropriateness of the program content for this age group. The value terms 'quality' and 'appropriateness' are a part of how 'merit' is being defined in this evaluation.

[5] The term 'worth' refers to a program's quality and importance within a *given context* (Davidson 2005; Lincoln and Guba 1979; Stufflebeam and Shinkfield 2007). A typical question addressing worth might be: 'Is this program high quality and also addressing a need?' The protective behaviours program may have merit, but perhaps the school already teaches children the necessary skills in their existing curriculum.

was defined in predominantly value-free terms, referenced to the assessment of goals. Scriven, according to Stufflebeam (2013) argued that the achievement of goals:

> was not necessarily the same as producing meritorious outcomes, because the developer's goals could be biased, corrupt, or too narrow to reflect the full range of important outcomes (p. 78).

Rather, evaluation's constant goal was to assess something's value.

Scriven's contribution, specifically drawing attention to the value-based nature of evaluation, is evident in many scholars' definitions. For example, professional evaluation is understood to be a systematic investigation against one or more values such as merit/quality, worth or significance (Alkin, Vo, and Christie 2012; Fournier 1995b; Hertting and Vedung 2012; House 1995; Hurteau, Houle, and Mongiat 2009; Kirkhart and Ruffolo 1993; Lincoln and Guba 1979; Mark, Henry, and Julnes 2000; Schwandt 1997a; Scriven1995; Vestman and Conner 2006).

In evaluation, the term 'systematic' differentiates informal from more formal and professional types of inquiry such as program evaluation. Evaluations done systematically are 'relevant, designed and executed to control bias, kept consistent with appropriate professional standards, and otherwise made useful and defensible' (Stufflebeam and Shinkfield 2007, p. 28). Weiss (1999) refers to evaluation as 'the systematic assessment of the operations and/or the outcomes of a program or policy compared to a set of explicit or implicit standards' (p. 4). The Australasian Evaluation Society refers to evaluation as a type of critical thinking that 'involves systematically collecting, analyzing and using information to enhance learning and decision making' (Australian Evaluation Society n.d.). Likewise, the Program Evaluation Standards, developed by the Joint Committee on Standards for Educational Evaluation (JCSEE) define evaluation as 'the systematic investigation of the quality of programs…' (Yarbrough et al. 2011, p. xxv).

In Australia, the AES Guidelines for the Ethical Conduct of Evaluation refer to evaluation as the 'systematic collection and analysis of information to make judgements, usually about the effectiveness, efficiency and/or appropriateness of an activity' (Australasian Evaluation Society 2013). This definition aligns with the Australian Government's outcomes and outputs framework, developed in the late 1980s (Department of Finance and Administration 2000), suggesting that, in

Australia at any rate, professional evaluation has traditionally been closely linked with the agendas of its government. Here, merit has a relatively narrow focus on 'effectiveness' and 'appropriateness'; likewise, worth is equated with 'efficiency'.

In contrast, New Zealand's professional evaluation association, ANZEA, offers two broad definitions of evaluation, the first being that of the North American Program Evaluation Standards, and the second based on an ANZEA Evaluation 101 Workshop held in 2014, in which evaluation is defined as: 'The process and product of reaching robust, defensible evaluative conclusions logically derived from evidence ... evaluation always focuses particularly and intentionally on the quality, value and importance of things' (Superu and Aotearoa New Zealand Evaluation Association 2015, p. 15).

The term 'logically derived' in ANZEA's definition alludes to the reasoning process that underpins all evaluation (discussed further in Section 2.3.3). Greene (2011) succinctly sums up this process by suggesting the primary purpose of evaluation is to: '…render judgments of merit and worth that are grounded in defensible empirical evidence and argument and that are anchored in chosen values' (p. 85).

The definition of evaluation adopted in this study draws on two concepts outlined in the discussion above that are considered essential for program evaluation. These are that the evaluative inquiry (i) results in judgements about merit or worth and (ii) involves a reasoning process. Hence, evaluation is defined in this study as the assessment of a program's merit or worth. Evaluative judgements are grounded in empirical evidence and argument and anchored in defensible values, which are referred to as criteria.

2.2.1 Program evaluation

Scriven (1991) has referred to the 'big six' fields in which professional evaluation occurs: product, personnel, performance, proposal, policy and program evaluation. In government, the fields of policy and program evaluation are closely connected. The 'public' component of public policy refers to the policies of elected governments. Policies set the direction for how resources in a sector, for example education and environmental, will be allocated, and what aspects will receive priority. A public program is a particular enactment of a policy. It is a set of activities funded or administered by government that seeks to achieve particular outcomes for recipients and other impactees (Fitzpatrick, Sanders and Worthen

2011). In this study the term 'program' refers to both government policies and programs. The terms 'program evaluation' and 'evaluation' are used interchangeably in this study to refer to evaluations of policies and programs funded partially or fully by government.

The researcher focused on program evaluation in Australia and New Zealand. An overview of how program evaluation is conceptualised and practiced in these two countries is included in Chapter 3.

2.3 Values in evaluation

2.3.1 Purpose

This section of the literature review includes an introduction to values in evaluation. Its purpose is to understand why values and valuing have received so little attention to date. It outlines the origins of a so-called fact-value dichotomy that has had long-standing consequences for evaluation. It then introduces the literature on evaluative reasoning and the logical process by which one can reason from values, also referred to as criteria, to reach an evaluative judgement.

2.3.2 Historical context: a fact-value dichotomy

For much of the discipline's history, discussions about evaluation quality have focused on methods for collecting data. This emphasis has been demonstrated in Kallemeyn, Hall, Friche and McReynolds (2015) study that examined articles published in the *American Journal of Evaluation* (AJE) and *Evaluation*, a journal supported by the European Evaluation Society. The study examined the number of times the terms 'method', 'evaluation use' and 'valuing' were used in 171 papers in the two journals during the period 2008–2011. The findings showed that both journals emphasised methods, with valuing receiving the least attention by scholars.

A significant contributor is a lingering value-neutral view of the social sciences that has persisted in separating facts from values; the former being viewed as objective; the latter as subjective and not able to be validated or empirically confirmed. Given the fundamental role of values to evaluation, it is important to understand this fact-value dichotomy and how it came about. In doing so, the researcher reconstructs the arguments of Hilary Putman, a Harvard philosopher who gave a series of lectures at the Northwestern University School of Law in 2000, referred to as the

Rosenthal Lectures that were subsequently incorporated into a text 'The Collapse of the Fact/Value Dichotomy' (Putnam 2002). It should be emphasised that Putnam was not alone in critiquing the positivists' arguments on truth. Others (e.g. House and Howe 1999; Schwandt 1989; Scriven 1986) have provided similar accounts of the dichotomy and its legacy on scientific social inquiry.

2.3.2.1 *Origins of a value-free social science discourse*

Putnam begins by noting the fact-value dichotomy rested on another philosophical dichotomy in which statements were classified as either analytic or synthetic. Analytic statements were either definitional truths, for example 'all bachelors are unmarried', true by virtue of their meaning. Synthetic statements were descriptive truths, for example 'all robins have feathers', true because they are descriptive facts (by virtue of observation; one can see that a robin has feathers).

In the 1950s, philosopher and mathematical logician, Willard van Orman Quine, was instrumental in demonstrating that many statements about the world sat outside this strict analytic-synthetic dichotomy. The example given is that of mathematics which Quine argued could not be known through empirical methods and neither was it a true by virtue of its meaning. Putnam argued that the analytic-synthetic distinction was logically required to support a second dichotomy referred to as the fact-value distinction. Quine's successful critique of the first analytic-synthetic dichotomy was significant in collapsing the second dichotomy.

For most of the 20th century the prevalent view was that statements of fact were viewed as being 'objectively true' and able to be 'objectively warranted' (Putnam 2002). Values were viewed as subjective, not able to be scientifically proven and therefore not able to be systematically incorporated into scientific inquiry. This notion was advanced in particular by Max Weber's value-free doctrine in which it was argued that judgements of fact are one thing; judgements of value another (Scriven 1991).

Several philosophers have since demonstrated that facts and values are often entangled. House (2001) and House and Howe (1999) suggest most statements exist on a continuum; some as pure fact, for example 'all adult humans start off as embryos' and some are based on personal taste, for example 'Fiji is a better holiday destination than Bali'. Along the continuum are many other statements that are a blend of fact and value. For example, a statement such as 'that product is too

expensive' requires the evaluator to define the value term (expensive) and collect factual information to determine whether the particular product in question does indeed cost too much. Likewise, Putnam has used the example of a 'cruel' teacher. In declaring his son's teacher to be a very cruel person, he argued he was not simply describing the person, but also evaluating him. The term 'cruel':

> simply ignores the supposed fact/value dichotomy and cheerfully allows itself to be used sometimes for a normative purpose and sometimes as a descriptive term (Putnam 2003, p. 35).

2.3.2.2 *Consequences of the fact-value dichotomy for evaluation*

The consequences of positivists' attempts to define truth statements (even though short lived) has resulted in a lingering view in some circles that values have no place in scientific inquiry (Crane 1989; Schwandt 1989). The result is that some evaluators have adopted a view that the role of evaluation is to 'investigate questions of what is, not what ought to be' (Schwandt 1989, p. 12). The following comment, from an AEA member describing their approach to evaluation, exemplifies such a position:

> I typically collect data, write-up results, and provide interpretations and recommendations, and *then let the client decide what the criteria should be* (Ozeki et al., 2019, p. 3, italics in original).

There are at least two consequences of such a position for the development of criteria. The first is that if the evaluator limits their role to description of data, then they are logically not engaged in evaluation. In effect, the so-called evaluator hands over the central task of evaluation to a client who may well have a vested interest in the program's success or non-success. A second, follow-on, consequence is that the client is at risk of making evaluative judgements based on an incomplete set of values (criteria).

2.3.3 Evaluative reasoning

As a generic term reasoning 'involves inferring a "conclusion" (claim assertion) from one or more premises' (Beyer 1985, p. 17). In relation to evaluation, the reasoning process involves both value and factual premises that culminate in an evaluative claim. This is articulated in what has been referred to as a general logic of evaluation (Fournier 1993; Schwandt 2018b; Scriven 1994, 1995, 1996, 2002,

2007b, 2012). The logic of evaluation comprises five steps, as follows. Step one involves identifying criteria, the dimensions of value against which the goodness of the evaluand will be assessed. In the language of argument, this provides the first of two value premises. Step two is the standards, which focus on levels of performance for each dimension of value, for example 'excellent' versus 'poor'. Standards provide the second of two value premises. Step three involves measuring performance (thus comprising factual premises) and then in Step four the results are compared against the standards (to make an evaluative claim). Step five is the final step of the logic, where the performance on all dimensions is synthesised into an overall evaluative judgement, as indicated in Figure 2.

1. Establish criteria: on which dimensions of value must the evaluand do well? (Value premise)
2. Set performance standards: how well should the evaluand perform? (Value premise)
3. Measure performance (Factual premise)
4. Then compare with the standards: how well did the evaluand perform? (Evaluative claim)
5. Synthesis and integration into a judgement: what is the merit (intrinsic value), worth (value to a group) or significance (overall impact) of the evaluand? (Evaluative claim)

Figure 2: Logic of evaluation

In the evaluation literature the general logic is most often attributed to Professor Michael Scriven. However, the logic can also be found in the work of other philosophers, including that of Rescher (1969) and Taylor (1961). By example, Taylor's (1961) logic is conveyed as five steps as shown in Table 1. Scriven's articulation of the logic follows the same sequence as Taylor's, although it misses step 3, in which the parameters for comparison are set. Both logics serve to demonstrate the intertwining of values and evidence to reach an evaluative judgement. Most importantly, without criteria and standards, one can only ever arrive at descriptive or explanatory claims (Smith 1987).

Table 1: Comparison of two logics of valuing

Taylor's 'logic of the evaluation process' (Taylor 1961, pp. 9–10)	Scriven's logic of evaluation (Scriven 1991)
1. Adoption of a standard or set of standards for evaluating (ranking or grading) the evaluatum[6]. These are the 'good making' characteristics. (For a family car, one criterion is seating capacity.)	Identifying criteria of merit
2. Operational clarification of the standards. (A car with five seats is the minimum; a car with less is unacceptable.)	Setting standards for each criterion
3. Specifications of the class of comparison. (Cars with at least five seats, within a certain price range.)	
4. Determining the good-making and bad-making characteristics of the evaluatum. (Collection of data on family-sized cars within a certain price range.)	Collecting data
5. Deducing, from 2 and 4, the degree to which evaluatum on the whole fulfils or fails to fulfil the standards. (In the case of value rankings this would ideally be done for every member of the class of comparison. The members would then be ranked in order according to the varying degrees to which they fulfilled or failed to fulfil the standards.). (The car that meets or exceeds the standards is judged to be the ideal candidate.)	Deducing, from 2 and 4, to arrive at an evaluative judgement.

With regard to evaluation, Shadish et al. (1991) refer to the logic as a persuasive 'meta theory', noting:

> In its bare-bones form, it is difficult to think of a value statement that could not be subjected to it. To reject the logic of valuing is to reject rational analysis of values entirely (p. 95).

Evaluations can result in different types of evaluative claims, of which two are introduced here. The first is a claim of personal preference and is generally associated with informal or everyday evaluation. Declaring a culinary dish to be 'as good as the one your grandmother made' sits squarely in this category. Subjective claims such as this example are made against individual taste, preference and 'wants', and are difficult to justify (Scriven 2007b, p.10).

[6] Taylor uses the term 'evaluatum' to refer to the evaluand, or object being evaluated.

The second type of claim differs from personal preferences in that it is possible to justify the claim through the identification and selection of all relevant values. Scriven calls this second type of claim 'essential', noting:

> When we talk of a change as 'really significant' or a doctoral thesis as 'truly excellent', we are stressing that a truly careful evaluation will reveal the result we are claiming. This sense of value is the one that professional evaluation seeks to uncover, and it is the one that evaluation as a discipline is all about… (Scriven 2007b, p. 10).

Essential evaluative claims are the type that professional evaluators strive to produce. To avoid accusations of personal bias or subjectivity, evaluators must pay close attention to establishing criteria and standards, as it is against these that evaluative conclusions, or claims about merit, worth or significance, will be made. In the public sector, the focus of this study, defensible evaluative conclusions are particularly important as the stakes are likely to be higher. Public funding is involved, and policies and programs have the potential to impact on a great number of people, including those most vulnerable in society. The high stake nature of public sector programs serves to emphasise the importance of well-reasoned evaluation that produces credible and defensible evaluative judgments (Chelimsky 1995; Ryan 2003). Specifically, the identification and management of criteria and standards becomes critical as evaluators engage with different interests and understanding of what matters in a specific context.

Nunns et al. (2015) demonstrate how the logic of evaluation can be conceptualised as an argument structure referred to as evaluative reasoning, and comprising five elements: (i) evaluation objectives and questions; (ii) criteria; (iii) standards; (iv) a warranted argument; and (v) evaluative conclusion/judgement.[7] In their conceptualisation of evaluative reasoning, criteria 'provide the grounds on which the evaluator reasons towards an evaluative conclusion/judgment' (Nunns et al. 2015, p. 142).

The element of a warranted argument in evaluation is not new. House (1977) discusses evaluation as a type of argument based on informal logic, contending that evaluators are rarely, if ever, able to provide definitive proof. Rather, the

[7] It is not clear why 'evidence' is not included in this conceptual framing of evaluative reasoning.

evaluator's role is to provide 'the credible, the plausible, and the probable' through setting out a convincing argument (House 1977, p. 5). However, he does not provide detail on the elements required to provide such an argument, other than to stress the importance of context and audience. Fournier (1993) also refers to the role of context in her concept of 'working logics'. She argues that evaluators must first have a clear understanding of the evaluand in order to know what good/bad, success or non-success might look like. An evaluation of a product, for instance a family car, might include criteria concerning its functionality, such as the number of people it can seat. However, functionality may be less relevant in an evaluation of a 'quit smoking' program. Alternatively, the term 'functionality' may be defined differently depending on the evaluand.

Fournier (1993) also describes the argument structure that underpins the logic of evaluation. Drawing on the work of Toulmin, she identifies six features that work together to justify claims in an evaluation. They include:

1. Claims – about merit, worth or significance that are to be taken as acceptable and legitimate.
2. Evidence – comprises the grounds and provides the factual premises on which the rest of the argument is based.
3. Warrants – link the evidence to the claim by appealing to some authority that is context dependent.
4. Backings – give additional support to the warrant.
5. Exceptions – the conditions under which the warrant would not hold.
6. Qualifiers – indicators of the claim's strengths.

Fournier is concerned primarily with the argument structure for the evaluation overall, that is, the reasoning that occurs at a high level to reach a defensible evaluative conclusion. This thesis proposes that evaluations comprise many sub-arguments, one of which is an argument that justifies the criteria. Therefore, each criterion would need its own supporting argument. The following example of

criteria proposed for an evaluation of a STEM[8] program, taken from Greene et al. (2006), explains this point.

In it, the independent evaluators describe their approach to identifying and justifying criteria using a hypothetical STEM program. Criteria are implicit in evaluation questions and address three areas: (i) the contextual relevance of the program design, (ii) the quality of the program in terms of its science, pedagogy and equity interrelationships, and (iii) how the program advances the interests of traditionally underserved groups. The client has no problem with the first two areas of their inquiry but challenge a focus on equity and social justice in their proposed assessment. The evaluators respond by arguing:

> … these are essential dimensions of all STEM educational programs that have been given little critical attention in the MUR context to date, that the diversity of the program remains low, and that other elements of the program will also be carefully and thoughtfully assessed (Greene et al. 2006, p. 66).

In this respect the evaluators make two claims to support the inclusion of criteria on normative ethical principles of equity and social justice. The first is that these values are part of the definition of this type of program. Their claim serves to justify a focus on these values. Their second claim is that this is an aspect of the program that needs attention. The evaluators support this claim with evidence on current diversity within the program. This evidence serves to both justify the second claim and provide additional weight to the first. The evaluators' argument puts them in a stronger position to include criteria that the client seems reluctant to include. The purpose of this illustration is to demonstrate that the defensibility of criteria can be strengthened if a convincing argument is provided for the criteria included in an evaluation.

The need to attend to argumentation in evaluation is recognised in professional evaluation standards. Both the US Program Evaluation Standards and the Aotearoa New Zealand evaluation standards[9] align explicit evaluative reasoning with the

[8] STEM stands for science, technology, engineering and math.

[9] The AES does not have its own set of evaluation standards.

concept of accuracy and trustworthiness. For example, the New Zealand evaluation standards note:

> Evaluative reasoning clearly links results to the information and analyses (the evidence), the evaluation criteria of quality (merit, worth, importance and/or significance), and any appropriate warrants (ANZEA and Superu 2014, p. 22).

2.4 Evaluation criteria

2.4.1 Purpose

This section of the literature introduces evaluation criteria, and characteristics of a criterion that are pertinent to the development of defensible criteria are identified. These characteristics provide a conceptual framework for both phase one and two of the study. Current debates about how evaluators can best approach the development of criteria are then discussed.

2.4.2 Introduction to criteria

In evaluation, scholars use the term 'criteria' to refer to the values against which judgements are made. For example, Fitzpatrick et al. (2004) refer to:

> The inclusion, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit) in relation to those criteria (p. 5).

The Oxford English Dictionary (OED)[10] defines a criterion (the singular of criteria) as '…a distinguishing mark or characteristic attaching to a thing, by which it can be judged or estimated'.

Given its centrality to evaluation it is somewhat surprising that so few evaluation scholars and texts provide a definition of 'criteria'. Where one is provided, most reference Jane Davidson, who defines criteria as:

> the values (i.e. principles, attributes or qualities held to be intrinsically good, desirable, important and of general worth) which will be used in an evaluation to judge the merit of an intervention. (Davidson 2013, p.i)

[10] OED, accessed online May 2018

In both these definitions, the terms 'properties', 'attributes', 'qualities', 'distinguishing mark' and 'characteristics' are synonyms for a criterion (singular) or criteria (plural).

The terms 'criteria' and 'standards' are sometimes used interchangeably in academic literature (e.g. Weiss 1998). This is also the case in the OED, where an alternative definition of criterion is that of 'a test, principle, rule, canon, or standard'.

However, in this study, and in evaluation generally, it is useful to consider criteria and standards as separate aspects of the valuing process. Criteria define the dimensions of quality and goodness as relevant for a program and its context. Evaluators must be able to identify and justify all relevant criteria, as they provide one of two value premises against which a defensible evaluative claim is made. In evaluation, standards outline the level of performance for a criterion, and provide the second value premise against which evaluative claims are made. Standards may be expressed quantitatively, for example, a number grade or rank, or qualitatively, such as poor to excellent (Davidson 2005). The focus of this thesis is on criteria only.

2.4.2.1 *Criteria as values*

In evaluation the term criterion (singular) or criteria (plural) is used to refer to the values that bear on an evaluand, and that will be used to estimate the quality of something. The diversity and all-pervasive nature of values is captured in the following sample of terms collected during a review of the literature for this study: desires, wants, satisfaction, frustration, preferences, attributes, utility, means, conditions, correctness and incorrectness, integration and disintegration, health, survival, adjustability, freedom, justice, law, duty, virtues, progress, beauty, truth, respect for difference, kindness. As indicated by the above sample, values can have positive as well as negative connotations.

Values are contextual, in that the social, intellectual, environmental or historical situation can influence how they are understood. Gephart (1981) demonstrates this with the example of 'gold', an:

> element that has value, not because of some intrinsic character, but because of mind sets about it, about its color, its tensile strength,

> malleability, etc. If its colour is what someone wants or needs, gold has value in that setting. If its tensile strength meets someone's need, gold has value in that need. In some instances, gold has symbolic value. It represents something important to people. In this sense it has value created in and by the mind and mental interactions (p. 15).

Value as a noun is also used to reference categories of things. For example, in the public administration literature, Pollitt (1993) distinguishes between democratic, general ethical, professional and people values, while in the evaluation literature, (Scriven 2007b) distinguishes between market value, essential value, public value, contextual value, illustrative and exemplary value. Used as a verb, value refers to something being judged as *having* worth (Magendanz 2003). In Gephart's (1981) description of gold, 'value' is used as both a verb and a noun, for example 'I value gold'.

Both the verb and noun forms of 'value' are relevant to the development of criteria, as demonstrated by the following example. A decision about whether one wears spectacles or contact lenses may depend on general values such as (i) lifestyle, (ii) comfort, (iii) convenience, (iv) cost and (v) aesthetics. An individual may have a mental model of what each of these values means. But if a professional evaluator is assessing a service that provides aids for vision-impaired people, a more precise description of each value may be required so that there is a clearer understanding of how each is being conceptualised. This description may draw on other values. For example, the value 'lifestyle' may include values such as 'work' and 'recreation'. In this respect, criteria are more precise than general values. A decision about which criteria are relevant to the decision making involves the process of valuing, therefore using 'value' as a verb.

Criteria convey what is important or what matters in a specific evaluand and its context. In some cases, criteria may include values that are *inherent* to the thing being examined. Examples of such values are relatively easy to identify in objects such as a wristwatch, where 'timekeeping' is integral to the meaning of a watch. In government-funded programs it can be difficult to identify the inherent values of an evaluand.

Ethical values are an example of values that are not inherently connected to an evaluand but still may inform criteria as they identify what is important or what matters in an evaluand. As Scriven (1986) notes:

> …one could hardly do program evaluation without looking at personnel policies within the program; would not unjust policies – perhaps extremes of gender discrimination – constitute grounds for downgrading or even condemning a program? (p. 14).

Likewise, the following example of an evaluation of pavement[11] surfaces on a state highway demonstrates how values can be difficult to define, as they can be considered both part of the meaning of a good evaluand and incorporate ethical values. In the example, the evaluator has included a criterion of 'effectiveness' which is further defined as 'wet-road skid resistance'. The evaluator argues that a road's ability to reduce car skids in wet conditions is part of the reason we have a pavement surface in the first place (otherwise, why not pave the road with metal chips). Wet-road skid resistance may therefore be considered part of the inherent function of a road pavement, particularly in areas where there is high rainfall. However, part of the reason we care about wet-road skid resistance is that it provides safer conditions for drivers, and safety is a public value that implicitly or explicitly underpins government funding of roads.

To summarise, for the purposes of this thesis, criteria define or describe the dimensions of value on which quality or success is judged. Criteria may comprise dimensions of value that are part of the meaning of the evaluand or they may describe what is valued in a particular program.

2.4.3 Characteristics of defensible evaluation criteria

This study is concerned with identifying the characteristics of defensible criteria. The concept of 'characteristic' is used here to refer to the elements that are required for criteria to be considered defensible. The concept of 'defensibility' is defined by the OED as 'the quality of being defensible in argument; justifiability'[12].

[11] In the transport sector, the term 'pavement' refers to the surface of a road (Gribble 2018).

[12] OED accessed May 2015

As noted in section 2.3.3 evaluation scholars have argued that defensible evaluative conclusions rely on strong evaluative reasoning. Such reasoning requires attention to all elements of an evaluative argument not just those related to methods and findings. Taylor (1961) illustrates the importance of attending to the value premises in evaluative reasoning through an imaginary philosophical dialogue between two people, in which A begins with a claim:

> A: This is a good painting.
>
> B: Why?
>
> A: Because, for one thing, its colors are harmonious.
>
> B: Why does that make the painting a good one?
>
> A: Because having harmonious colors is a good reason for judging paintings of kind K to be good, and this painting is of kind K.
>
> B: I admit that this painting is of kind K and that its colors are harmonious, but why is having harmonious colors a good reason for judging paintings of kind K to be good?
>
> B's last question may be stated in another way – why is having harmonious colors a valid or appropriate [criterion][13] for judging paintings of kind K? To answer this question satisfactorily, A must show not only that he, but that anyone (and therefore B), is justified in adopting the given [criterion] for judging paintings of kind K. That is, he must give good reasons why paintings of kind K are correctly included in the range of application of the standard. (p. 81)

B's last question requires A to justify the inclusion of 'harmonious' colour as a relevant criterion for judging the painting to be 'good'. That is, B is asking A to provide reasons why 'harmonious colour' ought to be included as a criterion for judging K-type paintings. Taylor's argument is as applicable to program evaluation as it is to philosophy. For an exploration of the aspects of criteria that need to be justified, the researcher turns to Scriven (1994a) criteria for criteria.

[13] Taylor uses the term 'standard' to refer to a criterion.

2.4.3.1 *Criteria for a set of criteria*

Scriven developed a list of criteria for a set of criteria as part of his work on evaluation logic and reasoning and informed by his experience developing checklists for evaluating products[14]. His first list, published in 1994, comprised five criteria and this was later updated to seven (Scriven 2007a).

The original list included the following conditions that he considered essential to development of a set of criteria: (i) only criteria of merit should be included, that is, properties that are part of the meaning of the concept of a good evaluand of the type being evaluated; (ii) the list should be comprehensive, that is, all the significant criteria for the evaluand should be included; (iii) criteria should be clearly described so one can determine the presence of the criteria; (iv) they should be described at the same level of simplicity or complexity; (v) criteria should be non-overlapping (Scriven 1994a). His later revision of the properties (Scriven 2007a) included two additional conditions: (vi) the list should be concise, and (vii) confirmable.

The first condition of a set of criteria is that of criterial status. Scriven (1994, 2007a) differentiates between criteria of merit (CoM) and indicators[15] in what he refers to as the 'tough relevance requirement' (Scriven, 1994, p. 374). In criteria of merit, Scriven is seeking to identify the essential characteristics that define what it means to be a good X. Thus, a criterion of merit is:

> one that bears on the issue of merit, sometimes very heavily (so that a failure on that criterion is fatal) but often just in the sense of being one of several that are highly relevant to merit, although not – in itself – absolutely essential (Scriven 2007a, p.7).

In explaining the difference between criteria and indicators, Scriven draws from product evaluation the example of a wristwatch (Scriven 2007a). He argues that timekeeping is essential to the meaning of a watch and is therefore a criterion *of merit*. In contrast, the brand of a watch may be tenuously connected to merit – it indicates quality but cannot be relied on. A person wanting to purchase a good

[14] Michael Scriven was asked about the origins of the properties in April 2018. Dr Scriven's response was conveyed to the researcher in an email dated 31 March 2018.

[15] Scriven also refers to primary or direct indicators (by which he means CoM) and secondary or indirect indicators which are empirically connected to CoM (Scriven 2007a).

quality watch may search for a prestigious brand, equating a high-end name with quality. In the case of a brand such as Rolex, Scriven argues there is sufficient evidence to show that many other watch brands will outperform on a criterion of timekeeping.

Scriven suggests indicators should be avoided, if possible. His concern is that indicators can be tenuous in their validity and open to manipulation. He concedes that indicators may need to be used in some evaluations, for example, if there is a lack of time to develop a set of criteria. However, in these cases, he suggests the conclusions need to reflect this limitation. His argument is that an evaluation relying on indicators can rarely, if ever, reach the same level of defensibility as one that includes criteria of merit.

A second example of the difference between criteria of merit and indicators is provided by the researcher, based on evaluation of a disability program. In this hypothetical example, the evaluator is tasked with assessing the support provided by community workers to disabled people. An indicator of success could come from the administrative data collected by program staff, in this case the number of times the community workers and disabled people meet over a set amount of time. The indicator has a tenuous link to the concept of support and should be treated as unreliable, as it says little about the disabled person's experience of the support they received. A preferable approach would be a criterion that defines the attributes of support for this context: what should 'good support' look like in this context?

The second condition of a set of criteria is that it be comprehensive. Here, the concern is that all relevant dimensions of value are included, in order to create as complete as possible understanding of quality or success in a particular context. Scriven's (1990, 1995) rationale is that a partial list of criteria might distort a judgement if it misses an important criterion. For example, one might argue that an evaluation of watches would be incomplete if it focused on brand, cost and size, but did not attend to the dimension of timeliness.

Likewise, Stake and Schwandt (2006) argue for comprehensiveness, in the sense that quality should include not just a focus on impact, but also its implementation. Thus, comprehensiveness may refer not just to the evaluand itself but to the different stages of a program. This is illustrated here with an example. An

evaluation of a homelessness program might result in a judgement that initiative has had negligible effect on the housing situation of homeless people for whom the program is intended. If the evaluation focuses just on outcomes, it may miss important process-related factors that are not occurring, for example, homeless people's ability to pay for accommodation or access to counselling support to address mental health issues.

A second way in which the concept of comprehensiveness might be understood concerns the inclusion of different conceptions of value that might inform how quality or success is understood. In this sense of the word, comprehensiveness refers to the concept of inclusiveness of values, or values pluralism. Such an approach recognises that there can be multiple perspectives on quality, success or goodness. Some scholars have advocated for values pluralism in evaluations of publicly funded programs (e.g. Greene et al. 2006; Mark and Shotland 1985; Shadish et al. 1991). As Shadish et al. (1991) contend, a commitment to values pluralism:

> increases our concern lest any legitimate stakeholder group – either disadvantaged or advantaged – be excluded… evaluations require pluralistic input on values perspectives, just as they require pluralism in methods (p. 183).

Likewise, House and Howe's (1999; 2000) deliberative democratic evaluation approach attends to comprehensiveness via values pluralism, through adherence to three principles: inclusion of all relevant perspectives; dialogue so that all relevant views are 'properly and authentically represented' (p. 94) and deliberation.

The third condition of a set of criteria is that each criterion be clear or "comprehensible" (Scriven 2007a, p. 4) The rationale is that if criteria are not well described they risk being invisible or unclear, which makes it difficult to assess whether all relevant criteria have been included in an evaluation.

The fourth condition of a set of criteria is that each criterion be described at the same level of commensurability (Scriven 2007a). By example, Scriven describes criteria as having equal levels of generality, beginning with abstract headings that then cascade into three further levels, each of which further unpacks the level above it. The main purpose of commensurability is to ensure that concepts are nested at the same level of specificity across a set of criteria.

A fifth condition is that criteria be non-overlapping (Scriven 2007a). Here, Scriven is concerned that criteria are distinct and independent, to avoid double-counting. The sixth condition is that criteria should be concise, that is, avoid superfluous criteria (Scriven 1994, 2007a). The final condition is that criteria be confirmable, that is, measurable or reliably inferable.

2.4.4 Explicit versus implicit criteria

This next section of the review examines two broad approaches to criteria development. In the first, criteria are explicit, outlined in detail and used to guide the evaluation from the start; in the second, criteria remain implicit.

Some scholars argue for explicit management of criteria on the basis that such an approach can best support evaluative reasoning. For example, Davidson (2013; 2005) and King, McKegg, Oakden & Wehipeihana (2013) promote the use of rubrics, in which criteria are visually represented alongside standards and descriptors of performance, then used as a tool to reach transparent and explicit evaluative judgements. Often, stakeholders are engaged in identifying the range of values, with the intention of understanding what matters (Davidson 2005; King et al. 2013). The criteria and standards that emerge from this process provide a guiding framework for the evaluation as they identify what information needs to be collected (Scriven 2013b).

Explicit identification of criteria is also consistent with the evaluation standards endorsed by the Australasian Evaluation Society (AES) and those of the Aotearoa New Zealand Evaluation Association (ANZEA). The New Zealand evaluation standards, for example, suggest explicit and transparent criteria are fundamental to trustworthy results (Superu and Aotearoa New Zealand Evaluation Association 2015).

Other scholars are concerned that explicit approaches to criteria development can result in a misrepresentation of a program's quality. For example, Bamberger, Rugh and Mabry (2006) have argued that programs are generally too complex and context dependent 'to be accurately rendered by comparison with a few criteria' (p. 299). Stake and Schwandt (2006) have also raised concerns about quality being too narrowly defined. They argue that most evaluators do not pay sufficient attention to the constructs that define quality for a particular evaluand. Instead, they observe evaluators using a few easy-to-articulate indicators that do not fully represent

quality. Their response has been to advocate for an emergent and implicit approach to establishing criteria (Stake 2004; Stake and Schwandt 2006). In such an approach, quality is described in the form of narratives:

> This view emphasizes grasping quality in experience-near understandings, that is, in the language and embodied action of those who actually are undergoing the experience of a program or policy (Stake and Schwandt 2006, p. 408).

The resulting judgements, they argue, draw on a richer, more nuanced and multifaceted understanding of quality, worth and significance.

Similarly, Eisner (1976, 2004) drew on the concept of connoisseurship and criticism to suggest that quality can be understood only when one has extensive expertise in a subject. Experts reach holistic judgements based on their extensive and intimate knowledge and experience of a subject.

An implicit approach to managing criteria has its own challenges. First, it relies on the evaluator having significant content expertise in order to recognise quality, or lack thereof. It is this expertise that provides the lens through which quality is understood. This may be possible for evaluators who assess the same type of program again and again in the same contexts; but is less appropriate for generalist evaluators who are required, in the course of their work, to evaluate different types of programs operating in different contexts. A second, related, challenge is that relying on the evaluator to know quality when they see it can lead to allegations of subjectivity. Even if qualified and experienced, the evaluator brings their own biases and personal context to bear on how they understand quality. It is their understanding of value that shapes the evaluative description. Different evaluators, using such an approach, could produce alternative narratives without being required to justify their selection of values.

In practice, implicit and explicit approaches may be viewed as complementary, with each having a critical role in developing a comprehensive set of criteria. Explicit criteria can help stakeholders discern gaps due to what has remained implicit or not previously considered, while implicit criteria can surface in explicit form as more is learned about the evaluand (Martens 2018). While recognising and supporting this complementary interplay, the current study has a focus on the development of explicit criteria.

With regard to fully describing criteria, both Sadler (1985)and Scriven (1990) refer to the use of hierarchical approaches to move from the abstract to more specific levels of detail. Sadler (1985) notes that criteria may be presented as abstract terms at a high level, as for example, in single words such as 'effectiveness' or 'relevance', and then more detail is provided at a lower level, with a view to clarifying meaning to support the reasoning process. Another advantage of describing criteria hierarchically is that it can be easier to show their relevance to merit for an evaluand and ensure that dimensions of criteria are independent of each other. This can serve to avoid double counting (Scriven 1990; 2007; 2013b).

Both Scriven (1990) and Fournier (1993) note that if definitions are left at too broad a level, it can be difficult to connect criteria to a meaningful level of performance that can be observed and measured. Accordingly, Scriven advocates for the use of checklists to identify all relevant dimensions of merit, in order to avoid the Rorschach effect, that is, the tendency to see what one wants to see in the data. Rather than making a wide-sweeping judgement based on an abstract criterion, the evaluator is forced to consider performance on each dimension.

2.4.5 Justifying criteria

To reach a defensible evaluative conclusion, evaluators need to be able to justify the criteria they have included. According to Shadish et al. (1991) sources used to identify and select criteria include the program goals, claims by stakeholders about program achievement and factors they consider are important for understanding success or failure, logic models that explain how inputs relate to processes and outcomes, and the degree to which material needs are met. Davidson (2005) and Scriven (2013b) identify a similar, although much broader, list of sources for identifying criteria, of which a selection is included in Table 2.

Table 2: Sources for identifying criteria

Types of sources	Potential purposes
Cultural standards	Understand what is culturally relevant or what has a fit with local norms
Definitions	Understand elements of a program that are essential to its meaning
Ethics	Identify ethical dimensions such as fairness, equity, duty, rights.
Expert judgement	Include subject matter expertise

Types of sources	Potential purposes
Fidelity to specifications	Identify the proposed archetype or exemplar
Institutional perspectives	Derive criteria based on a broader institutional view, e.g. from a World Health Organisation (WHO) perspective
Legal requirements	Understand legal compliance
Needs assessment	Identify what is essential for something (or someone) to function at a satisfactory level
Organisational goals	Strategic goals of the organisation/s funding/implementing the program
Professional standards	Relevant professional guidelines that apply to the evaluand
Program goals	Expected outcomes
Research literature	Identify relevant academic knowledge
Scientific/technical	Program content corresponds to current knowledge (usually accessed via referenced literature)
Stakeholders	Understand stakeholders' interests

Note: List drawn from Davidson (2005) and (Scriven 2013b)

Davidson (2005) and Scriven (1991) are cautious about using program goals or objectives as the sole source for identifying criteria of merit. Their reasoning is that goals may reflect the expectations of certain powerful interests rather than the genuine needs of a target group. Even when needs are known, compromises may have to be made to take account of political sensitivities, input from lobby groups representing private interests and competing requirements for resources. This results in goals and objectives that are sub-optimal or patched together as a consequence of having to accommodate numerous conflicting interests.

2.4.6 Approaches to developing criteria

Scholars approach the development of criteria in different ways. Scriven (2013b) establishes criteria by assessing the targeted community's needs, and through consideration of other possible sources. For Scriven, the primary function of a list of sources is to serve as a checklist. His focus is on identifying and selecting all the essential dimensions or characteristics that distinguish between a good/bad, successful/non-successful program. Drawing on all relevant sources is how evaluators identify 'all the features and consequences of both the intervention and the means for its achievement', Scriven (2016, pp 13–14). He refers to such

elements as criteria *of merit*. While Scriven advocates evaluators to consider all sources, he does not accept them at face value. For instance, he argues that the evaluator has an obligation to assess whether stakeholders are knowledgeable and well informed for their views to be considered as legitimate. Scriven has also been a vocal critic of program goals as the primary source of criteria, noting that it can be futile to know if goals (that certain stakeholders value) have been achieved if in fact they are not worth achieving (Scriven 1972). Goals may be politically motivated or based on a compromise of different interests rather than oriented around the actual needs of affected communities. Further, a focus on goal achievement can ignore important side effects. It is for these reasons that Scriven has promoted the use of goal-free evaluation whereby programs are judged independently of their stated or implied goals. Shadish et al. (1991) agree that goals can be an unreliable source of criteria, adding that they are 'often vague, contradictory, or latent' (p. 457).

Greene's approach has been to advance the interests of stakeholders who are least well served through promoting values of justice and equity (Greene 2011). Similarly, House and Howe (2000) suggest that criteria be oriented around ethical perspectives that focus on whether the program is 'right', 'fair' or 'just'.

Shadish et al. (1991) have categorised the different approaches to valuing as either descriptive or prescriptive. In descriptive valuing, evaluators limit their role to describing the values held by stakeholders; judgements are then made against stakeholders' perceptions of merit, worth or significance. In prescriptive valuing, certain normative ethical values such as equity or rights, are promoted by the evaluator, as advocated in deliberative democratic evaluation (House and Howe 2000).

While eliciting descriptive values can facilitate evaluation by addressing the political context in which it is practised, Shadish et al. (1991) suggest that prescriptive approaches can provide more solid and justified grounds for understanding value for a program. This is because:

> The justification and internal explication of prescriptive theories usually have been carefully worked out over centuries of philosophical thinking… (p. 48)

Further,

> Prescriptive theories give evaluators a critical perspective and intellectual authority that descriptive theories cannot match. They broaden evaluators' understanding of good social programs by broadening their understanding of what is good for the human condition generally (p. 49).

The prescriptive theories that Shadish et al. (1991) refer to are located in moral philosophy. Such theories are concerned with normative ethics and the study of what is good and bad. Both Shadish et al. (1991) and House (1980) note that such theories are evident in scholars valuing approaches even when theorists themselves are not explicit about them. For example, Scriven's definition of need, which is central to how he understands criteria of merit, 'implies endorsing assumptions of egalitarian theories of ethics' (Shadish et al. 1991, p. 49). When criteria are based primarily on program goals and decision makers' perspectives, House (1980) suggests that the underpinning ethical assumptions are utilitarian. Meanwhile, his preferred approach, deliberative democratic evaluation, aims to incorporate the interests and perspectives of all legitimate stakeholders, including those with the least economic and political power.

While ethical theories clearly have relevance to evaluation and the task of sourcing values, Shadish et al. (1991) argue there are difficulties with prescriptive valuing that are insurmountable for the evaluation practitioner. They describe these difficulties as follows. First, they argue there are multiple competing theories about value and as such there is no agreement about which is best. In relation to Rawls' theory of justice, for example, there are many credible alternatives:

> … justice is just one moral concern in evaluation, along with human rights, equality, liberty, and utility (p. 456).

A second, related argument, is that democracy fosters pluralism of values. The authors argue that public programs generally encompass the values of multiple stakeholders, and therefore criteria should not be dominated by one conception of good, however persuasive the theory. A third, and very serious concern, is that the evaluator's values may colour which conceptualisation of good is adopted. The risk here is that the evaluation findings may be dismissed as subjective if the evaluator is seen to have privileged certain values.

For these reasons, Shadish et al. (1991) suggest it is more practical to use descriptive valuing, that is, assess against the values held by stakeholders. Indeed, they argue that this is the approach most evaluators adopt anyway, although they provide no empirical backing for this assertion.

Descriptive approaches to valuing, that is, describing the values of stakeholders, are evident in the work of many evaluation scholars, including Greene (2005), Hall et al. (2012), Mark et al. (2000), Strike (1979) and Lincoln and Guba (1986). However, examination of different scholars' approaches to descriptive valuing indicates that criteria may still informed by the evaluator's stance, that is, a prescriptive approach, albeit in a more implicit manner. For example, Patton's (2008) utilisation-focused approach is concerned with supporting decision makers or 'intended use by intended users' (p. 37). In such an approach there is a tendency for the priorities, and thus values, of decision makers (as the users of the evaluation) to determine criteria. In fact, House (1980) assigns Patton's utilisation-focused evaluation as aligning to a managerial (elite). Meanwhile, Lincoln and Guba (1986) see the evaluator's role as negotiating or mediating between different stakeholders' value positions to ensure all stakeholders' interests are represented fairly. Their approach aligns with democratic principles and values pluralism, which Shadish et al. (1991) argue is:

> the best defense against the tyranny of our individual value biases… is achieved by listening to all the legitimate stakeholder groups (pp. 183-184).

Greene (2001) has also endorsed a democratic pluralist approach to valuing, arguing:

> …multiple stakeholder interests must be included and respectful dialogue among those interests must be pursued in order to render meaningful evaluative claims about the evaluand (p. 184).

While there are clear merits in including multiple stakeholders' values into the way criteria are conceptualised, there are also challenges. As with prescriptive valuing, there is the issue of deciding the relative importance of different perspectives, for example, where there are divergent values, whose should be prioritised or given more weight? Greene suggests this can be dealt with through dialogue; however,

she provides little practical guidance about how one might defend certain values over others in order to reach an evaluative judgement. Chelimsky (1998) draws on her experience working in the U.S. General Accounting Office, and observes that agreement amongst stakeholders was often difficult, especially when officials were concerned that an evaluation could hurt their interests. She argues that the evaluator needs to take ultimate responsibility for valuing, stating:

> (t)he evaluator has no vested interest in the program, or at least much less interest than the stakeholder has, and the evaluator's purpose is the public good, which may or may not be congruent with the stakeholder's good (Chelimsky 1998, p. 45).

Critics of descriptive valuing also argue that such an approach assumes stakeholders can identify all relevant dimensions of quality for a program, that they are well informed and knowledgeable, and that the information they provide is reliable and trustworthy. As Scriven (1991, 2016) warns, the evaluator has an ethical obligation to attend to all relevant dimensions of quality and not just those the program decision maker wants to know about:

> The heart of the matter for the professional evaluator is, 'How good or bad is the program (including its components) – more specifically, what is its quality, comparative cost-effectiveness, and/or significance – in this context?' Answering that question is not complete if you leave the client with much of the hard part to handle (Scriven, 1991, p. 13).

House and Howe (1999) have also criticised Shadish et al's. (1991) emphasis on descriptive valuing, arguing that practicality is insufficient as a criterion for a valuing process: '…practicality as a primary criterion for evaluation means evaluators may serve whatever ends clients or policy makers endorse' (House and Howe 1999, p. 36). They also argue that evaluators can be right even when there is no agreement amongst stakeholders. By example, they refer to Brown v. Board of Education in which the justices ruled, despite 'rancorous disagreement' that racial segregation of children in public schools was unconstitutional (House and Howe 1999, p. 49). In this example, legal sources provide a justification for a certain ethical stance. Further, while there are multiple ways of understanding good, House and Howe suggest that not all values are on a par:

> For example, one person's right to non-discrimination is different from another's desire to retire in Boca Raton. And should the two come into conflict, the right to non-discrimination should take precedence over the desire for an extravagant retirement (House and Howe 1999, p. 49).

The challenge, it would seem, is to justify certain ethical values through a reasoning process. But just how might evaluators do this? Justifying certain values over others may be relatively straightforward when using House and Howe's example of non-discrimination versus a right to an extravagant retirement, but in many government programs there are competing and equally worthy values. Schwandt (2002) identifies the challenges inherent in this question with the example of a program designed to provide support to disabled people. He asks:

> What metavalues are important here: Fairness and equity in services provided? Absence of any additional harm in the act of providing the service? Efficiency? Physical appearance of health care personnel? (Schwandt 2002, p. 141–142).

Normative ethical perspectives provide a potential pathway to address this challenge.

2.5 Normative ethical perspectives

2.5.1 Purpose

This section of the review provides an overview of normative ethical perspectives and their relevance to identifying and selecting criteria. Its purpose is to provide an understanding of the different perspectives that inform prescriptive valuing that were introduced in the previous section. Normative ethical perspectives provide the theoretical foundation for a conceptual framework introduced in section 8.3.

2.5.2 Introduction to normative ethical perspectives

Normative ethics is a branch of western philosophy that deals with the rightness or wrongness of actions (Popkin and Stroll 1993). The philosophy literature describes three competing ethical perspectives on how moral questions should be answered: consequentialist, deontological and virtue-based (e.g. Boston, Bradstock, and Eng 2010; Braham and van Hees 2014; Cavanagh, Moberg, and Velasquez 1981; Perrett

and Patterson 1991), with different theories nesting under each, as illustrated in Table 3.

Table 3: Examples of theories that align with different normative perspectives

Perspective	Example of theories
Consequentialist: Assesses right, wrong, good or bad on the basis of its consequences	Utilitarianism: Maximise utility, happiness, welfare; the greatest good for the greatest number (Popkin and Stroll 1993)
Deontological: Acts are deemed right, wrong, good or bad in all circumstances regardless of the consequences	Duty-based: All people are created equal, no person should be discriminated against before the law (Popkin and Stroll 1993)
	Rights-based: Protect the ethical rights of those who will be affected by an intervention (Bromell 2012)
	Fairness or justice: People are to be treated equally or proportionately (Harper and Stein 1992)
Virtue-based: Concerned with inter-personal relationships. Rejects universalism and argues that care is always contextual (Kristjansson 2006)	Caring should be a foundation for ethical decision-making (Noddings 1984)

Consequential ethical perspectives assess the goodness, or otherwise, of a program primarily on the basis of its consequences, or outcomes (Boston et al. 2010). One of the most well-known consequentialist theories is utilitarianism, which views actions as right, wrong, good or bad when they produce the greatest possible outcomes for the greatest number of people (Childress 2009). A program is successful if, for example, it increases wellbeing or lessens suffering. The focus is on securing the best consequences for the group as a whole.

Defining criteria of merit through a consequentialist lens would include an emphasis on short- or longer-term outcomes and impact. A potential limitation of a consequentialist perspective is that it may downplay, or ignore outcomes achieved (or not achieved) for a small, but significant interest group. This is because the focus is on achievement of the greatest good for the greatest number. By example, an evaluation of a program intended to support disabled people to move from residential care into the community might include a criterion that defines 'effectiveness' as the program achieving its intended outcome for as many people as possible. The evidence may show that the program works for the majority of

participants; specifically, those with low or medium-level support needs. Based on this conception of success, the program could be hailed as a success by government decision makers. However, viewing success through this lens may serve to disadvantage a small group of people who require additional support to achieve the same outcomes. Some scholars get around this potential limitation by assigning some of the priority to those most disadvantaged (Hooker 2002).

With a deontological perspective, acts are deemed right or wrong in all circumstances regardless of the consequences (Childress 2009). The focus is on whether something is consistent with certain agreed ethical principles (Boston et al. 2010). This raises the issue of what principles should count as important, and if there are conflicts between relevant principles, how these should be resolved. Deontologists use decision-rules for resolving ethical conflicts, for example, giving absolute priority to a one ethical stance, or using some kind of weighting. Using the disability example above, the evaluator might describe success as the program enabling all disabled people who want to move out of residential care and into community-based accommodation to do so. This value stance could be justified by a rights-based principle.

Beyond rights, Newman and Brown (1996) describe two other broad principles that are aligned with a deontological perspective. In the first, duty takes precedence as the way success (or non-success) is viewed. What is important is a sense of duty based on expectations, commitments and obligations. Evaluators looking for duty-based perspectives might focus on existing expectations and contractual agreements. Such a contract, for example between government and a provider, may then frame performance criteria. A potential limitation of a duty-based principle is that it may over-emphasise a managerial perspective and risk neglecting the interests of other important interest groups.

The other principle within a deontological perspective emphasises concepts of equity, fairness and social justice. Here, the focus is on the least advantaged, as demonstrated by House's work (House 1995). Gender and ethnicity are also important attributes traditionally related to social justice and equity. Drawing on a principle of equity, criteria might be tilted in the direction of the needs and interests of those most disadvantaged. The evaluator may need to look beyond outcomes to examine the political and social conditions that have led to these outcomes and how the program takes these conditions into account. A potential

limitation of an equity-based principle is that some people may adopt a sense of entitlement to certain outcomes (Newman and Brown 1996).

A virtue-based perspective stresses the importance of developing good habits of character, such as generosity, self-respect and courage. Aligned with a virtue perspective is a principle of ethics of care, which emphasises the importance of relationships, trust and responsiveness to need. It emphasises connectedness, for example through concepts such as social responsibility. In contrast to the other perspectives, ethics of care is always contextually applied. Using the disability program example, an evaluator might frame success as person focused, for example the extent to which the program is able to support the different needs and desires of disabled people to live as independently as possible. A potential limitation of ethics of care is that it can be ambiguous.

2.5.3 Relevance of normative perspectives to the development of criteria

The relevance of normative perspectives to the establishment of criteria is evident in the writing of several evaluation scholars, most notably those with a background in philosophy. Scriven (2007b, p.2) considers normative ethics to be a "specialized domain" of evaluation in that its subject area is central to the task of valuing. In one article he notes that an understanding of duty is essential to personnel evaluations (Scriven 1994), implicitly suggesting that a deontological perspective be considered when defining criteria of merit. House (1980a) observes that normative perspectives are often tacit and far from direct, yet they significantly affect how evaluations are shaped. Specifically, the way in which 'good' or 'success' is defined depends on conceptions of value that different stakeholders hold.

More recently, both Schwandt (2018a) and Stame (2018) have argued for more ethically-oriented approaches to evaluation that take account of different conceptions of value. Both scholars argue that evaluators have a moral responsibility to surface broader conceptions of 'right' or 'good' than those articulated by the evaluation commissioner. Stame asks this question directly:

> '…what would ethical criteria look like if they also had to support the agency and development of people? They would undoubtedly have to avoid the usual trap of setting out fixed targets that encouraged yes/no (success/failure) judgments' (Stame 2018, p. 447).

Missing from the literature is practical guidance on engaging with consequentialist, deontological and virtue-based perspectives. Where ethical approaches to criteria development are discussed, the focus is primarily on engaging with multiple stakeholders. For example, Greene et al. (2006) and Shadish et al. (1991) aim to include multiple perspectives on 'good' or 'success' that are rooted in concerns for social justice or a moral obligation to democratic inclusion. Consideration of all relevant perspectives is consistent with U2 'attention to stakeholders' and 'U4 'explicit values' of the Program Evaluation Standards (Yarbrough, Shulha, Hopson and Caruthers 2011) and the second principle: ethic of care of the Evaluation Standards for Aotearoa New Zealand (Superu and Aotearoa New Zealand Evaluation Association 2015).

2.6 A gap and an opportunity

As established in this background review, selecting and justifying relevant criteria is critical to defensible evaluative reasoning, especially so in evaluations of publicly funded programs. Yet much of the scholarly literature has focused on reasoning at a meta level, with little investigation of the individual elements. Beyond a few resources, the identification and selection of criteria – or values – is a particularly underdeveloped area of evaluation practice. Where criteria are discussed in the literature, scholars' claims about how to best manage criteria have little empirical backing. Claims about use of prescriptive or descriptive approaches to valuing are a case in point. What approaches are used in current program evaluation practice? Which approach can best support the development of defensible criteria? A review of research on evaluation published between 2005 and 2014 identifies that very few empirical investigations address questions related to values and valuing (Coryn, Wilson, Westine, et al. 2017). As Alkin et al. (2012) note:

> 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).

The review of the literature identified three studies that have included a focus on criteria. In the first, Hurteau et al. (2009) examined 40 evaluation reports for evidence of judgements. Of the 20 reports that included a judgement, all but one (95%) were determined to have included justified criteria on the basis that they included reference to a source. Likewise, Nunns et al. (2015) found justified criteria

in a third of the reports they examined. Again, justification was defined as the evaluator including reference to a source, for example the sponsor and/or stakeholders being engaged in defining the criteria. Both two studies raise questions about the characteristics of defensible criteria. Specifically, is mere reference to a source sufficient for criteria to be deemed defensible? If not, what else is required of criteria to support the evaluative reasoning process?

A third study involved a survey of American Evaluation Association (AEA) members to investigate their knowledge, perceptions and applications of the logic of evaluation in their evaluation practice (Ozeki et al. 2019). The findings indicated that many participants had no clear understanding of the meaning of criteria, evidenced by many using the terms criteria and standards interchangeably. Yet the majority indicated they frequently or always use them when making an evaluative judgement. As the authors of this study note, the findings raise a concern about whether evaluation practitioners are conducting evaluation or simply applying social science methods to address questions of interest to their clients.

From the review comes an opportunity to investigate the characteristics of defensible criteria for Australian and New Zealand public sector evaluations. A key gap in knowledge is understanding how criteria are currently managed in program evaluations. The review has raised a number of questions that could be examined, including: What evidence is there of explicit criteria in practice? How do evaluators justify criteria in program evaluations? How can this aspect of practice be better supported? What theory-informed tools might bridge the gap between theory about valuing and the development of defensible criteria in evaluation practice?

2.7 Aims of this study

The overall aim of this study is to identify the characteristics of defensible criteria for Australian and New Zealand public sector evaluations. Two further aims are to understand how criteria are currently managed in Australian and New Zealand evaluations of public sector policies and programs, and to explore the development of a theory-informed tool for establishing defensible criteria in evaluation practice.

In this thesis, the term ‘program evaluation’ is used to refer to both public sector policy and program evaluations.

2.8 Research questions

This study has one central research question:

1. What is required to develop a defensible set of evaluation criteria for program evaluations?

In addition, the study includes two secondary research questions:

2. How are criteria managed in Australian and New Zealand program evaluations?

3. What is an evidence-informed conceptual framework for defensible criteria that can be operationalised for evaluation practice in the public sector?

2.9 Bounding the study

Williams and van't Hof (2014) note that every study has boundaries:

> You cannot do everything, consider everything, see everything, record everything (p. 20).

This statement is pertinent for two reasons. The first relates to the topic of this study. The issue of what to *consider* is at the heart of decisions about evaluation criteria and the topic of this study. How ought the evaluator go about the task of considering what criteria to include?

The second reason relates to the geographic bounding of the study to evaluation practice in Australia and New Zealand. The decision to incorporate both countries in the study, while excluding other countries with a similar democratic history and tradition of program evaluation, is as follows. Both Australia and New Zealand share a common colonial heritage and have governments based on the Westminster system. From the late 1980s until 2018 the two countries shared a professional evaluation association, the Australasian Evaluation Society (AES). This association has been pivotal in facilitating professional learning that has been attended by evaluators from both sides of the Tasman Sea. In this respect, the study provides an opportunity to pool together the collective wisdom of professional evaluators on a topic that is critically important to all those engaged with program evaluation. There are also key differences between the two countries that may contribute to differences in how evaluators approach criteria development. For

example, the researcher's assumption is that evaluation practice informed by Treaty of Waitangi[16] principles of partnership, participation, and protection may result in variances in the way some program evaluators in New Zealand identify criteria; for example, more explicit engagement of cultural expertise and attention to culture in the way criteria of merit are framed. In this respect, the study provides an opportunity for comparative analysis of practice between the two countries.

2.10 Contributions of the study

The expected contributions that this study will produce to fulfil the academic needs consistent with the requirements of a PhD are as follows:

Contribution to knowledge: The study will contribute to theories of evaluation, defined as 'general accounts of valuing, such as what the nature of valuing is, how it is done, or what its logic is' Shadish (1994, p. 351). Specifically, the study will extend the body of evaluation-specific knowledge to describe the characteristics of defensible criteria. This will extend understanding of the different elements of criteria development that evaluators need to consider when undertaking program evaluations.

Contribute to evaluation practice: To the researcher's knowledge, this is the first study to examine, in depth, how criteria are managed in program evaluations. The study will describe what evaluators make assessments against, the sources they most commonly use, and those they consider important, to establish criteria. It will also examine how criteria are organised and presented in program evaluations. Understanding how criteria are currently managed in practice is an important step toward improving Australian and New Zealand program evaluation practice.

In addition, the study will explore an approach for systematically identifying and selecting relevant criteria that is theory based and able to be implemented in practice. A conceptual framework will be developed for identifying dimensions of value based on normative ethical perspectives. Resources to support evaluators to engage with the framework will be developed and field tested. The study will also describe what happens when evaluators engage with a theory-informed tool to

[16] The Treaty was signed by the Crown and Māori in 1840. It is considered to be the founding document of New Zealand. The Treaty has profound implications for how public sector evaluation is practised in New Zealand.

develop criteria for a current evaluation. The framework will encourage evaluation practitioners to be more thoughtful about how values are applied in an evaluation and provide a way to think about ethical values and their role in how criteria are conceptualised.

2.11 Chapter summary

This chapter has examined the existing literature relevant to the development of defensible evaluation criteria for program evaluations. Evaluation is, by definition, judgemental and values based. This is articulated in the general logic of evaluation. The dimensions of quality or goodness against which judgements are made are referred to as criteria.

The literature identifies three characteristics considered essential to the development of defensible criteria. The first concerns identifying and defining the essential aspects of quality or success for a program and the second is that a set of criteria be comprehensive. A third draws on concepts from argumentation, specifically that criteria be based on authoritative sources. Of these, criteria that draw explicitly on normative ethical theories, that is, a prescriptive approach to valuing, appear to provide solid and justified grounds for understanding value for an evaluand. The three characteristics provide the basis for a theoretical framework for understanding the characteristics of defensible criteria for Australian and New Zealand program evaluations.

Within the context of Australia and New Zealand this study aims to: (i) identify the characteristics of defensible evaluation criteria for public sector evaluations; (ii) examine how criteria are developed in current public sector evaluation practice, and (iii) develop an evidence-informed conceptual framework for defensible criteria that can be operationalised for evaluation practice in the public sector.

In Chapter 3, aspects of the Australian and New Zealand program evaluation context are described.

3 Program evaluation in Australia and New Zealand

3.1 Purpose

This chapter provides an overview of key historical, political events that have informed how evaluation is conceptualised and practised in Australia and New Zealand. Its purpose is to provide contextual information relevant to the study for subsequent interpretation of the research findings.

In many western democracies the demand for program evaluation has become more evident since the 1960s, in response to increased government investment in social services and a need to measure the outcomes of those investments (Lunt and Trotman 2005; Rossi, Lipsey and Freeman 2004; Shadish and Luellen 2005; Sharp 2003). While this demand can be viewed as a global phenomenon in western democracies, at a country-level there are distinct historical and political factors that have informed how evaluation has been conceptualised and practised. It is therefore important to describe aspects of the Australian and New Zealand public sector evaluation context relevant to this study to inform understanding of the findings.

3.2 Program evaluation in Australia

In Australia, formal evaluation activity has been traced back to the years following World War Two, in particular related to agricultural extension studies (Sharp 2003). However, over the following 30 years, evaluative activities tended to be *ad hoc*, initiated in response to an incident or through chance encounters while on overseas' trips. For example, the experience of thalidomide[17] led to the establishment of the Australian Drug Evaluation Committee in 1963. The Committee's role included formulating measures for evaluating the potential effects of drugs being used in Australia. In another example, a group of Australian educators visited the U.S. in the early 1970s and heard about the use of evaluation to foster development and improvement in schools. Their visit was followed up by

[17] A drug taken by pregnant women around the world to alleviate morning sickness, and which resulted in many thousands of infants being born with phocomelia (malformation of the limbs).

an invite, from the Headmasters College of Australia, for six U.S. educators to conduct evaluation workshops in most Australian capital cities (Herbert n.d.).

Two inquiries into Australian public administration, resulting in the Coombs Report in 1976, and the Baume Report in 1979, paved the way for more widespread endorsement of program evaluation (Sharp 2003). The Coombs Report recommended an agenda of reform, including performance audit and the adoption of new public management (NPM) principles, while the Baume Report provided successive governments with a recommended definition of program evaluation, namely 'the process of thoroughly and critically reviewing the *efficiency*, *effectiveness* and *appropriateness* of any program or group of programs'. (Sharp 2003) notes there were various attempts following these reviews to implement more systematic program evaluation across government, but this did not occur until the late 1980s, except in agriculture extension and in education. A key focus of the NPM approach was to foster a greater focus on 'value for money', with an emphasis on program outcomes, as well as access and equity (Ryan 2003). As part of this approach, services previously run by public servants were contracted out, and from the late 1980s Commonwealth-funded programs in Australia had to have mandated annual program evaluation plans and reports.

By the late 1980s, program evaluation had become well established in the Australian public sector. As demand for evaluation grew, so did the need for professional evaluation support. The AES was established in 1986, the same year as the AEA[18], and five years after the world's first professional association, the Canadian Evaluation Society, was established. By 1990 there were at least three graduate diplomas in program evaluation being offered in Australia, one at the University of Melbourne and the others at the Phillip Institute of Technology and LaTrobe University (Sharp 2003).

In the mid-1990s, the Australian Cabinet established a requirement for all programs to be reviewed by both the Department concerned and the Department of Finance. This was followed, in 1997, by an Outcomes and Outputs Framework that necessitated all federal programs to demonstrate *appropriateness, efficiency*

[18] The Evaluation Research Society and the Evaluation Network merged in 1986 to become AEA.

and effectiveness. It is interesting to note the alignment between these criteria and those in a definition used by the AES, in which evaluation is viewed as the:

> systematic collection and analysis of information to make judgements, usually about the effectiveness, efficiency and/or appropriateness of an activity (Australasian Evaluation Society 2013, p. 5).

During this period there was an emphasis on outsourcing the delivery of government programs. As programs were contracted out, agencies turned their focus to scrutinising the performance of services through evaluations, with the intention that this inform decisions about the next round of contracts (Mackay 1998). More than a decade later, the Public Governance, Performance and Accountability Act 2013 was introduced. This also placed greater emphasis on accountability, and reporting results of programs to Parliament, relevant ministers and the public. Anecdotal evidence suggests this Act has resulted in an increased demand for external evaluators to help develop monitoring and evaluation (M&E) frameworks and provide support with M&E capacity building[19].

A contextual factor discussed less in the Australian literature is that of evaluation affecting Aboriginal and Torres Strait Islander people. Since 2008 and the introduction of the Closing the Gap[20] framework there has an increased focus on evaluations of indigenous-specific programs. However, these evaluations have been criticised for their lack of 'rigour and independence' (Department of Finance and Deregulation 2010, p. 12). Moreover, Aboriginal and Torres Strait Islander communities have criticised many evaluations as being of limited usefulness: 'The evidence about what works, including for whom, under what circumstances, at what cost, and why, remains scant' (Empowered Communities 2015, p. 90). In 2019 the Australian Government's Productivity Commission was asked to develop a whole-of-government evaluation strategy for all Australian government agencies for programs affecting Indigenous Australians. As part of this strategy, the Commission is expected to establish a principles-based framework, identify priorities for evaluation and set out its approach for reviewing agencies' conduct of evaluations against the strategy. The United Nations (UN) Declaration on the Rights

[19] Personal communication, Anne Markiewicz (Australian evaluation consultant), 21 April 2018

[20] The Council of Australian Governments (COAG) pledged in December 2007 to close key gaps in outcomes between Indigenous and non-Indigenous Australians.

of Indigenous Peoples provides a further context for this work. The UN Declaration was signed by the Australian Government in 2009.

3.3 Program evaluation in New Zealand

In New Zealand there is also evidence of some evaluation being conducted in the post-war years, including a strong focus on evaluation in the agricultural sector (Lunt and Trotman[21] 2005). However, it was the economic and social upheavals of the 1970s that prompted a series of national committees, conferences, reviews and taskforces that in turn led to an increase in evaluation activity, particularly evident in the social and education sectors. For example, a 1976 Task Force on Economic and Social Planning recommended better evaluation of programs. This led to the establishment of the New Zealand Planning Council, which later noted that existing policies tended to go unchallenged while new proposals were more thoroughly interrogated (Lunt and Trotman 2005).

The New Zealand State Sector Act 1988 and the Public Finance Act 1989 were introduced, which placed greater emphasis on accountability, and NPM principles were embraced, including private sector styles of management and the contracting out of funded services (O'Flynn 2007). In the health sector, for example, between $2 billion and $4 billion is contracted to Not for Profit organisations each year to deliver services that were once delivered directly by government agencies[22]. A key difference between the two countries is that New Zealand does not have the same legislative requirement to report results of nationally funded programs. The impact of this on the demand for external (or internal) evaluation in New Zealand is not known.

There has, however, been increased legislative recognition of the principles of the Treaty of Waitangi since the 1980s. The Treaty is generally accepted as the founding document of New Zealand (The Department of Prime Minister and Cabinet 2008). While its articles are not enshrined in law, the principles have been included in various pieces of legislation. In terms of evaluation, the Treaty can be regarded as a defining point of difference between evaluation in New Zealand and

[21] Ian Trotman is a Fellow of the Australasian Evaluation Society (AES) and founder of the Wellington Evaluation Group (WEG).

[22] http://www.health.govt.nz/new-zealand-health-system/key-health-sector-organisations-and-people/non-governmental-organisations (downloaded 01 April 2017)

Mathea Roorda #579498

Australia. For example, Article Two of the Treaty refers to tino rangatiratanga, or the right to exercise self-determination, which in practice includes 'control over what is evaluated, how and by whom…' (Kerr 2012, p.8). Numerous guidelines for engaging Māori stakeholders and conducting culturally responsive evaluation and research were developed from 1998. These guidelines in turn have contributed to an emphasis on culturally responsive evaluation practices, for example selecting evaluators with the necessary cultural, language and subject competencies, engaging knowledgeable Māori stakeholders in evaluation planning and design, assessing Treaty of Waitangi implications of the evaluand, analysing data for Māori compared to non-Māori, and reporting results back to Māori stakeholders.

To support the development of more systematic culturally competent evaluation practice, advocates for a New Zealand-based professional association saw a need for an organisation that:

> would better serve and represent the unique needs, values, obligations and context of the Aotearoa New Zealand evaluation community (Aotearoa New Zealand Evaluation Association n.d.).

The Aotearoa New Zealand Evaluation Association was duly established in 2006. The following year the first cohort of evaluators completed New Zealand's first (and only) formal evaluation qualification from Massey University.

During the 2000s a new discourse entered the New Zealand public sector, epitomised by terms such as evidence-based policy making and evidence-informed policy (Gluckman 2013b; Smith and Sweetman 2009), and where evidence of 'what works' had a narrow focus on program consequences. The literature on evaluation would suggest this discourse extended beyond New Zealander's borders (Sanderson 2010; Vedung 2010).

Part of this discourse has also focused on what *counts* as quality evidence. For example, Sir Peter Gluckman has used his position as the Chief Science Advisor to the New Zealand Prime Minister to argue that experimental methods are better equipped 'to protect the objective nature of science' (Gluckman 2013b, p. 12). This positivist perspective on what counts as knowledge has not gone unchallenged.

Kaupapa[23] Māori evaluation approaches, for example, bring to the fore Māori philosophy and values.

Most recently, an alternative 'braided rivers' model has been promoted, that aims to draw on both Māori and western worldviews in evaluation and research contexts (Social Policy Evaluation and Research Unit and Social Policy Evaluation and Research Unit 2018). This conceptual view of different types of knowledge being separate but in dialogue explicitly recognises taken for granted assumptions about knowledge, for example, that there are alternative and equally important ways to understand the world. This acknowledgement of other epistemological ways of knowing is important, but also has its challenges as it is not just a case of braiding one methodology with another. Within both western and Māori paradigms there are likely to be multiple worldviews, and in the context of valuing, different perspectives on what matters.

3.4 Concluding comments

Program evaluation now has an established role in supporting government oversight and evidence-based decision making in both Australia and New Zealand. As noted, the demand for evaluation has been influenced by public administration policies and NPM thinking which have focused on program quality, efficiency and effectiveness. In New Zealand there has been an increased focus on cultural appropriateness of programs since the 1980s, while in Australia this focus is only now beginning to gain momentum.

On the supply side, while the AES's mission is to strengthen the theory, practice and use of evaluation, commentators including Perrin (2003) and Sharp (2003) have observed that in both countries evaluation practitioners tend to value practical knowledge over theoretical knowledge. For example, the most recent account of public sector evaluation in New Zealand is an edited book that focuses on practice-related topics such as managing evaluations, tools for working with stakeholders and getting evaluations used. The result is what Sharp (2003) has politely referred to as an eclecticism in evaluation methods and practice in Australian evaluation practice. Similarly, Nunns (2017) sought the opinion of evaluation experts to help

[23] The concept of 'kaupapa' refers to 'ground rules, customs, the right way of doing things'. The term 'kaupapa Māori' encapsulates Māori desire to 'affirm cultural philosophies and practices' (Pihama, Cram and Walker 2002, p. 30).

understand the shape of evaluation practice in New Zealand. One expert's perspective was less than flattering. They noted that evaluation practitioners in New Zealand placed less emphasis on evaluation theory, and that local practice tended to be 'informal…less systematic and lacking rigor…We have a much less rigid notion of what is involved in an evaluation' (Nunns 2017, p. 155).

3.5 Chapter summary

This chapter has described distinct historical and political factors that have informed how program evaluation has been conceptualised and practised in Australia and New Zealand. In both countries, program evaluation is relatively young, emerging as a distinct field of work in the mid to late 1980s, as a result of an increased need for public accountability. Until 2006 evaluators shared one professional association. A distinctive feature of New Zealand is its founding document, the Treaty of Waitangi. This has resulted in the development of Kaupapa Māori evaluation approaches and an expectation that all evaluators work in culturally responsive ways. In comparison, Australia has no similar Treaty, and a focus on a culturally responsive approach to program evaluation has only recently begun to gain momentum. In both countries, evaluation practitioners have tended to value practical knowledge over theoretical knowledge, resulting in practice that has been criticised as less systematic and lacking rigor. This information provides context for the findings in phase one of the study, specifically the way in which practitioners approach the development of criteria.

4 Methodology

4.1 Purpose of this chapter

The previous three chapters introduced the research problem, the research questions and positioned the research within the literature on evaluation reasoning and evaluation criteria. This chapter describes the research methodology, including the researcher's experience, the research design and strategy for conducting the study.

'Methodology' encompasses the 'strategy, plan of action, processes or design' (Crotty 1998, p. 3) that underpin decisions about how the research process proceeds. Methodology also incorporates the researcher's philosophical assumptions about the:

> nature of the social world; the nature of the knowledge one can legitimately have about this world; what is possible and what is important to know; value commitments regarding the purpose and location of our work in society; and the methods preferences that enable enactment of these assumptions (Greene 2002, p. 259-260).

The purpose of this chapter is to be explicit about the methodological choices made by the researcher in undertaking this study.

4.2 Researcher's context

In Chapter 1 the researcher introduced the evaluation context in which she has worked since 1998. Her academic training and her experience as an evaluator inform how she conducts research. Ontologically, she embraces the idea of multiple realities that are in turn shaped by different social, political, cultural, economic, ethnic and gender perspectives. A brief introduction to the researcher's Pākehā[24] context is provided here. Its purpose is to illustrate how context is multi-layered and to show that the researcher conducts her craft from within a particular social construction that is not fixed.

A fundamental question for a researcher working in New Zealand and Australia is: 'What does it mean to be Pākehā or a white person engaged in an activity that

[24] New Zealander of European descent

impacts on indigenous peoples?' The researcher's insight into her cultural values did not start with this study (see Roorda and Peace 2008; Torrie, Roorda, Peace, Dalgety and Bailey 2015) but deepened as the researcher engaged with the literature on normative ethical perspectives (Chapter 8). She has come to understand that philosophical perspectives such as duty based or consequentialist are not abstract concepts confined to texts but integral to how she views 'good' and 'bad' in different situations, and familiar elsewhere in the everyday world around her, for example in her relationships with family and friends. In this way she now understands that to be Pākehā or a white researcher is to think and act in ways that are culturally connected to particular philosophical origins, sometimes referred to as 'western'.

Being a Pākehā/white researcher also locates the researcher in a historical and political context. In both Australia and New Zealand there has been a long history of doing research and evaluation 'on' and 'to' Māori (New Zealand's indigenous population) and Aboriginal and Torres Strait Islander peoples (Australia's indigenous population). This led to Māori, Aboriginal and Torres Strait Islander peoples being wary about research and researchers. In New Zealand public sectors, there is now more explicit recognition that programs, whether economic, environmental or social, have a cultural dimension and are bound by obligations under the Treaty of Waitangi (Moewaka Barnes 2003). There are expectations on Pākehā researchers and evaluators to be culturally competent (Superu and Aotearoa New Zealand Evaluation Association 2015). Likewise, in the Australian public sector, there is greater acknowledgement that evaluation activities must be respectful, and take into account the values, of Aboriginal and Torres Strait Islander peoples. For example, the Australian Government's Indigenous Advancement Strategy Evaluation Framework (2018) notes:

> High quality evaluation incorporates Indigenous perspectives in refining knowledge about outcomes and in broader framing of the definitions of program success. Credible evaluations, underpinned by ethical practice, generate high quality evidence that can be used for policy and program design into the future (Department of Prime Minister and Cabinet 2018, p. 8).

The journey to cultural competence is ongoing, personal and challenging. For the researcher, the process has been dialogic; being conscious of one's own values and beliefs, listening carefully and acknowledging that there are multiple ways of seeing and understanding the world (Torrie et al. 2015).

The researcher's beliefs about knowledge align with constructivism. Constructivists argue that knowledge is socially constructed, dynamic, culturally and historically contingent (Guba and Lincoln 1989; Sanderson 2002; Schwandt 1998). A constructivist conception of research requires researchers to be explicit and systematic with regards to how context and values influence an inquiry (Dahler-Larsen 2001). The researcher's role is to bring these multiple perceptions 'into conjunction…with one another and with whatever information…can be brought to bear on the issues involved' (Guba and Lincoln 1989, p. 142).

4.3 Initial conceptualisation of the study

This section includes a description of the initial research design which was planned as a mixed methods study consisting of a preliminary literature review followed by a three-phase Delphi study. Also documented are the reasons for not proceeding with this design but choosing instead to explore the topic via alternative methods.

4.3.1 Review of the literature

As noted in Chapter 1, this study began with a central question: what are the characteristics of defensible criteria for Australian and New Zealand program evaluations? The first step in the study was a review of the evaluation literature. The literature identified theoretical perspectives on the development of criteria but revealed an overall paucity of information about the development of criteria. Hence, an initial strategy was to draw on the knowledge of experienced evaluators to better understand what defensible criteria look like, and how they are managed, in the context of Australian and New Zealand program evaluations.

4.3.2 Delphi technique

An adaptation of the Delphi technique was identified as a useful method for gaining insight into the research problem. Developed by the RAND Corporation in the 1950s to solicit and integrate the collective wisdom of a panel of experts (Dalkey and Helmer 1963) this technique generally involves an iterative process of data collection and analysis designed to converge the opinions of experts into group

consensus through a series of rigorous questionnaires interspersed with controlled feedback (Okili and Pawlowski 2004).

The success of the Delphi technique relies on careful selection of the expert panel. A selection criterion for Delphi is that respondents have knowledge and experience with the issues under investigation, as noted by Sobaith, Ritche and Jones (2010). The researcher sought to meet this criterion by recruiting evaluators with at least five years' experience conducting or managing program evaluations.

The first round of the Delphi study proceeded through administration of an online survey. The findings provided useful information to inform an understanding of the research problem but also indicated that many respondents were not knowledgeable about the development of evaluation criteria. It was decided that engaging respondents in a second round would not advance an understanding of the research problem. The findings from round one of the Delphi study were therefore reconceptualised as an online survey of evaluators and the first phase of a mixed methods study.

4.4 Mixed methods strategy for this study

This section describes the mixed methods strategy that subsequently informed this study. A brief overview of mixed methods is provided, followed by a discussion of how mixed methods were used in this study.

Mixed methods research (MMR) is an approach used to gather:

> quantitative (closed-ended) and qualitative (open-ended) data, integrating the two and drawing interpretations based on the combined strengths of both sets of data to understand research problems (Creswell 2015, p. 2).

Mixed methods may be used for a range of reasons, including that initial results of a study need to be further explained or investigated, or that the overall research objective is best addressed with multiple phases. The use of triangulation, through enhancing their credibility and trustworthiness by comparing different types of evidence from different sources, can help to overcome the limitations of relying on one method, or it can help the researcher to gain a deeper understanding of a phenomenon. Both these purposes are evident in the following definition, in which mixed methods are viewed as an approach that intentionally integrates elements of qualitative and quantitative methods (for example, at the data collection, analysis

and reporting stages) 'for the purposes of breadth and depth of understanding and corroboration' (Johnson, Onwuegbuzie and Turner 2007, p. 123).

Mixed methods designs can be distinguished in several ways, including the degree to which methods interact with each other. For example, a study that comprises different methods may be mixed during data collection, data analysis and/or at the interpretation stage of a study. A mixed methods study may also be implemented concurrently, sequentially, or a combination of both. Relative status given to quantitative and qualitative methods is a distinguishing feature of mixed methods. In a study, either one method is dominant, or they are given equal status.

Traditionally, research questions emerge in response to problems identified within the literature and practice (or both), and it is these questions that then determine the methods to be used in the study. However, Plano Clark and Badiee (2010) identify circumstances in which new or additional questions may be incorporated at a later stage of a study. Examples of such circumstances include complications with implementing a particular method, an unexpected finding or a new development within the context of the study.

4.5 Design of the study

The current study was conceptualised as a two-phase emergent mixed methods design, with the second phase responding to what was learned in the first phase of the study. This section begins with a brief overview of the study design and then provides more detail about each phase.

4.5.1 Overview of the study design

As illustrated in Table 4, the study began with a central research question and one sub-research question (RQ#2) in phase one. The aim of this sub-research question was to understand how evaluators in Australia and New Zealand manage criteria for program evaluations. This question was addressed through an online survey of program evaluation practitioners in Australia and New Zealand (Chapter 5), and a review of published Australian and New Zealand program evaluations (Chapter 6). Both the online survey and document review were conceptualised as concurrent studies, conducted and analysed independently of each other. At the analysis stage the findings were in discourse with each other, with the findings from the review of evaluation reports triggering further, and deeper analysis of the online survey

results. In Chapter 7 key findings from both studies are considered together for the purposes of expansion, that is, to extend the breadth of the inquiry and provide a more comprehensive understanding of the research topic (Greene, Caracelli and Graham 2008).

The status of analysis differed for each study. As illustrated by the notations[25] in Table 4, a quantitative analysis approach was dominant in the online survey and a qualitative analysis approach was dominant for the review of evaluation reports.

Table 4: Overview of the study design

Phase	RQ#	Data sources	Methods	Analysis	Location in study
Both	1	Literature on evaluation and normative ethics	Document review	QUAL Thematic analysis	PART ONE 2
Phase 1	1, 2	Program evaluators (Australia and New Zealand)	Online survey	QUAN/qual Descriptive statistics Factor analysis Thematic analysis	PART TWO 5
		Program evaluation reports (Australia and New Zealand)	Document review	QUAL/quan Content analysis Thematic analysis	PART TWO 6
Phase 2	1, 3	Conceptual framework and development of a tool			PART THREE 8
		Ethics scholars and practitioners (Australia and New Zealand)	Document review	QUAL Thematic analysis	PART THREE 8
		Program evaluation practitioners and case examples	Qualitative interviews Document review	QUAL Thematic analysis	PART THREE 9
Central research question: RQ#1 What is required to develop a defensible set of evaluation criteria for program evaluations? **Sub-research questions:** RQ#2 How are criteria managed in Australian and New Zealand program evaluations? RQ#3 What is an evidence-informed conceptual framework for defensible criteria that can be operationalised for evaluation practice in the public sector?					

[25] The notations QUAN and QUAL designate the priority given to quantitative and qualitative analysis in each component of the study. QUAL indicates a qualitatively driven focus, whereas QUAN denotes a quantitatively driven focus. Lower case letters imply less emphasis, e.g. a component of the study that is QUAL/quan designates the qualitative data has major emphasis and the quantitative less emphasis.

Findings from the online survey and review of evaluation reports indicated a gap between theory on evaluation criteria and the development of criteria in practice. This resulted in the formation of RQ#3 which aimed to develop an evidence-informed conceptual framework for defensible criteria that could be operationalised for evaluation practice in the public sector.

In phase 2, a conceptual framework was developed, informed by literature on normative ethics. This was then extended into a practice-based tool to support evaluators develop defensible evaluation criteria for program evaluations. A criteria matrix handbook was developed to support evaluators to engage with the tool. The handbook was reviewed and then field tested. A qualitative analysis approach was used for both the review of the tool and field testing. The methods used to undertake both these tasks are reported in Chapters 9 and 10 respectively.

At the conclusion of phase two the findings from all components of the study were considered in the context of existing literature, with the intention of informing the central research question (RQ#1) and introducing new and fresh insights into the research problem (Greene et al. 2008).

Sections 4.5.2 to 4.5.5 introduce the methods used in each part of the study. These are subsequently expanded as follows:

- Practitioner perspectives on developing criteria: Chapter 5
- Criteria in evaluation reports: Chapter 6
- Development of a conceptual framework and a criteria matrix handbook: Chapters 7, 8 and 9
- Practitioner engagement with normative ethics: Chapter 10.

4.5.2 Survey of evaluators (Chapter 5)

The first component of the study comprised an online survey of 137 Australian and New Zealand evaluators. The survey informs an understanding of both RQ#1 and RQ#2. As noted in the review of literature (2) criteria are informed by sources. Thus, one aim of the survey was to identify the sources practitioners commonly used to identify dimensions of merit, the sources they considered important, and the factors that influenced decisions about which sources were used. The survey also explored practitioners' views about the characteristics of defensible criteria

for program evaluations. The survey comprised both closed and open-ended questions and quantitative and qualitative analyses. Further information about the survey design, including the sampling frame and approach to analysis, is included in Chapter 5.

4.5.3 Document review (Chapter 6)

The second component of the study was a review of 47 published evaluation reports using qualitative and quantitative data analyses. Again, the purpose of the review was to inform both RQ#1 and RQ#2. The sampling strategy for the document review is discussed in Section 6.2.1.

Meta-evaluation, an approach used to systematically review evaluation reports (Cooksy and Caracelli 2009) was considered for this part of the study. However, the focus was not to critique evaluation reports, but to describe what criteria looked like in current practice. A decision was therefore made to use evaluation reports as a data source and to examine them using constant comparative analysis, content analysis and thematic analysis. Each of these analyses is discussed in Section 0.

The online survey and the document review complement each other. The survey findings gave the researcher access to practitioners' views and perspectives, which provided insight into their decision-making process for identifying and selecting criteria. In contrast, information about the evaluators' decision making with regard to criteria was rarely evident in the review of published reports. Few reports, for example, documented the sources used to identify and define criteria. However, the reports did provide the opportunity to examine criteria in the context of specific evaluands, something that was not possible in the survey of practitioners. Thus, triangulation served to overcome the limitations of individual methods and it served the purpose of providing a deeper understanding of the phenomenon of criteria development in program evaluations.

The findings from the online survey and the review of evaluation reports provided valuable insight into the management of criteria in current program evaluation practice (RQ#2). The review of evaluation reports particularly informed RQ#1, serving to highlight the importance of criteria that include all relevant dimensions of value.

4.5.4 Developing a conceptual framework and a criteria matrix handbook (Chapters 8 and 9)

Analysis of the online survey and evaluation reports, alongside the literature, identified a gap between theory and practice. RQ#3 aimed to bridge this gap by developing an evidence-informed conceptual framework for defensible criteria that could be operationalised for evaluation practice in the public sector. The framework is informed by literature on normative ethical perspectives. This literature is not new to evaluation scholars; a number explicitly draw on specific to ensure a focus on equity and justice (e.g. Greene 1997; House and Howe 2000; Mertens 2009; Mertens, Bledsoe, Sullivan, and Wilson 2010). However, few, if any, scholars consider the three dominant normative perspectives of consequentialism, deontology and virtue-based ethics as a way to systematically consider dimensions of merit from a range of perspectives. The researcher believed considering them together might assist in capturing multiple value perspectives on understanding concepts of quality and success that could then inform criteria.

The researcher had to bridge a gap between the conceptual framework and a tool that could be used by practitioners. The solution was a handbook, which provided step-by-step instructions for identifying dimensions of value based on normative perspectives and then using these to develop a justified set of criteria.

The handbook was reviewed by drawing on the Evaluation Checklists Project validation process. First, two ethics experts reviewed a set of statements about normative perspectives and principles. Second, the draft handbook was reviewed by nine Australian and New Zealand evaluation practitioners. The researcher analysed their feedback and made changes as appropriate, in readiness for the tool to be field tested.

4.5.5 Practitioner engagement with a normative framework (Chapter 10)

The final component of the study also informs RQ#3. Three evaluation practitioners field-tested the criteria matrix tool in the context of one of their current evaluations. Specifically, the intention was to explore what happened when evaluation practitioners engaged with normative perspectives to identify dimensions of value.

Field testers were asked to provide their completed criteria matrix and set of criteria to the researcher. They were also asked to reflect on the process of using the tool by providing written responses to a set of five questions (see 'Fieldworker feedback', located at the end of the criteria matrix handbook in Appendix F) and through participating in a follow-up qualitative interview.

The material developed by each field tester (referred to as case examples) was analysed along with field testers' written and verbal reflections of using the tool. A broader analysis of the conceptual model and field testing of the tool also informs the central research question.

4.6 Ethical considerations

The project was approved by the Melbourne University Human Research Ethics Committee (Ethics ID 1340793.1). A request for amendment to the study was approved for phase two of the study.

Procedures for protecting participants' rights and identities were followed as required for conducting ethical research. Survey participants were provided with information in a written Plain Language Statement. Participants could access additional information about the study by contacting the researcher or the Executive Officer, Human Research Ethics, at the University. No requests for additional information were requested of the researcher and no complaints were received.

All identifying information about participants, including email addresses, is kept on a computer available by password only. According to ethical requirements, data will be retained for five years after thesis completion and then destroyed.

4.7 Chapter summary

This chapter has described key philosophical assumptions underpinning the study. It outlines the rationale for a mixed methods design and discusses an overview of the research design, data sources and methods. More detail about the methods used in each component of the study, including data collection procedures, sampling and analysis, is included at the start of Chapters 5, 6 and 7 (phase one) and Chapters 8, 9 and 10 (phase two).

PART TWO: CRITERIA IN CURRENT EVALUATION PRACTICE

Part two of the thesis comprises phase one of the study, which has a focus on management of criteria in current Australian and New Zealand program evaluation practice. Two research questions are investigated:

RQ#1 What is required to develop a defensible set of evaluation criteria for program evaluations?

RQ#2 How are criteria managed in Australian and New Zealand program evaluations?

Two perspectives on criteria in current evaluation practice are presented. The first – practitioner perspectives on developing criteria – arises from the findings of an online survey of 137 evaluators working in program evaluation in Australia and New Zealand (Chapter 5). The second – criteria in evaluation reports – arises from the findings of a review of 47 published evaluation reports written or commissioned by public sector agencies in Australia and New Zealand (Chapter 6).

5 Practitioner perspectives on developing criteria

5.1 Purpose of this chapter

This chapter examines the development of evaluation criteria from the perspective of 137 Australian and New Zealand evaluation practitioners who responded to an online survey. Evaluators with at least five years' experience were recruited through the AES and ANZEA and the researcher's evaluation networks in Australia and New Zealand. The chapter begins with a discussion of the survey design, sampling protocol, survey responses, analysis, after which the findings are presented and discussed.

5.2 Research objectives

The specific research objectives of this component of the study were to:

- investigate how evaluators in Australia and New Zealand identify criteria
- examine factors that influence the way criteria are developed and justified in Australian and New Zealand program evaluations.

5.3 Survey design

5.3.1 Overview of the design

The survey was developed using SurveyMonkey, a web-based survey tool. The survey comprised 12 closed and open-ended questions (Appendix A). As illustrated in Figure 3, the questions were designed to explore any differences between how evaluators managed the development of criteria in practice and what they considered ought to happen.

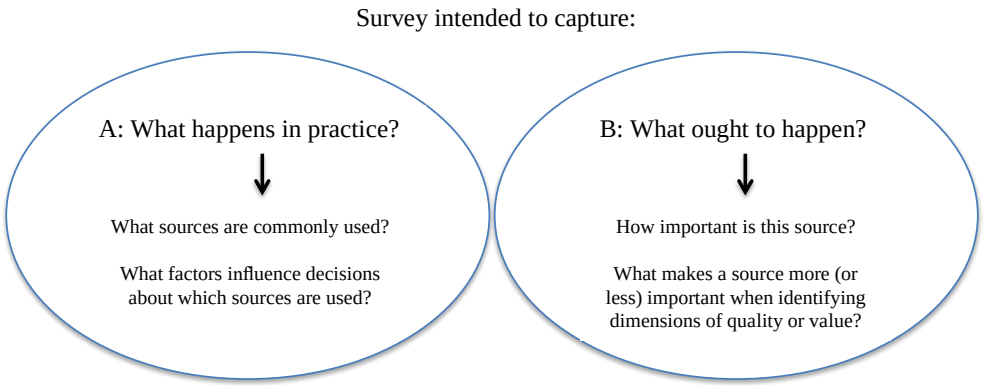


Figure 3: Overview of survey focus

5.3.1.1 *Questions 1–5*

The first five questions focused on obtaining demographic information. One purpose of these questions was to provide a framework for later analysis, to explore differences and similarities between evaluators working in a range of contexts. A second purpose was to include only participants who had experience in the public sector. This purpose is discussed further in Section 5.3.3.

5.3.1.2 *Questions 6–11*

Question 6 asked respondents to identify whether they assessed against key evaluation questions, criteria of merit, performance standards and/or something else. This question was intended to test the assumption that not all evaluators assessed against explicit criteria. In an implicit way, this question also acknowledged, to respondents who did not routinely include criteria of merit in their evaluations, that assessments could be made in less explicit ways and that their experience and views were also being sought.

The key question of interest to this study (Question 7) began with a preamble and instruction as follows.

> **Sources used to decide what to focus an evaluation on**
>
> Below is a list of potential sources evaluators can use to identify the dimensions that will distinguish (for example) a good or well-functioning program/policy from one that is poor.
>
> Please rate these items as follows. (A) Reflect on the evaluations you are involved with and identify the most common sources used to decide dimensions of quality and value. (B) Give your opinion about importance of these sources. Are there some you regard as more important than others?

For Part A (how often they used each source) respondents could select from: 'never', 'seldom', 'often' or 'always'. For Part B (importance of a source) the options were: 'unimportant', 'less important', 'important' or 'very important'. Responses were required for all 17 items in Part A and B.

In Question 8, respondents were asked to add any sources missing from the options provided in Question 7. The final open-ended questions (Questions 9–11) asked respondents to identify factors that influenced decisions about which sources they used, as well as their views about what made a source more (or less) important.

Table 5 shows the alignment of the survey questions to the key research questions guiding the study.

Table 5: Alignment of survey questions to the key research questions

Survey instrument	Key research questions for the study
Survey questions 6, 7, 8, 9	RQ# 2: How are criteria managed in Australian and New Zealand program evaluations?
Survey questions 7, 10	RQ#1: What is required to develop a defensible set of evaluation criteria for program evaluations?

5.3.2 Sources of criteria

A list of potential sources for identifying dimensions of value was identified from a review of academic peer reviewed articles, evaluation texts and grey[26] literature on evaluation, including public sector guidelines sourced online. Search terms used to find potential sources included 'criteria', 'values' and 'success'. In evaluation texts and grey literature, the researcher reviewed sections relating to evaluation design to identify potential sources. The review continued until the point of saturation, when no new items could be added to the list of broad concepts.

The sources were initially framed as direct quotes lifted from the literature. As part of the process of creating a succinct list, they were rewritten as phrases to facilitate respondents being able to rate a relatively long list of items. For example, the United States Government Accountability Office (2011) included a statement that [evaluators must draw on]

> the laws, regulations, contracts, grant agreements, standards, measures, expected performance, defined business practices and benchmarks against which performance is compared or evaluated (p. 76).

This statement was shortened to the phrase: 'Government legislation, standards, regulations'.

Five experienced[27] New Zealand evaluators with public sector experience reviewed an initial list of potential sources. Sources considered too similar were removed or reworded, thereby reducing the initial list to independent sources. Following this,

[26] Grey literature is defined here as government, non-government and academic literature in print and electronic formats, but not controlled by commercial publishers.

[27] Each evaluator had between 8 and 10 years' experience evaluating public sector initiatives.

three PhD colleagues, two PhD supervisors and an experienced evaluator[28] also reviewed the list. The final list of sources comprised 17 items. The list of sources included in the survey is outlined in Table 6, along with references to the literature from which they were derived.

Table 6: Sources for developing criteria as identified in the literature

Sources	References
Program, policy objectives	Department of Finance Australia, (1994); Arens (2005); Hertting and Vedung (2012); Lachausse (2011); O'Connor-Fleming, Parker, Higgins and Gould (2006); United States Government Accountability Office (2011)
Evaluation contract manager	Personal conversation with evaluation colleagues, (March 2014)
Stakeholders affected by the program/policy	Department of Finance Australia, Department of Finance Australia (1994); Greene (1997); Guba and Lincoln (1989); Oakden (2013a; 2013b); Thomas and Palfrey (1996)
Literature on best practice	Department of Finance Australia (1994); Lachausse (2011)
Content expertise	United States Government Accountability Office (2011)
Ethical guidelines or standards	Ministry of Social Development (2004)
Lobby groups	ACT Government (2010; Department for Communities and Local Government (2009)
Cultural expertise	Aotearoa New Zealand Evaluation Association (2011); Julnes (2012); Ministry of Social Development (2004); (Kawakami, Aton, Cram, Lai and Porima 2008).
Policy documents	Department for Communities and Local Government (2009)
Evaluation commissioner	ACT Government (2010); Anderson (1979)
Program staff	Brandon (1998); Thomas and Palfrey (1996)
Policy staff responsible for developing the initiative	ACT Government (2010)
Government's strategies	Greene (2009, p. 53) Department of Finance Australia (1994; Stufflebeam (2001)
Government legislation, standards, regulations	Department for Communities and Local Government (2009; Dodaro (2011); Shipman (2012); Stufflebeam (2001)
Political minister	ACT Government (2010); Amies and Department of Finance Australia (1994); Thomas and Palfrey (1996)

[28] This evaluator had more than 10 years' experience evaluating public sector initiatives.

Sources	References
Evaluator	European Commission (2004); House (1980a)
Needs assessment	Amies and Department of Finance Australia (1994); Davidson (2005); Lachausse (2011); O'Connor-Fleming et al. (2006); Scriven (2011); Stufflebeam (2001)

5.3.3 Sampling strategy

To be eligible for inclusion, participants had to be knowledgeable about public sector evaluation practice. The study also aimed for diversity so that similarities and differences could be explored between practice in Australia and New Zealand and by the evaluator's workplace. Do evaluators, for example, approach the selection of criteria in a different way if they work inside or outside a public sector agency? The selection criteria and rationale are outlined in Table 7.

Table 7: Rationale for the survey selection criteria

Selection criteria	Rationale for selection criteria
Evaluation experience in the public sector	After five years, evaluators would be likely to have a reasonable understanding of the public sector context in which they were working and have gained practical experience developing and undertaking evaluations. Practitioner experience was considered more important than evaluation-related academic qualifications, as the intention was to identify what happened in practice.
Mix of managers and evaluation practitioners working in different settings	The intention was to explore differences between evaluators working in a range of contexts, including different workplaces, evaluation roles and countries.

A two-pronged recruitment approach was considered most appropriate for identifying potential respondents. The first involved inviting members of the AES and ANZEA to participate in the survey. At the time the survey was conducted (May 2014), the AES had approximately 1,000 members and ANZEA 360 members[29]. Neither professional organisation was able to provide demographic data on members' primary workplace or how many years they had worked in the field of evaluation. It was therefore not possible to target members who matched the inclusion criteria for the survey. This limitation also meant it was not possible to

[29] Administrative staff at the AES and ANZEA offices provided the researcher with information about the number of people with membership to their respective organisation.

estimate the response rate against the number of AES and ANZEA members who worked in the public sector and had had at least five years' program evaluation experience.

Members of both associations were invited to participate in the survey via an information sheet (Appendix B). For privacy reasons the information sheet was prepared by the researcher and then emailed to members of the two professional evaluation associations by their internal administrative staff. Both professional organisations included New Zealand members, so it is likely some evaluators received more than one invitation.

The second recruitment strategy involved sending a personalised email to 59 of the researcher's evaluation colleagues[30] in Australia and New Zealand, both to encourage them to participate in the survey and to request they forward the information sheet to other evaluation colleagues who might not be members of either the AES or ANZEA. This strategy was considered appropriate as the researcher has longstanding relationships with evaluators in both Australia and New Zealand.

5.3.4 Survey responses

During the 15 days the survey link was open there were 197 responses. Multiple responses per computer were allowed and a check of IP addresses showed no duplicates. Twenty-five respondents said they had less than five years involvement in evaluation and so were automatically taken to an exit page where they received the following message:

> Thank you for your willingness to participate in this survey. Unfortunately you do not qualify to complete the questionnaire. We greatly appreciate your interest and your willingness to share your expertise and opinions.

A total of 137 respondents met the experience threshold and completed the compulsory questions, that is, all questions up to and including Question 7. Of these respondents, 80 (58 percent) worked in Australia and 57 (42 percent) in New Zealand. Overall 68 percent of respondents had 10 years or more evaluation experience, as shown in Figure 4

[30] Emails were sent only to colleagues personally known to the researcher.

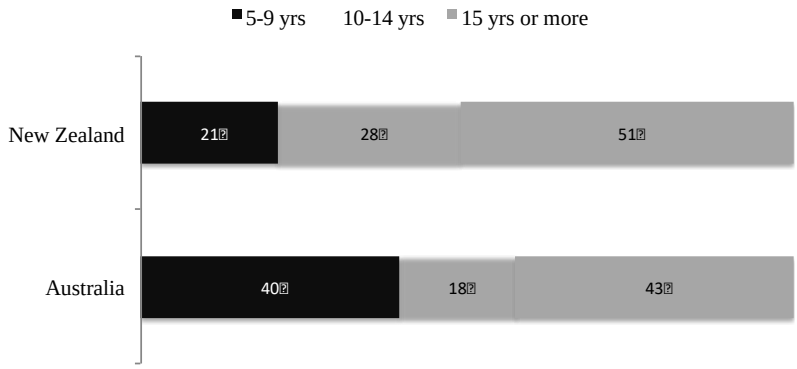


Figure 4: Respondents' evaluation experience (%)

The majority of respondents worked in either a government agency or in private practice[31] Eleven respondents worked in more than one organisation and were therefore coded as working in multiple workplaces, to facilitate exploration of differences by sector.

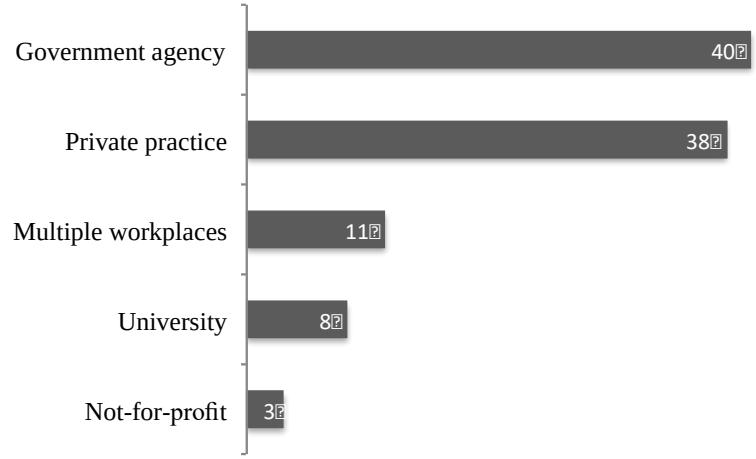


Figure 5: Respondents' workplace (%)

Program evaluation was the primary work focus for 85 percent of respondents, as shown in Figure 6. Respondents were asked also about their role within an organisation. Almost 90 percent of respondents indicated they conducted evaluations and 45 percent[32] commissioned evaluations.

[31] The term private practice is used here to refer to evaluators who are self-employed contractors or work in privately owned companies.

[32] The percentages do not add to 100 as some respondents both conduct and commission evaluations.

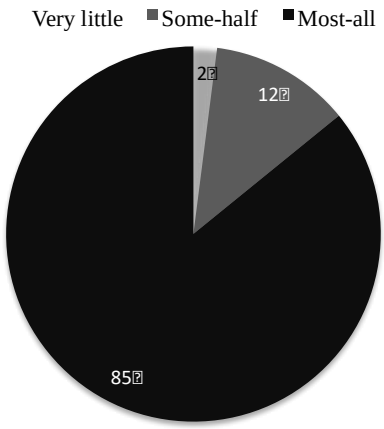
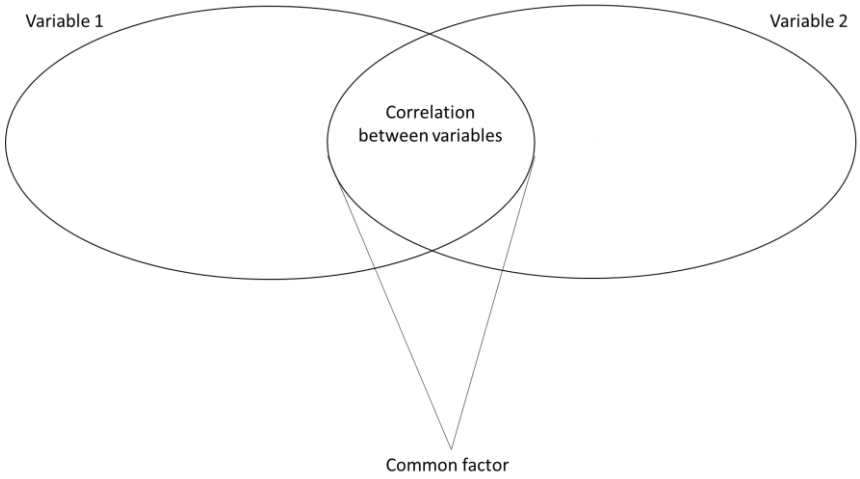


Figure 6: Extent to which evaluation is a focus of respondents' work (%)

5.3.5 Data analysis

The data set was exported from SurveyMonkey and 'cleaned' in Excel. This involved removing 35 respondents who did not complete the compulsory questions. The survey analysis began with descriptive statistics using SPSS, a statistical analysis program. This included frequencies for each of the closed-response questions and cross tabulations for Question 7 (use versus importance). Question 7 was then analysed against demographic data. The approach was purposefully exploratory, to allow themes, contradictions and puzzles to emerge.

An exploratory factor analysis (EFA) was then conducted to identify any underlying patterns or factors that could explain the interrelationships between the 17 sources. The idea behind factor analysis is that for a collection of variables there is an underlying set of variables called factors. In the EFA conducted for this study, correlations between the 17 variables (sources) were sought, and where this occurred, a common factor was proposed.



Source: Punch (2005, p. 126)

Figure 7: Exploratory factor analysis

Used commonly in the fields of education and psychology, factor analysis is considered a particularly useful method for interpreting self-reporting questionnaires as it provides construct validity of self-reporting scales (Williams, Onsman and Brown 2010). It is also useful for theory generation.

Although sample size is important, there are varying opinions about an appropriate minimum sample size for an EFA. For example, Comrey and Lee (1992) identified N=100 as poor, N=300 as good, and N=1000 or more as excellent. Others argue that a sample size of 50 or 100 is adequate (Hair, Anderson, Tatham and Black 1998; Sapnas and Zeller 2002). The sample size in this study was 137, thus poor according to Comrey and Lee (1992) and adequate according to other authors.

In this study, principal components analysis was used as the extraction method. The Kaiser Rule assisted in determining factor extraction. This rule is a commonly used procedure for determining the number of factors to extract, recommending that any factor with an eigenvalue greater than one be considered significant (Williams et al. 2010).

5.3.5.1 *Qualitative data*

The responses to the open-ended questions in the survey were coded using Nvivo, a qualitative data analysis program. Responses were coded initially by question number, and then into broad themes as they emerged. For example, Question 10 asked: What makes a source more (or less) important when identifying dimensions of quality or value/evaluation criteria? Responses were first analysed to determine

what ideas were being conveyed and to label each new idea with a code. The aim was to develop provisional categories grounded in the data rather than assign data to existing (top down) categories.

Some responses included multiple ideas and were thus coded to more than one category. As eight respondents did not respond to Question 10, the analysis is based on 129 responses. Once initial coding was completed, responses in each category were checked, by mentally asking: 'Is this the same broad idea as the previous response/s or a different idea?' As a result, some responses were coded to additional categories or shifted to a broader category. Sub-themes within a category were also identified.

Table 8: Illustration of coding framework for Question 10

Example responses	Provisional category *			
	A	B	C	D
The end users (policy makers, operational staff, ministers) of the evaluation are very important. Understanding what they want to do with the information shapes what we then look at. We also consider what type of information they will see as credible.	1			
The Government wants to know whether it has done what it said it would do…	1			
I aim to consider competing perspectives and question dominant views.				1
Sources that directly and logically relate to the intervention. From this, it is a matter of identifying those groups most affected by the intervention (especially negatively affected) to determine where harm might arise.		1	1	

*Provincial codes: (A) Managerial. (B) Relevance. (C) Justice and equity. (D) Democratic.

The data coded to each category was then analysed, guided by the question: What do the responses (collectively) reveal about the category? This procedure was repeated for each category.

5.4 Findings

The survey findings are discussed in four sections. The first section examines how respondents make evaluative assessments. The second and third sections identify sources commonly used to develop dimensions of quality and factors that can determine which sources are used in practice. The final section includes an analysis

of sources that respondents consider to be more, or less, important for shaping how they identify dimensions of quality.

5.4.1 How respondents make evaluative assessments

As noted in Chapter 2, criteria should arguably be evident in every evaluation where an evaluative judgement is being made. To test this assumption, the survey asked respondents to indicate whether they made assessments against key evaluation questions, and/or criteria of merit and/or performance standards. The results are in Table 9.

Table 9: What respondents make assessments against

Respondents assess against:	No.	Percentage
Key evaluation questions	132	96
Criteria of merit	61	45
Performance standards	56	41

Almost all respondents (96 percent) said they used key evaluation questions to guide their assessments. Of these respondents, 42 percent indicated they used key evaluation questions without subsequently developing criteria or performance standards. Only 45 percent of respondents said they used explicit criteria of merit, many in conjunction with key evaluation questions and performance standards.

One possible explanation for these findings is that some evaluators do not consider judgement to be essential to evaluation. Such a view certainly seemed to align with this respondent's comment:

> I don't work this way (developing criteria). I use a hermeneutic approach. I also don't judge the program, I answer the questions agreed with the Commissioner.

One other respondent also admitted to not reaching evaluative judgements in their work:

> (I) feel uncomfortable to admit we don't often go as far as assessing if our projects are good, or good enough.

While their reason for not reaching evaluative judgements is not known, the comment suggests this may be an acknowledged weakness in their practice, rather

than a decision to not assess being based on a fundamental difference in how they understand evaluation.

From the qualitative comments included in the survey responses, it was clear that some practitioners assessed against criteria of some form, but did not refer to them as such. Criteria may be implicit as illustrated in the following respondents' comments:

> Indicators are derived from known expected patterns (but not necessarily in the form of criteria of merit).

> Ideally questions are linked to the theory of change or logic and developed with key partners.

Some respondents used the terms criteria and standards interchangeably or confused one with the other. For example:

> I tend to develop an overview of the range of outcomes we're looking for and related indicators that we are expecting to change but do not identify evaluation criteria for what no, low, medium, high shifts might look like. This emerges more from the data. My experience has been that people expect much larger changes than actually happen. Therefore I would not use this type of pre-determined criteria unless it is backed up by prior research or evaluation, which shows that the expected changes are reasonable.

5.4.2 Sources commonly used in evaluation practice

Respondents were asked to reflect on public sector evaluations they had been involved with and to identify the most common sources they used to decide dimensions of quality and value. They were also asked to identify the factors that influenced decisions about which sources were used. The way the questions were phrased purposefully took into account that respondents might not use explicit criteria in their evaluation practice.

This section presents the analysis of these two questions. The findings begin with a broad overview of sources commonly used, or not, when developing public sector evaluations. Next, similarities and differences between respondents working in different geographic and workplace contexts are described.

As noted earlier, the survey included a list of 17 sources that might be used to identify relevant criteria for a public sector evaluation. Respondents were asked to rate the extent to which they commonly used that source by selecting from one of four responses: always, often, seldom or never.

Two sources, political minister and lobby groups, were seldom or never used by more than two-thirds of respondents (Figure 8). All other sources were used (always or often) by at least 57 percent of respondents. Of all the sources, program/policy objective/s was the only one that more than three quarters (86 percent) of respondents always used. Needs assessment was a source where there was divergent use: 57 percent said they often or always used this source, while 43 percent said they seldom or never used it to develop criteria.

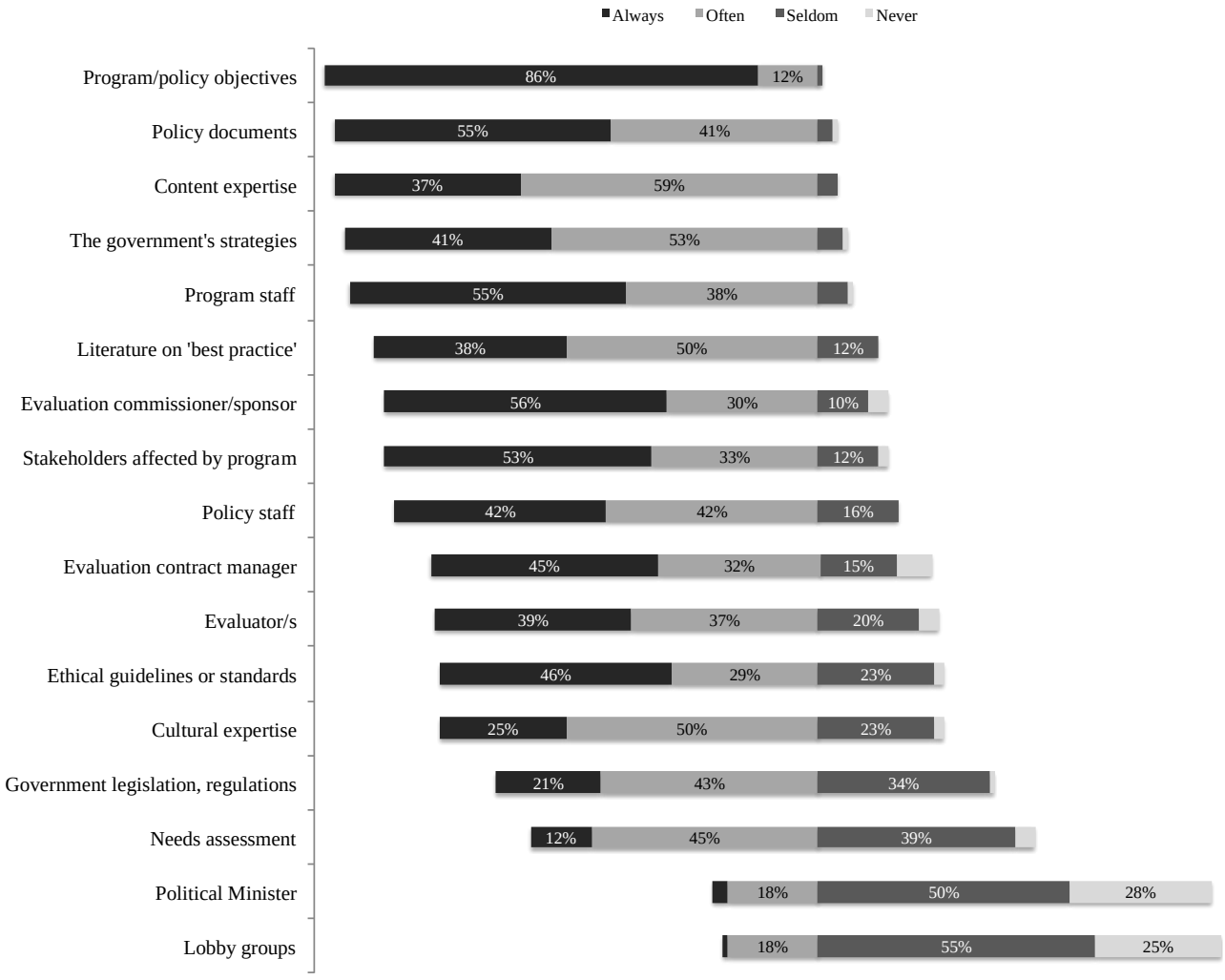
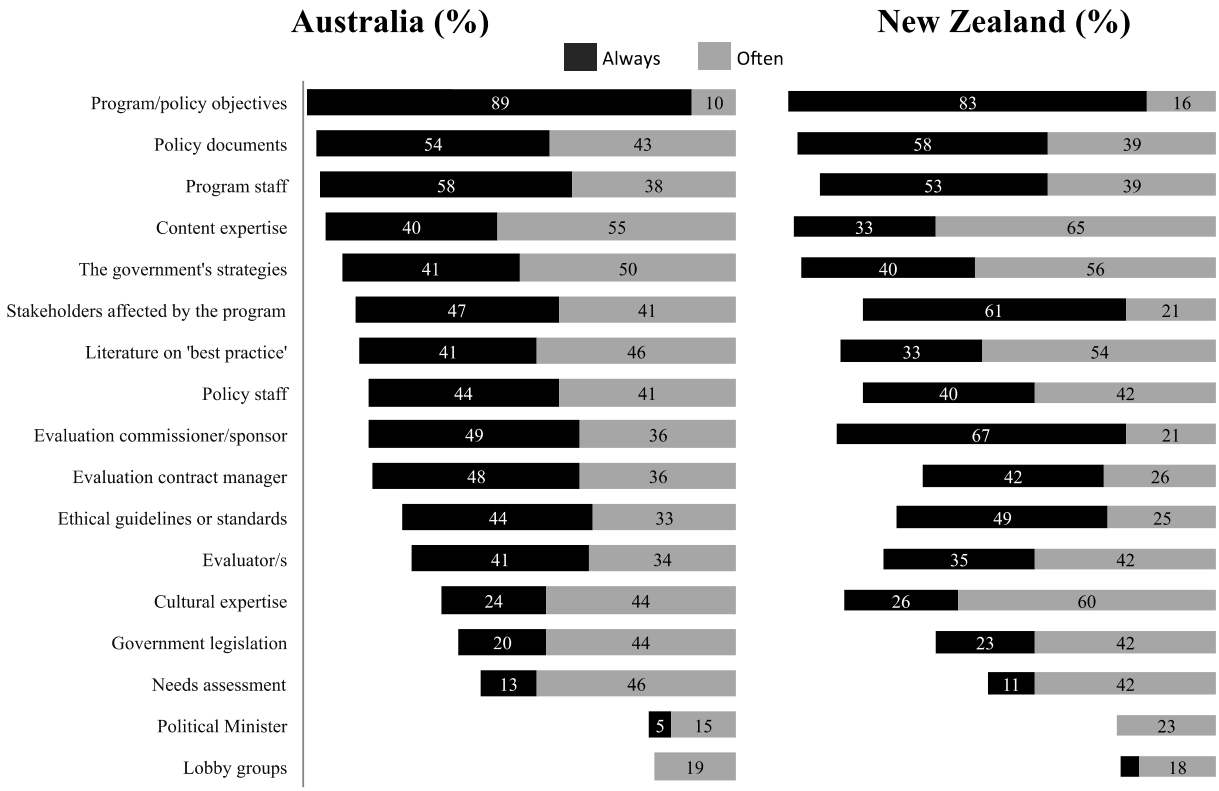


Figure 8: Extent to which respondents say they commonly use a source

An analysis of sources used by evaluators working in New Zealand (n=57) versus Australia (n=80) identified comparable findings for all but one source. New Zealand



respondents were 18 percent more likely to use cultural expertise (always or often) than their Australian counterparts (Figure 9). However, this difference was found not to be statistically significant.

Figure 9: Sources commonly used by respondents in two countries

The findings also showed that almost all respondents working in both government agencies and in private practice used program objectives as a key source for developing criteria. The only significant difference[33] between the two groups was that those working within a government agency were 31 percent more likely to use the political minister (always or often) than those in private practice.

Analysis of use by evaluators working in other sectors was difficult to interpret, as the numbers in each group were small. The findings indicated that program objectives were a commonly used source for respondents in all sectors. The data

[33] Significance confirmed by an independent T-test

indicated that those working in multi-agencies were more likely to use lobby groups as a source than respondents working in other sectors.

5.4.3 Additional sources identified by respondents

In addition to the items included in the survey, respondents were asked to identify other sources they had used to identify dimensions of quality for an evaluation. These sources are listed in Table 10 along with the number of respondents who identified the same source.

Table 10: Additional sources identified by respondents

Source type	No.
Previous evaluations	6
Similar programs in other jurisdictions	5
Baseline data	5
Grey literature, e.g. government websites, industry reports	2
Program theory	1
Program documents, including contracts	1
Values of provider organisation	1
Literature on ethics, social justice, human rights	1
Prescribed international criteria (e.g. DAC)	1
Strengths assessment	1

5.4.4 Factors that influence selection of sources

To help understand the findings on the use of different sources, respondents were asked to identify the factors that influence decisions about which sources are used when they are evaluating a public sector program (Appendix A, Question 9). The analysis is based on 128 respondent comments, as seven people did not respond to this question and two comments were incomprehensible to the researcher. Four key influences were identified that could contribute to decisions on which sources to use in an evaluation. They are (1) feasibility, including time and budget constraints; (2) political imperatives, including political or agency influence over the evaluation scope; (3) relevance to the evaluand context and (4) trustworthiness of the source. These are discussed in order of most to least mentioned by

respondents. It should be noted that many respondents identified more than one influence, thus the number of comments does not add to 128.

5.4.4.1 *Feasibility*

More than half (77) of respondents' comments referred to resource constraints as a key factor influencing decisions about sources used for developing criteria. Respondents said that stakeholders were sometimes not consulted if it involved travel to meet them or face-to-face interviews were required. Several respondents identified work-around strategies when it was not possible to include a source. One included the use of secondary sources, for example using existing documentation or literature to access stakeholder perspectives. A second approach was to prioritise sources, although the basis on which some sources were selected over others was not identified. There was a sense that evaluations had to 'fit the cloth', with one respondent commenting:

> (Public sector evaluation is) a diverse, challenging area in which you feel (often) that for each new evaluation you do your best, but due to time and resource constraints…you feel that the criteria may not be quite there, as in not quite complete/exact, but they will have to be good enough as you need to progress the evaluation.

5.4.4.2 *Political imperatives*

Decisions about what sources were used or not could be influenced by agency officials, for example the evaluation commissioner, or even the minister. Fifty-five respondents referenced this type of influence, referred to here as political imperatives. Some respondents viewed officials' expectations as a key determinant in deciding what was important in an evaluation, indicating that the client, or commissioning agency, had the final word on what sources were included or excluded. Examples of these types of comments included:

> Views of those commissioning the evaluation and the steering committee are very influential

> What the minister wants (which is usually relayed to me via senior managers).

Other respondents' comments indicate a tension between political imperatives and the evaluator's ability to access the most appropriate sources for scoping an evaluation. Respondents noted that agency control over an evaluation could occur particularly when the evaluand was of strategic importance or had a high profile. Where this was the case, evaluators might be required to access a wider range of sources or conversely be asked to restrict the type of sources used. As one respondent commented:

> If there is a contentious issue that the minister doesn't want externals to find out about, then the range of sources is limited accordingly.

Another aspect of agency control related to which sources were viewed as more credible by government stakeholders. For example, one respondent noted that decisions about the evaluation focus could be determined by:

> …pressures/anxiety from upper echelon stakeholders about the type of sources/methods deemed to be credible.

Similarly, another respondent commented on what they viewed as 'the politics of the day' having an influence on decisions about what was important in an evaluation:

> …the more generally pervasive influence of political conservatism, scientific positivism, and the ideology of new public sector management. These contribute to fetishizing facts, not looking for and/or misunderstanding causal mechanisms, over-emphasising efficiency, and often requiring a value for money analysis without adequate counterfactuals or consideration of externalities.

Respondents noted that the evaluation scope could be formally determined through a 'Request for Tender' document, leaving the evaluator with little capacity to engage in a discussion or negotiation about appropriate criteria. As one respondent said:

> I would love to have more capacity to tap the perceptions of the people affected by the policy/program and potentially affected by its evaluation, but this is rarely possible.

5.4.4.3 *Relevance*

Forty-two respondents identified the maturity of the program (whether it was still in development, fully implemented or completed), the subject area of an evaluand and evaluation approach as key influences on what sources were used in an evaluation. Respondents noted that sources required for a formative[34] evaluation might be very different from those required for an evaluation of a more mature initiative. For example, some sources could be considered less relevant in an initiative that had never been done before. As one respondent commented:

> You would not review the literature if the initiative was very new, with not much of an evidence base.

Another respondent commented that for summative[35] work they tended to focus on the 'appropriation level objectives' to determine the criteria for judging whether the intervention was effective in achieving its overarching objectives:

> From this we develop the intervention logic (here policy and programme staff are involved) to identify the intermediate outcomes to access.

The subject area and style of evaluation also determined the relevance of some sources over others. Participatory-type evaluations might give more emphasis to stakeholder input, and as one respondent commented, there might not be a relevant lobby group for the project being evaluated.

5.4.4.4 *Trustworthiness of the source*

A smaller number of respondents (16) identified the trustworthiness or quality of a source as influencing decisions about which sources were used. In their responses, some just noted 'quality' of the source as a consideration, while other respondents referred to terms such as 'integrity', 'credibility', 'authoritative', or 'reliability'.

5.4.4.5 *Evaluator's valuing stance*

The analysis of the qualitative data on influences identified that evaluators might also have a role in determining which sources were used. This finding emerged from a deeper analysis of the qualitative responses in which two broad views about

[34] 'Formative' refers to an evaluation conducted during the development or delivery of a program. It generally has an improvement focus (Mathison 2005)

[35] 'Summative' refers to evaluation that is done at the completion of a program (Mathison 2005).

valuing were identified. Both views regarded stakeholder engagement as important for understanding value; however, within the groups there were different emphases about whose values mattered most.

One cluster of comments indicated a government-oriented or client-focused approach that was focused primarily on serving the needs of evaluation commissioners or other agency stakeholders. Here the emphasis was on 'intended use by intended users' (Patton 2008). This perspective is demonstrated by the following comment:

> The end users (policy makers, operational staff, Ministers) are very important. Understanding what they want to do with the information shapes what we will then look at. We also consider what type of information they will see as credible.

A second cluster of comments indicated a strong justice or equity focus, suggesting the perspectives of minority groups ought to be prioritised when identifying value. This perspective is illustrated by the following comments:

> I always design an evaluation that ensures a focus on increasing equity and ensuring the methodology is inclusive of minority populations.

> Priority goes to those most affected by the evaluand.

There were also comments that indicated some respondents sought a more democratic approach to valuing. They articulated the importance of including the voices and perspectives of less powerful, often disenfranchised, stakeholders, while also trying to balance the interests of their more powerful, managerial stakeholders. The following quotes illustrate this balancing act:

> Ultimately, I work inside a management paradigm, but I aim to consider competing perspectives and question dominant views.

> I tend to think about what might happen at the end of the evaluation – who might be able to stand up and say 'Oi, that's not a reasonable way to make a claim about success or failure', where they might be coming from personally/politically…If the users of an evaluation feel like their frame of reference isn't reflected in the design, they may dismiss the whole thing.

What accounts for the difference in evaluators' stances? One possible explanation is that evaluators working within a government agency might be more invested in a utilisation focus that served the needs of government officials and would therefore give emphasis to their interests. To explore this potential explanation further, respondents' comments were coded into one of two categories, termed managerial versus democratic, based on their responses to Question 9, and then analysed by sector/workplace to explore any differences. No noticeable differences emerged; evaluators working in a public sector agency were just as likely to bring a democratic public-interest lens to their work as those working in private practice, universities or not-for-profit agencies.

To further explore the existence of different evaluation stances, the quantitative data on respondents' use of sources (Question 7a) was subjected to an exploratory factor analysis (EFA). This analysis clustered the sources into two factors as shown in Table 11[36].

Table 11: Sources prioritised by two factors

Source	Factor	
	1. Close connection to government	2. External to government
Evaluation commissioner	.579	.197
Policy staff responsible for developing the initiative	.552	.005
Policy documents	.644	-.069
The government's strategies	.591	.221
Government legislation, standards or regulations	.487	.281
Program or policy objectives	.510	.120
Evaluation contract manager	.615	.387
Lobby groups	.603	.307
Stakeholders affected by the policy or program	.100	.660
Ethical guidelines or standards	.136	.644
Cultural expertise	.153	.662
Program staff	.449	.518
Literature on 'best practice' for the intervention	.331	.475
Political minister	.291	.042

[36] After rotation, the factors explain 22 percent and 10 percent of the total variance respectively. Together the two factors explain 32 percent of the total variance. A third factor explained another 8 percent of the variance (based on three sources: the political minister, needs assessment and evaluator).

Needs assessment	.053	.329
Evaluator	.068	.080

Extraction method: Principal component analysis
Rotation method: Promax with Kaiser Normalization

Three sources (political minister, needs assessment and evaluator) aligned with a third factor. However, there did not appear to be any common features shared by these three sources. This factor was therefore set aside from the analysis. One other source, content expertise, did not align strongly with any one factor.

The two factors included in the analysis were each given a descriptor based on their main characteristics. For Factor 1, it was the connection to government that linked eight sources; hence the descriptor 'Close connection to government'. Factor 2 sources include the evaluation commissioner or contract manager, policy staff, policy documents, the government's strategies, legislation, standards, regulations, program or policy objectives, and lobby groups. At first glance, this final source, 'lobby groups', may seem an anomaly alongside the other sources included in Factor 1. However, lobbying is a form of advocacy with the intention of influencing legislation, regulation or other decisions made by government. Lobby groups may also be powerful stakeholders, with close alliances to political parties.

In contrast, the standout feature linking the sources in Factor 2 is their independence or distance from the machinery of government; hence the descriptor 'external to government'. Factor 2 sources include stakeholders most affected by a program, cultural experts, program staff, ethical guidelines or standards and literature on best practice.

There is an interesting alignment between Factors 1 and 2 and two of the evaluator valuing stances identified in the analysis of the qualitative responses. Specifically, Factor 1: Close connection to government would appear to align with a managerial or client-focused valuing stance, while the second factor (external government) appears to align more closely with a social justice or equity focus.

Further analysis was undertaken to explore whether any of the factors were associated with evaluators working in different geographic locations (New Zealand versus Australia) and/or sector (government, private practice, university, not-for-profit or multi-sector). This analysis involved first grouping sources into the two factors and then, for each respondent, calculating the total mean for each factor.

Respondents were then assigned to a factor according to whether their highest mean score aligned with Factor 1 or 2. Seventeen respondents had the same mean score on both factors and were thus set aside; the other 120 aligned with only one factor. This group was subjected to further analysis to identify if any demographic differences explained respondents' alignment to a specific factor. The findings indicated no differences based on where respondents lived. For instance, just over half the respondents in both Australia and New Zealand aligned with an 'external to government' stance.

Analysis by sector also showed little difference between those respondents working in government agencies and those in private practice. The analysis did, however, indicate that evaluators working in a university or not-for-profit organisation were more strongly aligned with Factor 2: External to government stance (Table 12).

Table 12: Alignment to a factor by sector/s where respondents work

	Factor	
	3. Close connection to government (%)	4. External to government (%)
Private n=45	29	56
Government n=49	29	56
University n=11	18	82
Not-for-profit n=4	0	100
Multi-agency n=11	33	40

Note: Percentages do not add to 100, as respondents who did not align with only one factor are not included in the analysis[37].

One possible explanation is that university-based evaluators have more resources to undertake literature reviews and/or they may be more sceptical about using government-based sources if they are not based on empirical evidence. Likewise, not-for-profit stakeholders may place more weight on the perspectives of those least enfranchised, that is, consider them more authoritative than government-related sources. Caution is required in interpreting this finding, as only a small number of respondents from these two sectors completed the survey.

[37] Respondents who loaded on more than one factor were working in the private sector (7), government sector (6) and multi-agency sectors (4).

5.4.5 Sources considered more, or less, important

This section examines respondents' views about the importance of different sources when developing criteria, and where applicable, contrasts these findings with respondents' ratings of actual use of sources. This is followed by an analysis of respondents' perspectives about what makes a source more, or less, important.

Figure 10 provides an overall comparison of respondents' ratings for importance against their use of sources. As this figure shows, a program's objectives were the source considered most important and were also most often used.

The analysis indicates a number of sources that were considered important or very important, but not always used in practice. These include needs assessment (20 percent of respondents were not using this source but considered it important or very important); the political minister (20 percent), government legislation, regulations (14 percent), cultural expertise (13 percent), ethical guidelines or standards (10 percent), stakeholders affected by the intervention (9 percent), lobby groups (8 percent) and literature on best practice (7 percent).

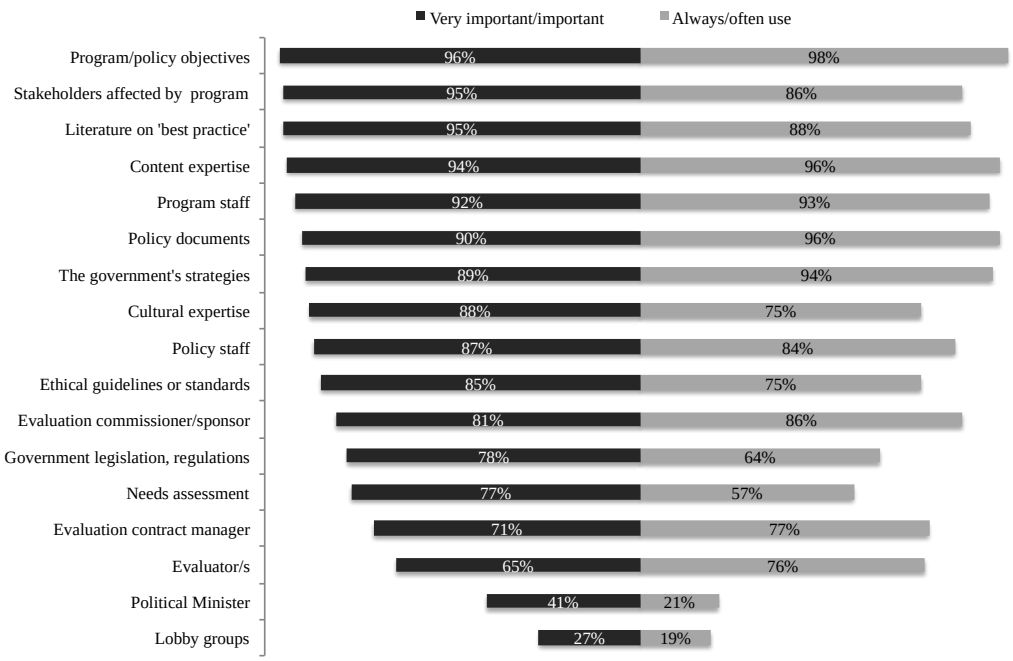


Figure 10: Importance versus use of sources

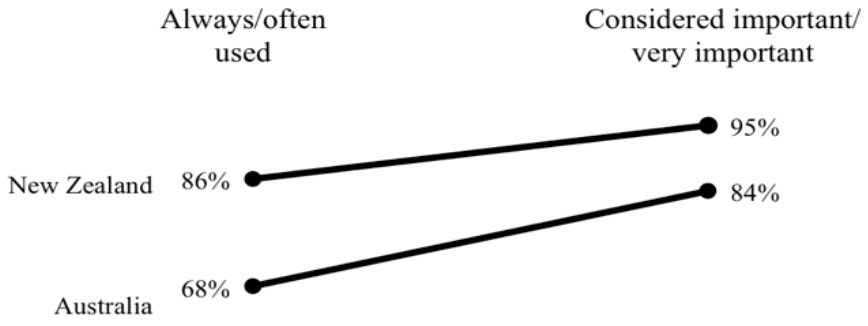
Two sources revealed a noticeable, though not statistically significant, difference between use and importance. The first was needs assessment, where 77 percent of respondents said they considered it an important/very important source, yet only

57 percent used it in practice. A potential explanation is that evaluators assume a needs assessment has been conducted as part of the policy development and thus incorporated into policy documentation. Alternatively, it may not be feasible to conduct a needs assessment within the available time and budget allocated for an evaluation. The commissioning agency may also have no interest, or see no benefit, in conducting a needs assessment if the program or policy has been instigated at a ministerial level.

The findings also indicated a difference between respondents' views about the importance of the political minister as a source (41 percent), and respondents' use of this source (21 percent). However, unlike needs assessment, there were contrasting views about the importance of the political minister as a source for developing criteria. While 41 percent of respondents rated the political minister as an important source, another 58 percent of respondents rated the political minister as unimportant/less important.

An analysis of sources considered important by evaluators working in New Zealand versus Australia identified comparable findings for all sources except one. Respondents based in Australia were 20 percent more likely than their New Zealand counterparts to consider policy staff as an important or very important source. The reasons for this are not known.

The findings also indicated cultural expertise was not accessed as often as respondents would like. As Figure 11 shows, this was the case for respondents in both Australia and New Zealand, although New Zealand respondents indicated they were more likely to use cultural expertise than their Australian counterparts, and also more likely to regard it as an important source for identifying criteria. The increasing recognition of the Treaty of Waitangi as New Zealand's founding



document, and the implications of this on public policy, is likely to account for at least some of this difference.

Figure 11. Comparison of use and importance of cultural expertise

5.4.6 Important factors to consider when identifying criteria

The online survey included an open-ended question (Question 10) asking respondents their perspective about what made a source more (or less) important when identifying dimensions of quality or value for program evaluations. The analysis identified five factors that might inform a respondent's decision about whether to use a source. To some extent these were the same factors that emerged from the analysis of Question 9, 'what factors *influence* decisions about which sources are used?' as well as the factor analysis, which focused on which sources were most commonly used by respondents. The convergence of themes is illustrated in Table 13.

Table 13: Convergence of themes identified in the data analysis

Themes/factors that might influence respondents' decisions about identifying/using a source to develop criteria		Factor analysis	Qualitative thematic analysis	
		Question 7a	Question 9	Question 10
Different valuing stances	Close connection to government/ managerial/client-based focus	x	x	x
	External to government/ democratic focus	x	x	x
	Justice, social equity focus		x	x
External factors	Feasibility		x	
	Political imperatives		x	
Context of evaluand	Relevance		x	x
Quality of source	Trustworthiness		x	x

The first two factors that emerge from the analysis of Question 10 focus on the trustworthiness and relevance of difference sources, as they relate to a specific evaluation. A feature of the last three factors is that they focus on the issue of whose values ought to be incorporated into criteria. Three distinct stances emerged from the analysis: a managerial focus, a justice and equity focus, and a democratic focus.

Table 14 shows the number of respondents who identified one or more factors as being important to the development of defensible criteria. Some respondents identified more than one factor in their response, thus the number of factors identified does not add to 129. The next part of this section describes each factor and gives examples from the data.

Table 14: Factors that inform which sources are used (n=129)

Factors	No.
Trustworthiness	56
Managerial/goal-based focus	48
Relevance	27
Justice and equity focus	14
Democratic focus	14

Trustworthiness: This factor captures ideas about the trustworthiness of a source and incorporates a range of terms used by respondents such as validity, authoritative, reliable, evidence-based and quality. Respondents' comments indicated different perspectives about what constituted a trustworthy source. Some respondents gave greater weight to stakeholders who had 'direct experience with the evaluand' or those who were 'well informed' or had a 'level of expertise in the field'. Others expressed caution about stakeholders' close proximity with an evaluand. For example: 'People involved in delivery have their own version that is not always correct.' In terms of written sources, peer-reviewed documents were more trustworthy than what respondents referred to as 'opinion pieces'. As one respondent noted:

> Evidence-based sources are most important as they enable comparisons against a benchmark or target.

Relevance: This factor refers to sources that are relevant to the context of the evaluand or type of evaluation. Some respondents referred in general ways to the

evaluand context, for example: 'Relevance (of a source) to the nature and context of the particular evaluand'. A specific example given by one respondent included an evaluation in the public defence agency.

> If (the evaluation is) for the delivery of criminal legal aid then obviously I need to refer to the legislation, whereas if it's adding a new service in the court for victims of sexual violence I might not need to.

The evaluation methodology may also influence types of sources, with some drawing value from the methodology rather than specific criteria of merit determined at the design phase of an evaluation. For example, one respondent noted:

> a cost benefit evaluation of a pharmaceutical intervention will be much less participatory than a programme evaluation. In this example the credibility of the pharmaceutical evidence is far more heavily weighted towards the outcomes from RCTs[38]. In programme evaluation I give less value to RCTs and more value to a wider range of evidence (including the different perspectives of stakeholders).

Managerial focus: This factor emphasises sources that satisfy the needs of the client or primary users of the evaluation findings. More than one third of respondents (56) included comments that indicated the respondent had a responsibility to include government-based sources in their conception of value. Many, but not all, respondents included policy makers, operational staff and ministers in their perception of primary users. As one respondent commented:

> Understanding what they want to do with the information shapes what we then look at. We also consider what type of information they will see as credible.

This factor also considers the goals of the evaluand to be important. Examples of respondents' comments include:

> Its centrality to the purposes, objectives or goals of the evaluand.

[38] Randomly controlled trials

Several respondents saw program or policy designers as a key source for identifying quality dimensions. For example:

> Programme designers are probably the most important –what they have designed and why is really the single thing that matters most, because that will tell you what they expected it to do.

Democratic focus: This factor was used to code a small number of comments (14) that indicated a focus on incorporating and balancing the values of different stakeholders. This was conveyed through comments such as:

> Balancing the needs of different stakeholders

> The need to ensure understanding quality from a range of differing perspectives.

Some respondents identified goals as important alongside other sources, thus indicating a link between a managerial and a democratic focus.

Justice and equity focus: A small number of respondents (14) specifically identified an equity focus as important. These respondents indicated that they aimed to prioritise the values of those most disadvantaged when identifying dimensions of quality. For example:

> I always design an evaluation that ensures a focus on increasing equity and ensuring use of a methodology that is inclusive of minority populations.

The three stances – managerial, democratic, and justice and equity focus – provide further evidence that respondents may approach valuing in different ways. The survey did not explore the reasons for these potential differences.

5.5 Theory-practice alignment

This section interprets the research findings on evaluation practitioners' perspectives of developing criteria in light of the existing literature. The discussion is presented in three sections. The first section discusses explicit versus implicit development of criteria. In section two the researcher discusses the types of sources participants say they most commonly use to develop criteria. The final section discusses the priorities given to different types of sources.

5.5.1 Explicit versus implicit development of criteria

Across Australian and New Zealand respondents, and of those responsible for managing or conducting evaluations in different workplaces, less than half the respondents said they evaluated against explicit criteria of merit. A more common approach was to use evaluation questions to guide an evaluative assessment. What might explain the difference between those who said they evaluated against criteria of merit and those who did not? While respondents were not asked to elaborate on their answers, two respondents did provide unprompted responses that offered a starting point for considering potential explanations.

The first respondent noted they did not consider judgement-making to be part of their role; instead they sought to answer questions as agreed with the evaluation commissioner. Such a view has been articulated by Alkin et al. (2011) as one where the evaluator provides the data but leaves the valuing to key stakeholders. The evaluator's role may be seen as that of providing scientifically based information. Underpinning this approach is a perspective that evaluation need not be about judgement. Such a view is evident in Chelimsky's (1987) argument that the evaluator's role is:

> a scientific one of bringing the best possible information to bear on a wide variety of policy questions (p. 26).

However, unless one is in the business of providing purely factual information based on descriptive research questions, it is not clear how such evaluators can avoid judgement based on criteria, even when they are implicitly framed. Evaluative questions, by definition, focus on values related to merit, worth or significance. As Shadish et al. (1991) note, nearly all theorists profiled in their text agreed that evaluation is about judgement, not just about describing programs.

A comment by a second respondent provided an alternative explanation for the absence of criteria in evaluation practice. The respondent agreed that evaluative judgements require criteria but noted that they lacked the confidence and skills to develop them. Such a view is consistent with a growing recognition that criteria are fundamental to evaluative reasoning (Nunns et al. 2015) and that evaluation requires evaluation-specific methods (e.g. Davidson 2005; King et al. 2013; Ozeki et al. 2019).

Mathea Roorda #579498

5.5.2 Types of sources used to develop criteria

A key focus for this component of the study was to understand more about the sources that informed criteria (see section 5.2). A list of potential sources was identified from the literature (Table 6). While it was expected that respondents would be familiar with many of these as sources for scoping an evaluation, it was anticipated that they might not necessarily associate them with the task of criteria development. Therefore, the preamble to the question about sources steered clear of the term criteria and instead asked evaluators about the potential sources they might use to identify the dimensions that distinguish (for example) a good or well-functioning program from one that is poor.

'Program objectives' was identified as the source most commonly used by respondents. This finding aligns with theorists' long-held assertion that evaluations are most commonly framed around a program's objectives or goals and the focus is on measuring attainment of these (Mark et al. 2000; Rossi et al. 2004). The findings from this study indicate that evaluators may draw on other sources to support or supplement criteria based on a program's objectives or goals. Respondents indicated they commonly drew on a range of sources, as appropriate to the context, including literature on best practice, content expertise, the professional views of program staff, policy documents and government strategies.

5.5.3 What sources are given priority?

One might expect that evaluators draw on sources relevant to the evaluand, the evaluation purpose and other contextual influences. This expectation was articulated clearly by respondents as an important characteristic of defensible criteria. Yet the findings indicate that the evaluator may unwittingly have a role in determining how criteria is conceptualised. First, the factor analysis identified two distinct stances in terms of respondents' prioritisation of sources. One stance would appear to prioritise government or client-based sources, in that most of the sources within this cluster are closely connected to the machinery of government. In effect, respondents who take a managerial approach commonly draw on a narrow range of sources from which to develop criteria. They look to policy documents, a government strategy and program objectives. Another stance emerging from the factor analysis prioritises a broader range of sources that

include the values of stakeholders affected by a program, cultural experts, ethics, program staff and literature on an intervention.

A second, thematic analysis of two questions, one about influences on decisions on which sources to use; the other about what makes a source more (or less) important, produced the same two stances that emerged from the factor analysis, reinforcing the finding that the evaluator might have some influence, either implicitly or explicitly, over which sources were prioritised. In addition, a third stance emerged in which some respondents aimed to incorporate a social justice or equity focus.

Similar stances are identified by Schwandt (2018) in his article on the normative characteristics of ethics in evaluation professionalism. He identifies three perspectives or ways that evaluators approach evaluation that then inform how they manage valuing. The first is the mainstream social trustee model, where the role of the evaluator is as a technical expert who works closely with the client to help them achieve a program's stated objectives. The second is an advocacy model, where the evaluator is closely aligned to those most affected by a program, and the focus is on engaging less powerful stakeholders in setting the evaluation agenda. A third approach, and the one clearly preferred by Schwandt, is a democratic professionalism model. Here the evaluator's role is that of co-designing, collaborating and co-creating knowledge with a broad range of stakeholders. A democratic professionalism approach to valuing is also consistent with guidance provided by the Program Evaluation Standards (Yarbrough, Shulha, Hopson and Caruthers 2011). Standard U2 refers to an expectation that the evaluator should attend to all stakeholders with a legitimate interest in a program.

The Program Evaluation Standards also include an expectation that evaluators develop self-knowledge of their own values and an awareness of others' values. As noted under *U4: Explicit Values*:

> Evaluations cannot afford to be seen as merely reflecting the vested interests of those who already have significant influence in the program and its evaluation (Yarbrough et al. 2011, p.38).

The use (or non-use) of cultural expertise provides a case in point with regard to how evaluators' values, as well as the broader cultural and political context, may influence selection of criteria. The researcher anticipated a significant difference

between New Zealand, versus Australian, evaluators' use of cultural experts as a key source for developing criteria. This difference was anticipated not because New Zealand evaluators share an inherently different set of values to their Australian counterparts, but because, as noted in section 3.3, the Treaty of Waitangi principles underpin many government policies and programs and one might therefore expect a greater emphasis on cultural input into criteria development in New Zealand evaluations. While evaluators in both countries have been challenged with developing culturally relevant and responsive approaches to evaluations, it is posited that increased recognition of the Treaty of Waitangi, and an increase in kaupapa Māori evaluation approaches, have served to normalise indigenous concepts, practices and protocols in New Zealand in a way that has yet to occur in mainstream Australian evaluation practice. It is therefore possible that the majority of evaluators in both countries draw on similar types of sources, and the modest difference indicated by the findings may be attributed to those respondents whose work is deeply immersed in, or informed by, kaupapa Māori approaches. However, the survey did not include a question about respondents' ethnicity, or the types of evaluation approaches being used, so this theory is untested.

5.6 Chapter summary

In Chapter 5 the researcher has provided the first of two perspectives about management of criteria in Australian and New Zealand program evaluation practice. As such, the development of evaluation criteria was explored from the perspective of 137 evaluators who manage or conduct program evaluations in Australia or New Zealand. Less than half the respondents said they evaluate against criteria of merit. A more common approach is to use evaluation questions to guide an evaluative assessment. Almost all respondents consider a program's objectives to be a key source for developing criteria. The findings indicate different evaluator stances in terms of prioritisation of sources. One stance prioritises managerial or client-based sources, while the other endeavours to incorporate a range of perspectives in the way criteria are conceptualised.

6 Criteria in evaluation reports

6.1 Purpose of this chapter

A second perspective on the management of criteria in Australian and New Zealand program evaluations is based on a review of 47 program evaluation reports in the public domain. All the reports were conducted or commissioned by Australian and New Zealand agencies responsible for social services, environment or agriculture and primary industries. To the researcher's knowledge, this is the first such study to include an in-depth examination of how criteria are managed in program evaluation reports.

The chapter is structured in four sections. In the first, the rationale for the document review is outlined, followed by a discussion of the research design. The findings from the review of evaluation reports are then presented, followed by an interpretation of the findings.

6.2 Research objective

The research objective for the document review was to describe how criteria are managed in the context of different types of evaluands and evaluation settings, including the types of sources used to define and justify criteria. Along with the findings from the online survey, this component of the study was intended to inform the research question: 'How are criteria managed in Australian and New Zealand program evaluations?' Both the survey and the document review provide different but complementary perspectives on the research topic. The findings from the survey gave insight into how evaluation practitioners approached the task of identifying and selecting criteria. However, a limitation of the survey approach was that criteria could not be examined in the context of specific evaluands. Thus, the research strategy for the next component of the study was to examine how criteria were managed in the context of completed program evaluation reports.

6.2.1 Sampling strategy

The study scope called for an approach that would allow for comparability of evaluation reports across two countries as well as different sectors and evaluand topic areas. This presented a challenge given the size of the Australian and New Zealand public sectors. At the Australian federal level alone (as at 2017), there

were 18 departments, while New Zealand's central government included 31 departments. Each agency also included multiple portfolios or topic areas. To allow for a broad focus, the researcher adopted a maximum variation sampling approach. This approach purposefully sampled for difference, a useful strategy when aiming for as much heterogeneity in the sample as possible (Patton 2002). The aim was to find comparable Australian and New Zealand public sector agencies that had diverse portfolios, thus allowing for maximum difference in evaluand topic areas.

To achieve maximum variation, while also maintaining a level of comparability across the two countries, agencies were identified at the highest level of government in Australia (federal agencies) and New Zealand (central government agencies). Agencies from both countries were then matched based on portfolios in common. A decision was made to exclude international development evaluations sponsored by the Australian and New Zealand aid agencies, as this would introduce an additional layer of context to the study such as the cultural contexts of recipient countries. Hence, all program evaluations in scope had an in-country (Australia or New Zealand) focus. The researcher examined websites of each of the remaining agencies and counted the number of portfolios common to each pair. Three pairings stood out as having the highest number of portfolios in common. These were the agencies responsible for social services, environment and agriculture and primary industries. As illustrated in Table 15 each pairing of agencies had six portfolios in common, and thus all were included in the sample.

Table 15: Final sample frame for reports

New Zealand	Australia	Sector	Portfolios in common
Ministry of Social Development	Department of Social Services	Social services	Children & families Disability Income support Parenting Women's safety Youth
Ministry for the Environment	Department of the Environment and Energy	Environment	Biodiversity Climate change Energy Land Marine Water

New Zealand	Australia	Sector	Portfolios in common
Ministry for Primary Industries	Department of Agriculture and Water Resources	Agriculture and primary industries	Biosecurity Fisheries Food safety Forestry Importing/ exporting Sustainable economic development Trade

Ideally the sample would have included all evaluation reports commissioned or undertaken by these agencies over a set period of time, and not just those publicly available. The risk, as noted by Bamberger et al. (2006, p. 127) is that published reports may comprise 'abbreviated and softened summaries' of evaluations. However, access to original evaluation reports, or those that have not been signed off for public release can be difficult. For example, Nunns (2017) sought to access evaluation reports directly from New Zealand evaluation units for a meta evaluation study but found that political scrutiny of public sector agencies at the time did not support accessing unpublished technical reports. The researcher accepted the limitation of public availability in the sample for this study.

To ensure currency, only publicly available reports dated 2010 to 2017 were included. A search was conducted on the website of each agency, as well as three research and evaluation clearinghouse websites: the New Zealand Family Violence Clearinghouse, Superu's 'The Hub', and the Australian Government's Closing the Gap Clearinghouse. A general web search was also conducted using the Google search engine. Search terms used included the name of the agency, the term 'evaluation' and the year (2010 through to 2017). Summary and draft reports were excluded from the sample. Where evaluation frameworks or plans documenting the methodology of a published evaluation were available, the researcher included them as an additional source of data.

Two new agencies were formed during the timeframe selected for the review. The Australian Department of Social Services (DSS) was established in 2013 and the New Zealand Ministry of Primary Industries (MPI) in 2012. The search for reports included the agencies that existed prior to the new organisations being established, and whose portfolios were incorporated into the new entities. No publicly available

reports were found for DSS's predecessor, the Family, Housing, Community Services and Indigenous Affairs (FaHCSIA), for the period 2010 – 2011. It is possible a decision was made not to upload older FaHCSIA reports to the new agency's website. Likewise, a search on MPI's website and on Google yielded no evaluation reports between 2010 and 2013 for evaluations conducted by the agencies it replaced, that is, the Ministry of Agriculture and Forestry, the Ministry of Fisheries and the New Zealand Food Safety Authority.

Full technical reports identified in the search were downloaded from the respective agency or clearinghouse websites if they included the term 'evaluation' in their title or there was clear evidence inside the report that the inquiry constituted a program evaluation[39]. The search yielded 47 full (i.e. not summary) reports of which 27 were published by Australian agencies and 20 by New Zealand agencies. All reports were included in the document review.

Table 16: Sector of reports (n)

Sector	Australia	New Zealand
Social	19	11
Environmental	4	4
Primary industries	4	5
TOTAL	27	20

More than half were evaluations of social sector programs (Table 16). The reason for the lower number of published reports from the environmental and primary industry sectors is not known. One possible explanation is that evaluative activities are not evenly distributed across the public sector. This appears to be the case in the US, where Peck (2018) identified the big three investors in evaluation as the US Department of Health and Human Services, the US Department of Education, and the US Agency for International Development. However, as noted in Section 3.2, Australia has legislation that requires all government agencies to undertake evaluations and reviews of its programs.

[39] A search on one New Zealand agency website yielded only two reports. The researcher contacted the evaluation manager who suggested using the term 'review' in the agency's web search function. This identified a further three reports; all included reference to 'evaluation' in the body of the report. For example, NZ72 notes: '(the team) *evaluated* the management and government of the program to assess whether accountabilities and processes are clear and adhered too (sic)' (p. 22).

The following three tables present additional characteristics of the reports included in the document review. The first, Table 17, includes an overview of the portfolio areas of reports included in the sample. There was a broad alignment of reports to the portfolio areas identified on the websites of agencies. However, no evaluation reports were found in the following areas: biodiversity and marine (environment sector) and food safety and fisheries (agriculture and primary industries). Further, the sample included reports in several portfolio areas that were not mentioned on the websites. These were financial services, migrants, refugees (social sector) and environmental research (environment). Across Australia and New Zealand, there were differences in the attention given to specific portfolio areas, which may reflect different government priorities over the timeframe they were commissioned.

Table 17: Portfolio areas (n)

Evaluand portfolio areas		Australia	New Zealand
Social sector (30)	Children & families	2	5
	Disability	5	0
	Financial services	1	0
	Income support	4	0
	Migrants	1	0
	Parenting	1	0
	Refugees	1	0
	Women's safety	0	3
	Youth	4	3
Environment (8)	Climate change	0	1
	Energy	1	2
	Environmental research	2	0
	Land	0	1
	Water	1	0
Primary industries (9)	Biosecurity	1	0
	Forestry	1	0
	Importing/exporting	0	1
	Sustainable economic development	1	3
	Trade	1	1

Although not explicitly seeking reports with different purposes, the sample does in fact include a spectrum of formative and summative reports (Table 18). Almost half the reports included a focus on both formative and summative, with the rest focused on either formative or summative.

Table 18: Evaluation purpose

Evaluation purpose	Australia	New Zealand
Formative only	6	6
Formative and summative	15	8
Summative only	6	6
TOTAL	27	20

The sample also included a range of author types. Half (24) the evaluation reports were authored by consultants, while the remainder included a mix of public sector agency, university or quasi-government authors as shown in Table 19.

Table 19: Evaluator's workplace

Evaluator workplace	Australia	New Zealand
Consultancy	16	8
Public sector agency	3	7
Public sector agency and university collaboration	0	1
University only	6	4
Quasi-government agency	2	0
TOTAL	27	20

6.2.2 Analysis strategy

The researcher's analysis strategy began with a broad focus, describing criteria as evident in the reports. Once located, the reports were re-read to locate any evidence of sources used to inform how criteria were defined. An inductive approach was employed comprising both content and thematic analysis, with the intention of interpreting and describing data to generate conceptual understanding of the research topic. Such an approach is appropriate and useful when little is known about a topic (Creswell and Plano Clark 2007).

Although this was not a grounded theory study, some components of this approach informed the analysis. For instance, an open-coding process, typical of grounded

theory research, was employed (Corbin and Strauss 2008; Creswell 1994; Creswell and Plano Clark 2007). Data were analysed by reading each report and sections of text were given a code. Subsequent text, if similar, was coded to the same code or if different, to a new code. Codes changed over the course of the analysis as deeper understanding of key concepts evolved.

6.2.2.1 *Content analysis*

Content analysis involves establishing a set of categories and coding instances of text into these categories. The result is a quantitative count of instances. One advantage of the method is that it is replicable, and its validity is easily established (Silverman 2006). In the reports, different value terms, for example, 'effectiveness' and 'impact', signalled the existence of criteria, even when not explicitly identified as such. The analysis initially focused on identifying value terms[40], and then returning to them to understand how, or if, they were defined and justified in specific evaluations. Value terms of interest were those that defined, or signalled, an aspect of quality or performance against which a judgement was made later in the report. In addition to examining them in context, the aim was to explore what types of criteria were most, and least prevalent across the sample of reports.

6.2.2.2 *Thematic analysis*

Thematic analysis focuses on identifying and recording patterns, or themes, within data. The thematic analysis of explicit and implicit criteria serves several purposes and includes both deductive and inductive approaches. An example of both approaches follows. An initial deductive analysis focused on whether criteria were explicit or not. A definition of explicit criteria was developed to guide the analysis. Criteria were deemed to be explicit if reports included the term 'criteria' to describe dimensions against which the program was being assessed. In some reports the term 'indicator' was used to refer to the same concept as a criterion, that is, a descriptor of merit. For the purposes of this study, such indicators were redefined as criteria.

Criteria were deemed to be implicit if evaluative judgements were made against value terms that were signalled in evaluation questions, objectives or elsewhere, such as the terms of reference or the evaluation methodology. To facilitate

[40] If a report included several statements using the same value word or term it was counted only once.

examination of implicit criteria, the researcher constructed definitions based on (a) the value term found in the report and (b) information found elsewhere in the report that facilitated understanding of the value term. In this study, the term 'value construct' is used to define these researcher-constructed criteria.

In this phase of the analysis, each report was coded as including either 'explicit' or 'implicit' criteria, and the location of criteria was noted. Very few of the reports included explicit criteria (refer Findings, Section 6.3). Rather than dismiss the remaining reports as having unjustified criteria, the researcher sought to locate the meaning of key value terms in reports and to identify sources that might have informed the basis against which judgements were being made, and how (implicit) criteria were being understood. In this respect, the focus was on coding the 'visible and surface content meaning' of terms, referred to as 'manifest' content (Young Cho and Lee 2014, p. 4). Identifying the manifest content required reading, and sometimes re-reading each report in its entirety, as definitions or explanations were rarely included in the same section as the value term. By example, an evaluation objective in one report stated:

> Compare [the program] against other like programs and evaluate the extent to which the [program] design remains *fit for purpose* or needs to adapt.

The value term 'fit for purpose' was coded as an implicit criterion because there was evidence, later in the report, of an evaluative judgement about this aspect of the program. Further information about how the term was being defined was found in a heading in the findings section of the report: 'The unique value proposition of the [program]'. The findings noted the program:

> …remains fit for purpose. The following … illustrates the unique position of the [program] compared to other …programs with similar aims (NZ70, p. 15).

In this case the value construct, that is, the value term and its definition, was defined and coded by the researcher as follows:

> Fit for purpose: Defined as the program being unique compared to other (similar) programs (NZ53)

Because the definition was reconstructed by the evaluator, based on information found in the report, the unit of text (that is, the value term and its definition) was referred to as a value construct.

Once the value constructs had been identified, the researcher's intention was to examine the types of values inherent in the constructs. An initial attempt was to undertake this component of the analysis using a deductive approach, informed by the literature on normative ethics. Specifically, the aim was to code criteria to one of five normative ethical perspectives, with the intention of exploring which perspectives, if any, were more prevalent than others. A pilot analysis, and feedback from an evaluation and philosophy scholar, suggested a deductive approach was not appropriate as it would be difficult to claim that a particular normative ethical perspective lay behind the choice of a particular criterion, especially when that inference was based on written reports without the ability to member check with the evaluation authors.

Consequently, the deductive approach to analysing value constructs was abandoned and a second, inductive analysis was undertaken informed by constant comparative analysis (Glaser and Strauss 1967). This approach typically has four steps: (1) comparing incidents applicable to each category; (2) integrating categories and their properties; (3) delimiting the theory and (4) writing the theory (Glaser 1967, p. 439). A form of constant comparative analysis, comprising the first two steps, was used to guide the inductive analysis of the criteria and value constructs.

The first step involved comparing each criterion or value term to assess whether it was similar or different from other value constructs across reports. Coding began with the value terms that indicated a criterion. By way of example, four reports included the term 'alignment'. A reading of each report identified similar ways in which this term was being defined, and hence were all assigned to a code labelled 'alignment with design'. Two other reports referred to alignment with government strategy and were coded as 'alignment with government strategy'. In this respect, the labelling of codes remained close to the text found in the reports to avoid definitions becoming too distanced from their original meaning.

A second step of analysis involved comparing and verifying codes. Criteria and value constructs were re-classified and broader themes established. This resulted

in some value terms being coded into separate categories. For example, some 'effectiveness' value terms were coded into a broader category called 'quality of implementation' while others became part of 'administrative performance'. In other cases, different value terms such as outcomes, impact and performance were coded into the same category as the way they were defined in reports suggested they were synonyms.

6.2.2.3 *Sources used to justify criteria and value constructs*

A final analysis focused on how criteria and value constructs were justified in the reports. Each report was read to identify sources that might have informed how they were being defined. By example, one report included explicit criteria in the form of an evaluative rubric. The criteria were displayed in the rubric at two levels. At the first level, criteria comprised abstract value terms such as 'impact'. At the second level, each criterion was defined in sub-dimensions that comprised short statements. The methodology section of the report identified that the rubric was developed in conjunction with ministry officials. Careful reading of the report identified reference to government legislation, which provided a rationale for the way 'impact' had been defined.

Criteria and or value constructs were coded as nominally justified if the report included a rationale for their inclusion. Where criteria were coded as justified, further analysis was undertaken to explore factors that supported this justification. No assessment is made about the validity of the justification. In the example provided above, the criterion of 'impact' was deemed justified because the report included a direct link to a source that had informed its definition (ministry officials) and, in turn, this definition was informed by government legislation.

6.2.2.4 *Code checking*

Miles and Huberman (1994) stress the importance of code checking, not only to aid definitional clarity but also to improve the reliability of coding. Code checking was particularly important for the analysis of the document review as criteria expressed in the reports were predominantly implicit. This approach is illustrated in Table 20.

First, a value term indicating a criterion had to be identified. This was recorded in column A, along with its location in the report. In order to understand how criteria

were justified, the researcher extracted the value term (usually a high-level abstract word as per column B). Next, the researcher searched for clues as to how this term was being defined or understood. In the case of the example here, the meaning of the term 'effectiveness' became apparent in the sub-evaluation questions (column C).

Table 20: Example of coding approach for the document review

A	B	C
Statement indicating a criterion, and its location in the report	**Value term**	**Definition and its location in the report**
		Value construct
'How effectively has [the program] been implemented?' Location: Evaluation question	Effectiveness	Encompasses: • Resource implications • Targeting suitable individuals and groups • Ability to transfer into and out of the program appropriately • Effect of the program on service providers. *Location:* These phrases appear in the sub-evaluation questions under the evaluation question.

An experienced public-sector evaluator and recent PhD evaluation scholar was recruited to code check a set of data. A comparison of codes produced by the code checker and the researcher identified an issue related to definitional clarity. The code checker rejected a number of statements included in the initial analysis (that is, column A, Table 20) on the basis that they considered the statement to be a descriptive research question rather than indicating a criterion. For example, one report included the question:

> Which new ways of working for [the agency] have worked well and not well, and why (considering different contexts, constraints and resources)?

While initially this looked like a descriptive research question, there was evidence of an implicit criterion in that the report later included an evaluative rating of 'very good' for this question. The findings section of the report described ways in which agency staff had worked with Māori landowners in six projects and included an evaluative claim that the different approaches worked well within their contexts.

Thus, the descriptive question, 'ways of working with Māori landowners' was used evaluatively, based on evidence in the report that this was one of the dimensions against which an evaluative conclusion was being made about the program's worth and value. The researcher discussed the reasons for including a value construct based on this statement, arguing that the report included an evaluative judgement about the program's ability to work well with these landowners. A second round of code checking based on this discussion resulted in agreement about all but two of the 166 statements (Appendix C, Statements 51 and 53).

6.3 Findings

6.3.1 Introduction

This section presents the findings from the review of published Australian and New Zealand program evaluations. The review contributes a second perspective on the management of criteria in Australian and New Zealand program evaluation practice. The findings also inform an understanding of characteristics of defensible criteria for program evaluations.

The research findings are structured in four sections. The first examines evidence of explicit criteria in evaluation reports. The second describes how explicit and implicit criteria are organised and presented in evaluation reports. The third examines the types and combinations of sources used to define and justify criteria. The fourth studies the dimensions of performance that evaluations focus on, with a view to exploring similarities and differences in the way abstract value terms are defined or understood, and to explore how, or if, evaluations provide assurance that programs are judged against all relevant criteria.

6.3.2 Evidence of explicit criteria

In the context of public sector programs, the evaluator's role is unequivocal. As Schwandt (1997) notes, evaluators have a different role to researchers. They are:

> obliged by the very nature of their work to make claims about the value of some practice, program, policy, project, or technology (p. 9).

These values are most clearly articulated as explicit descriptions of quality or performance, referred to as criteria.

In the 47 reports reviewed, six reports included explicit reference to criteria; that is, they used the term 'criteria' to describe the dimensions on which the program has been assessed. Such criteria were evident in both Australian (4) and New Zealand (2) reports, and across the three sectors. In five of the reports criteria appeared as textual definitions, either in the body of the report (3) or in a rubric (2). In one report criteria appeared as bullet points in a high-level concept map, with no further descriptors.

In the remaining reports, criteria were implicit either in indicators (4), evaluation questions (21), or where no evaluation questions existed, they were evident in the evaluation objectives, purpose or terms of reference (16).

Table 21: Types of criteria evident in reports

		Australia	New Zealand	Total
Explicit criteria	Explicit reference to 'criteria'	4	2	6
Implicit criteria	Include reference to 'indicators'	3	1	41
	Criteria implicit in evaluation questions	12	9	
	Criteria implicit in purpose, objectives, terms of reference or methodology	8	8	

6.3.3 Management of criteria in evaluation reports

The reports were further examined to understand how criteria were organised and presented. Here, the intention was to move beyond the use or non-use of the term 'criteria', and to examine in more detail how the meaning of criteria was conveyed. Existing literature on organisation and presentation of evaluation criteria was used to guide the analysis, with a view to comparing theorists' conceptions of criteria with what was evident in the sample of reports. Three strategies to managing criteria were identified: (i) use of hierarchal approaches to move from abstract criteria to specific levels of detail, (ii) use of report headings to explicate meaning, and (iii) use of comparators.

6.3.3.1 *Strategy one: Use of hierarchal approaches*

As noted in Chapter 2 (Section 2.4.4), both Sadler (1985) and Scriven (1990) refer to the use of hierarchical approaches to move from the abstract to more specific levels of detail in terms of defining criteria. This strategy for organising criteria was found in all six of the reports with explicit criteria and many of the reports where criteria were implicit in evaluation questions. Two illustrations of this strategic approach to organising criteria are provided. The first is a report with explicit criteria (NZ70), in which criteria are structured at three levels; the most abstract being 'value for money'. This term is defined as incorporating four 'performance dimensions': environmental, economic, cultural and social values. Each dimension is then further defined as illustrated in Figure 12.

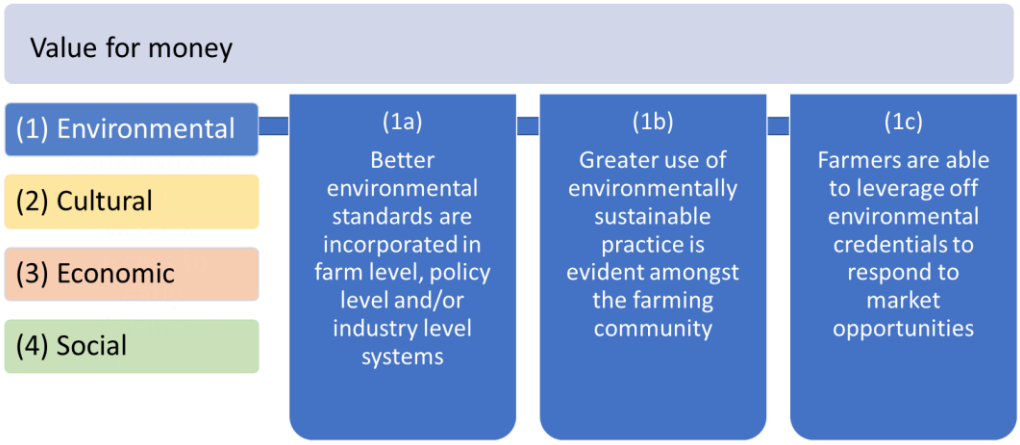


Figure 12. Hierarchical approach in an explicit criterion (NZ70)

The second example (Figure 13) is an implicit criterion that is evident in an evaluation question: 'How and to what extent has each of the pilots influenced the local and wider […] sectors?' This signals an expectation, that is, an implicit criterion, that the program will result in some kind of influence over the sectors in which it is engaged. Four bullet points that follow the question signal the range of influence the program is expected to achieve. In this respect, there is little difference between the organisation of criteria in a report where value terms are explicitly identified as criteria, and those where dimensions of merit are evident but left implicit. Both are managed in a hierarchal way; although the explicit criterion 'value for money' includes an extra level of detail that is not evident in the implicit criterion of 'influence'.

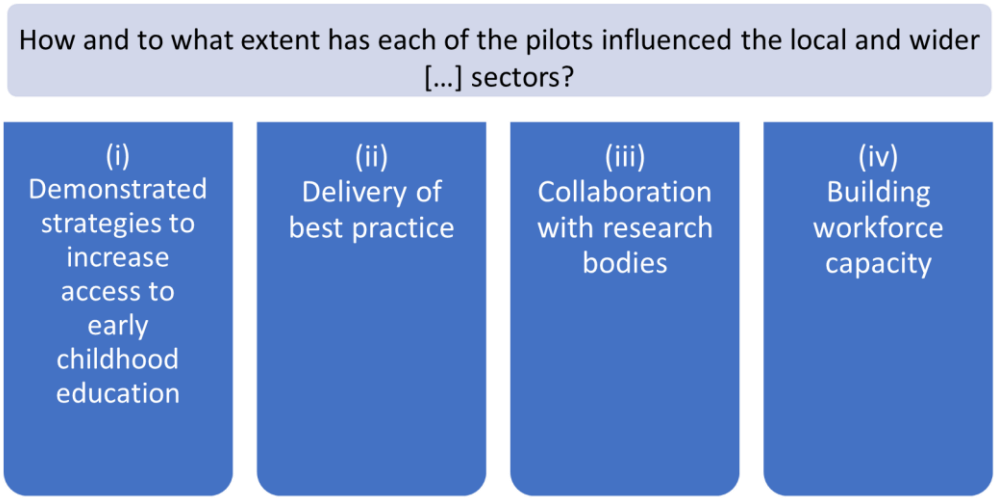


Figure 13. Hierarchical approach in an implicit criterion (AU20)

Analysis of reports where criteria are implicit in evaluation questions indicated that the hierarchical approach to defining implicit criteria was ad hoc; in 10 reports, some value terms were defined in more detail, for example as sub-questions or in bullet points following a question, as illustrated in Figure 13 and some value terms remain ambiguous. In such cases, the meaning of abstract value terms generally became clearer in the findings section of the report. In two reports, a criterion was indicated, but there were no findings related to the criterion, and no assessment was subsequently made. For example, one of the reports (NZ05) included a question: 'Is the program being delivered consistently?' The report was read several times and still it was not clear to the researcher how 'consistent' was being defined. The researcher's written reflection upon reading NZ05 was as follows:

> There are no findings about consistency of delivery in the report. Perhaps the question relates to the following? 'Investment into each young person has to be consistent across the group, and that dealing with a young person who is not coping with the [X[41]] program diverts that energy away from the others' (NZ05, p. 29). In any case, there is no assessment about whether this occurred.

[41] Names of programs have been removed to afford the evaluation authors some level of confidentiality.

6.3.3.2 *Strategy two: Use of report headings*

A second strategy employed by report authors was to use headings in the findings section of reports to implicitly frame criteria. This would appear to be a similar approach to that noted by Stake et al. (1997):

> Seldom are the criteria seen as direct criteria of merit but rather information categories from which interpretations are made (p. 92).

Such an approach was most noticeable in 18 reports where there were no evaluation questions. For example, AU14 included an evaluation objective: 'Assess the impact of the Package on access to services'. The way in which 'impact' and 'access to services' was understood became clear only when reading the evaluation findings, where section headings were structured in four parts: Part A: Outcomes for children and families; Part B: Awareness and registration; Part C: Early intervention (access, equity, affordability, quality and best practice); Part D: Package as a service model. The information included in each part implicitly conveyed how 'impact' and 'access to services' had been defined in this context.

6.3.3.3 *Strategy three: Use of quantitative indicators*

The third approach identified in the reports was the use of quantitative indicators that served as proxy criteria. In 12 reports, criteria were implicit in an evaluation question or the evaluation aim, and then further defined as quantitative indicators[42]. In Report AU03, for example, 'justice outcomes' were defined as 'measurable reduction in the rates of offending or recidivism by Indigenous Australians'; 'health outcomes' as changes in levels of marijuana, tobacco and alcohol used, physical, mental and sexual health' (p. 96). Another report (NZ42) described the evaluand and then introduced the methods with the statement: 'This study uses rich linked administrative data… and undertakes an impact evaluation using two separate quasi-experimental methods…' (p. 6) Here, the term 'impact' implicitly signalled the criterion of interest. Clusters of indicators were aligned with variables. For example, an outcome of 'early childhood education and health service engagement' had seven quantitative indicators of which the first included

[42] While quantitative indicators were evident in 12 reports, only four referred to the term 'indicators' (as noted in Table 21).

'maternal use of publicly funded addiction services in the first and second year following the birth of the children' (p. 19).

The challenges associated with using indicators as a reliable measure of success is illustrated with an evaluation of a boot camp program for youth offenders (NZ05). The program aim was to have youth successfully transitioned back into the community. The evaluation included an assessment of the extent to which this was occurring, defining 'successful transition' as the extent to which youth do not reoffend over a 12-month period, or commit only a single offence of a minor nature. This indicator of success implicitly ignored other potential dimensions that might be critical to successful transition, for example, engagement with family, employment or education. None of these other potential dimensions were discussed in the report. In addition, the indicator would seem to be an unreliable measure of success. For example, a decrease in youth arrests might indicate a different type of successful transition to that envisaged by the program planners, whereby youth learnt new strategies (maybe through engaging with other young people in the program), in how to avoid arrest.

In eight reports, criteria were established by the evaluation approach, for example experimental, quasi experimental or use of pre-post[43] measures. For example, Report AU27 included an evaluation objective: 'Assess the extent to which trial participants have and exercise greater control and/or choice'. The term 'control and choice' was defined in terms of differences in control and/or choice experienced by program participants in comparison with those in a control group.

In another report, NZ06, the educational outcomes of young parents who complete their secondary schooling at a teen parent unit were examined. 'Outcomes' were defined in terms of government-held administrative data. The availability of data therefore determined how 'outcomes' was defined. In this case, enrolment rates and educational qualifications of teen mothers in close proximity to a teen parent unit were compared with teen mothers in locations where there was no teen parent unit available.

[43] Examining whether participants in an intervention improve or regress during the course of the intervention

Mathea Roorda #579498

6.3.4 Sources used to define criteria

As noted earlier, the way in which abstract terms such as 'performance' or 'effectiveness' are understood in one program will be different from another. Criteria derive their meaning from one or more sources, as relevant to the context of the evaluand. Hence, the reports were examined to identify the types, and combinations, of sources used to select and define values. This section begins with a broad overview of the number, and type, of sources identified in reports that inform criteria selection.

6.3.4.1 *Overview of sources used to inform criteria selection*

All the evaluation reports in the sample included reference to at least one source. In seven reports only one source was evident; another 22 reports included just two sources.

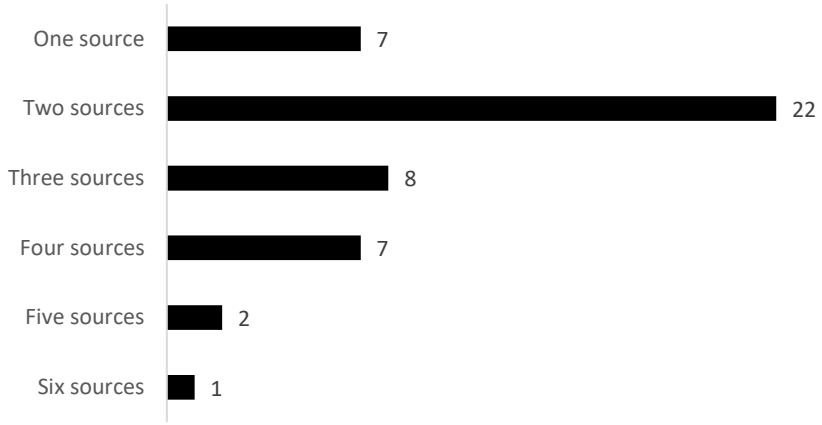


Figure 14: Sources used to inform criteria

The reports were examined to identify any differences between those that included explicit criteria and those where criteria were implicit. The six reports with explicit criteria included between one and three sources. Reports with implicit criteria referenced between one and six sources. There were no noticeable differences between the number of sources used in Australian versus New Zealand reports.

6.3.4.2 *Location of sources in reports*

Rarely was the link between sources and criteria clearly acknowledged. This was the case for reports where criteria were explicit as well as those where criteria were implicit. More commonly, a report included an introduction or background section that referenced sources such as literature on similar type programs, the

Mathea Roorda #579498

results of previous evaluations or a government paper that had informed the development of the program.

By way of example, the six reports with explicit criteria indicated a range of sources had informed the development of criteria (Table 22). Reference to sources was noted in the introduction, methodology and appendix sections of reports. Criteria were defined in methodology, findings or appendix sections. Only one report (NZ70) included information about both the criteria and sources in the same section of the report.

Mathea Roorda #579498

Table 22: Sources that inform selection of criteria

Criterion	Criterion expressed as:	Where in the report is the criterion defined?	Where in the report are sources located?	Sources that inform criteria selection
Impact (AU02)	Six bullet points describing 'desired results', expressed as descriptive statements	Methodology (program logic)	Introduction/background Appendix	Minister's speech Program objectives Relevant literature
Impact (NZ51)	Five dimensions in a rubric, expressed as one-word terms; then additional bullet points with indicators (i.e. evidence being sought)	Methodology	Introduction/ background	Legislation Public sector agency stakeholders
Performance (AU83)	Findings section starts with a definition of each dimension, expressed as sub-dimensions and a description of each sub-dimension.	Findings	Methodology	Funding agreement
Performance (AU13)	Descriptive statements of performance, with one or more quantitative indicators linked to each statement	Findings	Methodology	Program objectives Funding agreement Government report ('White Paper')
Alignment with standards (AU18)	Descriptive standards	Appendix	Introduction	National standards Government strategy Primary research
Value for money (NZ70)	Descriptive statements organised in a rubric	Appendix (methodology)	Appendix (methodology)	Government strategy Public sector agency stakeholders

Mathea Roorda #579498

6.3.4.3 *Types of sources identified in the reports*

Figure 15 includes a summary of the types of sources identified in the reports that either link to, or support criteria. There is some overlap between sources evident in the 47 reports reviewed, and those identified by the survey respondents as most commonly used. For example, analysis of both the survey and document review data indicate that program objectives were the most common source used to inform criteria. Five of the eight most commonly referenced sources were government based. In one third (16) of the reports, government-based documents were the only sources mentioned. They included one or more of the following: a minister's speech, program objectives, government strategy, a funding agreement or public sector agency stakeholders. In contrast there was less evidence of non-government sources being used to inform selection of criteria. The most prominent non-government source was relevant literature, evident in 18 reports.

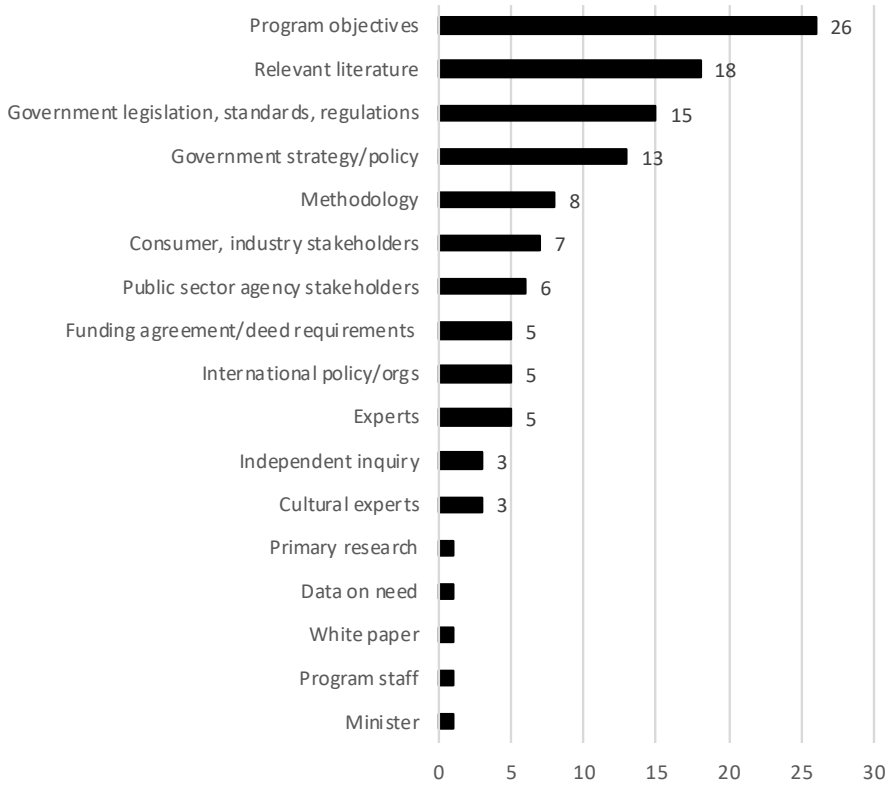


Figure 15: Sources that link to, or support criteria selection (n)

6.3.4.4 *Examination of reports with one or two sources*

This section examines the reports that included one or two sources; these comprised more than half the reports in the sample. In the reports that referenced

only one source, 'program objectives' was the only source (five reports); for the others it was a funding agreement for providers (one report) or relevant literature (one report).

Of the reports that referenced two sources, 10 included the program objectives plus another source as shown in Table 23.

Table 23: Criteria defined by two sources

Criteria defined by	Source providing support/justification	No.
Program objectives plus:	Relevant literature	3
	Methodology	2
	Standards	1
	Government strategy	1
	Public sector agency stakeholders	1
	Legislation	1
	Funding agreement	1

The inclusion of a second source provided context and, in some cases, support and implicit justification for criteria that were primarily based on program objectives. For example, Report AU03 included an evaluation question: 'Is the [program] meeting its aims and objectives?' The implicit criterion (meeting aims and objectives) was defined in the program objectives, a list of five bullet points related to expanding the diversity and availability of youth services and activities for youth. The second source was a coronial inquiry into the deaths by suicide/volatile substance of five young indigenous men. The inquiry concluded with a set of recommendations, including the need to implement culturally appropriate solutions to support indigenous youth. Accordingly, this second source provided a justification for a program that addressed the needs of indigenous youth, which in turn helped to justify the program aim. By comparison, where the program objective was the only source, the researcher was left wondering about its origin, and specifically, a question about whose interests had informed the program objective.

6.3.4.5 *Use of literature and stakeholders to justify criteria*

More in-depth analysis focused on two types of sources identified in the reports, relevant literature (18 reports) and stakeholders. The aim was to understand how,

or whether, these two predominantly non-government sources served to justify criteria in particular contexts. The findings of this analysis are as follows.

First, of the reports that included literature to inform explicitly or implicitly the development of criteria, six (AU02, AU15, AU20, AU21, AU24 and NZ05) used this source to review best practice approaches or outcomes achieved by similar programs operating in other jurisdictions or countries. In two of these reports (AU02 and NZ05), literature provided a critical perspective on program success that served to temper the expectations articulated in the official aims of each program. In the first (AU02) the evaluators stated that 'in order to evaluate [the program] it is important to clearly articulate the objectives of the policy and the criteria by which the success or failure of the policy in meeting the objectives set for it are to be evaluated' (p. 11). They noted that the objectives for the program had broadly been taken from a minister's speech and a government policy statement. The desired results, as articulated in these official sources, were listed in a high-level concept map. The evaluators presented a literature review of evaluations of similar type programs implemented in other countries. This review included an examination of the philosophical basis and justification for the type of program being implemented and outcomes that had been achieved and not achieved in other similar programs.

In the second report (NZ05), the evaluators of a military-style program for youth offenders noted it was 'widely accepted in the literature that traditional 'boot camps' do not work and in some instances can exacerbate offending behaviour' (p 11). They referred to aspects of the evaluand that had been modified to the traditional boot camp approach that potentially set this program apart from programs that had failed elsewhere. The literature implicitly served to justify a criterion that was focused on assessing whether these additional components 'worked well'.

Scriven (1986) has argued that criteria are best informed by needs of the affected population. Five reports (AU25, NZ06, NZ39, NZ42, NZ72) used literature to support the need for a program. For example, AU25 was an evaluation of a program providing financial services to low-income families. The report referred to literature that showed people on low incomes and those with poor credit history had limited options if they required additional funds.

The second in-depth analysis of sources focused on stakeholder input into the development of criteria. Numerous scholars note the importance of developing criteria with input from multiple stakeholders. Incorporating diverse perspectives serves to broaden rather than narrow how quality is understood (Greene 2005; 2011) and incorporate the perspectives of those who have historically been marginalised. Approaches that incorporate diverse value perspectives are evident in House and Howe's (2000) deliberative democratic approach and in the work of Hall et al. (2012). Stakeholder input into criteria selection also underpins a descriptive approach to valuing, where the aim is to include the values and interests of all relevant stakeholders. Such an approach is evident in Mark et al. (2000) notion of values inquiry and (Stake 1975) responsive evaluation approach.

Evidence of stakeholder input in the development of criteria was found in less than a third (17) of the reports. Stakeholders consulted in the design of an evaluation included consumer or industry spokespeople (seven reports); public sector agency staff (six reports), subject experts[44] (five reports), cultural experts (three) and program staff (one report). Some reports included references to more than one type of stakeholder.

A sub-set of the analysis of stakeholders focused on cultural input as a source for informing quality or success of a program. More than half (27) the reports in the sample included a focus on indigenous peoples, of which all but two were social sector evaluations. This focus was evident in different ways. In some reports, indigenous participants were identified as a significant group in the program characteristics, as for example: NZ05: 'To date, 60 per cent of [program] participants have been Māori,' (p. 18), and AU02: 'Overall, 91 per cent of those subject to [the program] are Indigenous' (p. xvii). In other reports, indigenous participants were a small percentage of participants, but identified as a priority group, as in Report AU09: 'This study focused on mothers in groups of special policy interest: single mothers, Aboriginal or Torres Strait Islander mothers, mothers from culturally and linguistically diverse backgrounds, and mothers employed on casual contracts or who were self-employed before the birth' (p. 4). As noted, three reports referred to the use of cultural experts. Two experts were members of the evaluation advisory group (NZ38 and NZ39) and one was a

[44] These included academics with expertise on gender-related issues (AU09, NZ39), disability (AU24), refugee settlement (AU28) and energy efficiency (NZ50).

member of a forum that helped develop national standards on which the evaluation criteria were based (AU22). In both the New Zealand reports, the experts were identified as indigenous. The Australian report referred to a migrant expert. None of the Australian reports with a focus on Aboriginal or Torres Strait Islander people made reference to cultural experts providing guidance to the evaluation. Evidence of cultural stakeholder input into the evaluation design, including criteria selection, was not evident in the remaining 24 reports.

6.3.5 What dimensions of performance do evaluations focus on?

Deciding what criteria to include in an evaluation is crucial. As Scriven (1990) notes: '…everything else hinges on it' (p. 17). Criteria help to define the focus of the inquiry, thus ensuring that data is collected on important dimensions of performance. Accordingly, the analysis of reports included a description of the dimensions of performance as presented in the evaluation reports. The analysis began by identifying all the value terms evident in reports, of which the most common are shown in Figure 16. The three most common terms, effectiveness, outcomes and impact, were found in similar numbers in both Australian and New Zealand reports. The other three terms, appropriateness, efficiency and alignment were more dominant in Australian reports. Of the New Zealand reports, 'appropriateness' appeared in one; 'efficiency' in two, and 'alignment' did not feature at all. There were no noticeable differences in terms used across sectors.

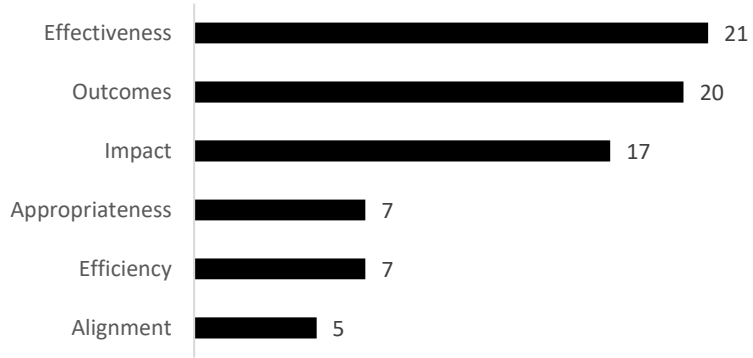


Figure 16: Most common criteria identified in the reports

As noted earlier, these terms were evident as explicit criteria or implicit in evaluation questions, the evaluation purpose, terms of reference or in the methodological approach. Many other value terms were also identified in small numbers, for example: non-discriminatory, equitable access, gender equity, choice

and control, work life balance and sustainability. Examples of values related to economic values included cost, cost effectiveness and economic feasibility.

6.3.5.1 *Identifying the meaning of abstract terms*

The criteria were coded, along with definitions evident in the reports and analysed to explore if terms had shared or different meanings. The value terms broadly clustered into one of three evaluation foci: (1) formative (2) summative and (3) future. A key finding was that abstract value terms were not self-evident in their meaning. Unless they were clearly defined in the report, they remained ambiguous to the researcher. Where defined, it was apparent that abstract terms, for example 'effectiveness', could have different meanings across reports. Terms were rarely defined when first introduced. Rather, a close reading of the report was required to identify how terms were being defined for a particular context. For example, a term's meaning might be assumed from the types of sub evaluation questions being asked, or from the headings in the findings section of a report.

An analysis of terms that clustered into one of the three evaluation foci follows. Each section begins with an example of the terms used in reports that aligned with these foci.

6.3.5.2 *Formative focus*

> Abstract terms with a focus on the quality of a program's design included: need, appropriateness and fit for purpose, reach, alignment, meeting aims and objectives, implementation as expected, non-discriminatory implementation, effectiveness

A sub-set of criteria focused on administrative performance, where terms included: performance, compliance, effectiveness of administration, appropriateness of governance structure, efficiency, cost, cost effectiveness.

Criteria that clustered under the formative category included a focus on the underlying rationale for a program (using the term 'need' or 'appropriateness') or on adaptions to program design (using the term 'appropriateness' or 'fit for purpose'). For example, AU28 used the term 'appropriateness' to ask whether the program was still valid and aligned with an identified need and with broader government policy. Two other evaluations (AU13, AU15) used the term 'appropriateness' to evaluate whether the design needed to be adapted, for

example, to take into account new or different geographic locations, age range of participants or types of providers.

Some criteria included a focus on reaching or targeting the right participants, for example:

> AU15: *Reach* (defined as reaching the eligible cohort of young carers)
>
> AU21: *Alignment* (defined as participants aligned with the eligibility criteria for the program).

In other reports, the term 'alignment' referred to alignment with existing standards or alignment with other parts of a system. For example, AU12 was a program aimed at teaching children personal safety skills. One of the value statements focused on how well the program integrated with the school curriculum and the 'broader education environment of the region' (AU12, p. 10).

A range of terms focused on process-related criteria, some of which focused on program quality and others on administration. They included whether program support was effective, appropriate, timely, of the right intensity or dosage, and whether staff had the capability to work effectively with participants. Such criteria also included a focus on whether individual program components were working well or whether components of a program were working well together. Administrative criteria focused on both financial and non-financial aspects of the program's operation. Value terms coded in this category indicated a focus on accountability. Two evaluations, for example, included a focus on performance and compliance against grant deeds.

Seven evaluations included the value term 'efficiency'. In some reports the term was not defined; in other reports the term was more tightly described. For example, AU28 described efficiency as having three dimensions: delivery of services, management of providers and the administration of the program (p. 85). It then further defined each dimension, describing delivery of services as program processes to manage expenditure on services, the relationship between program inputs and outputs [and] whether there were alternative models that would allow government to deliver the program more efficiently. Only one of the 47 reports (AU15) included the value term 'cost effectiveness', which was defined as 'the administrative costs per participant'.

Mathea Roorda #579498

6.3.5.3 *Summative focus*

> Abstract value terms in this category included: effect, impact, change, influence, performance, outcomes, benefits, maintenance of postvention gains, equity

Nearly half the reports (22) included a focus on outcomes or impact. Criteria in this category used a range of terms to describe the expected outcomes of a program on participants, the environment, industry or on other parts of a related system, for example 'the wider sector'. In contrast to value terms in the formative category, where meaning (if provided) varied according to the context of the evaluand, terms in this category could have the same meaning. For example, three reports (AU20, AU19, AU21) described changes anticipated on the wider sector variously using the terms 'influence', 'impact' and 'effect'. While different abstract terms were used in each report, all referred to change of the same type.

6.3.5.4 *Future focus*

> Abstract value terms in this category included: replicability, wider applicability and sustainability

Two evaluations use the terms 'replicability' and 'wider applicability' to refer to expanding their respective programs from a small pilot phase to a more permanent transition that incorporated a broader geographic focus. A third evaluation also had an expansion focus but was concerned with the program remaining financially sustainable if it continued beyond a pilot phase, and thus used the value term 'sustainability'. None of the environmental reports reviewed included a focus on sustainability.

6.3.5.5 *Criteria with a focus on culture*

As noted in section 6.3.4.5 more than half (27) the reports in the sample include a focus on indigenous peoples. The reports were relatively evenly split between Australia (14) and New Zealand (13). An analysis of the criteria in these reports was undertaken, with a view to understanding if, and how indigenous values were incorporated into understanding program quality. Guiding the analysis was an assertion by Kawakami et al. (2008) that evaluations involving indigenous peoples and communities:

> must promote and practice an indigenous worldview, including, but not limited to, consideration of indigenous identity, epistemology, values and spirituality (p. 230).

In the evaluations where indigenous peoples were a primary focus, there was little evidence of cultural input included in the evaluation design. Two New Zealand reports made reference to engaging cultural expertise, though no detail was provided about what this had entailed. One of the reports (NZ36) included a definition of 'effectiveness' that was defined as 'whānau satisfaction with the process and positive feedback about immediate outcomes in terms of safety, health, behaviour and education'. A second criterion of cultural appropriateness was defined implicitly in the evaluation findings as including use of Māori tikanga in the program, cultural support and supervision for staff and supporting participants to conduct whakapapa searches to engage with whānau and hapu.

In addition, one Australian report (AU22) included reference to engaging migrant expertise in the evaluation design. The evaluation included a criterion of effectiveness, defined in the findings section of the report as 'what is working well from both client and service provider perspectives, e.g. personalised service, sense of community generated by/within services, non-affiliation of service providers with past 'care' providers'. A second criterion focused on alignment of the program with a set of design principles, which had been agreed to by a forum that included cultural expertise.

Information included in several reports set up an expectation that the evaluation would include a focus on cultural effectiveness or appropriateness that was then absent in the report. Two illustrations are provided here, to show how these expectations were conveyed in reports. The first concerns an evaluation of a youth program (AU03). The need for the youth program is documented in the background section of the report, where there is reference to a coronial inquiry into a number of Aboriginal youth suicides. The inquiry noted the need for a 'culturally appropriate' solution to a problem. This sets up an assumption that the evaluation would include an explicit assessment of the cultural appropriateness of the program for this group of participants. However, there is no reference in the report to Aboriginal youth being a target group, and no reference to cultural appropriateness being a key requirement for the program, apart from a brief reference to the term 'diversity' which is not defined. Rather, the evaluation is

structured around three evaluation questions that are further defined in sub questions as identified in Table 24. These questions effectively serve to limit the evaluation scope to a review of governance, attainment or non-attainment of stated objectives and completion of written deliverable products.

Table 24: Criteria implicit in an evaluation of a youth program

Appropriateness further defined in the findings section of the report (as sub-headings): • Governance • Program infrastructure • Service implementation • Performance monitoring and reporting
Achievement of aims and objectives: The aims are to: • Expand the diversity and availability of existing youth services • Identify service delivery gaps • Implement new youth activities • Facilitate community capacity building to support ongoing activities • Engage and support young people to transition from childhood into productive adulthood in East Kimberley communities by significantly increasing access to and engagement in meaningful structured activities.
Impact refers to: • Preventing and reducing volatile and other substance use • Achieving better justice outcomes (reducing boredom, anti-social behaviour, contact with justice system, i.e. youth offending, recidivism and incarceration) • Attaining better health and life outcomes: including the resilience and wellbeing of young people and supporting improved education training and employment outcomes • Youth service delivery and infrastructure in the region, including the coordination of youth programs and services in and across the nine communities serviced by the [program] • Community safety.

A second illustration concerns the introduction of a mainstream primary school protective behaviour program in communities where the target population is predominantly Aboriginal (AU12). The report notes that the program involves nearly 2000 children in communities that include a high proportion of indigenous children. In one school, Aboriginal children comprise 90 percent of the roll. As with the previous example (AU03), this information sets up an assumption that the assessment of the program will include a focus on the cultural appropriateness of the program for Aboriginal children. The researcher examined the report closely, with the intention of identifying and describing any cultural values against which this program had been assessed. Implicit criteria were evident in two evaluation

questions. These focused on the effectiveness of the program in teaching protective behaviours to children, and the extent to which the program supports the aims of a broader program. Finding no mention of cultural appropriateness or relevance in the criteria, the researcher turned to the evaluative claims made in the report. Here, a judgement was made that:

> the program was inclusive of children with cultural or special needs (AU12, p. 5).

This evaluative claim further established an expectation that part of the assessment of the program included a focus on cultural appropriateness (as well as appropriateness for children with special needs). Evidence to support the evaluative judgement was found in the following statement:

> a number of principals and teachers noted that the program was appropriate for a diversity of children, including those for whom particular cultural sensitivities may be required as well as children with developmental delays or learning disabilities (p. 19).

A quote from a respondent clarified how appropriateness for children with cultural needs was being defined in this context:

> the animals [used in the program] are quite significant totems in the different [Aboriginal] nations for our kids, particularly here locally for our Kamilaroi kids. (Principal) (p. 20).

However, the report included no information that justified the principal or staff as cultural experts and did not acknowledge or identify other aspects of cultural appropriateness or relevance, such as language, that might be relevant dimensions of merit for an educational program of this type.

6.4 Theory-practice alignment

This section interprets the findings from the review of published evaluation reports in light of the existing literature. The findings are considered along relevant literature and structured in three sections. The first discusses how explicit and implicit criteria are organised and described in reports. The second reflects on strategies used in the reports to justify criteria, and the third examines evidence of all relevant perspectives being incorporated into a conception of value.

6.4.1 Organising and describing criteria

6.4.1.1 *Organising criteria*

Whether expressed as explicit criteria or implicit values in evaluation questions, the evaluation aim or in some other form, a key finding from the review was that all the reports included values against which the program was (or could be) assessed. In all reports, the values were introduced as terms that were highly abstract, that could have different meanings depending on the context. In those that included explicit criteria, these abstract terms were specified in more detail using a hierarchical approach. Thus, abstract terms might be defined as comprising several dimensions of merit, and each dimension was further explicated, somewhat similar to what occurs in dictionary definitions. This hierarchical approach to clarifying meaning was also evident in reports where values were implicit in evaluation questions or elsewhere. However, in many reports, the practice of specifying abstract value terms was not routine. Even within a report, some abstract value terms could be clarified while others were not.

Given the fundamental role of these values in providing the basis against which judgements are being made, it would seem important to give due attention to how abstract value terms are being defined. Providing clear and full descriptions of how abstract terms are being defined serves an important function of clarifying and communicating meaning. As noted by Sadler (1985):

> (b)y proposing an interpretation for a criterion at the next highest level, they enable others to assess the adequacy of a decomposition (p. 288).

Where abstract terms are spelt out in more detail, it might seem irrelevant whether they are explicitly signalled in reports as criteria or left implicit but well defined. The added benefit gained from explicitness is that of supporting the reader to understand the evaluator's reasoning process. The reader knows exactly how the evaluator got to their assessment, because the values against which a judgement is being made are identified as having a role in the reasoning process. Being explicit about how value terms are defined enables everyone to be on the 'same page' and provides a foundation for informed debate about the judgement. As such, being explicit would seem to be a necessary first step toward defensibility of criteria. That is, the reader has information about what values have been included, and against which a judgement or set of judgements has been made.

Mathea Roorda #579498

The researcher observed that reports included evaluative judgements even where value terms remained ambiguous. As Nunns et al. (2015) argue, the crux of evaluation quality is sound evaluative reasoning. When a key element of evaluative reasoning is missing, such as clearly articulated criteria (whether explicit or implicitly stated), the case to support an evaluative claim is weakened.

6.4.1.2 *Defining criteria*

Abstract terms such as appropriateness or effectiveness could have different meanings depending on the program. They did not necessarily follow dictionary definitions or those used in evaluation texts. Schwandt (2015) for example, suggests the term effectiveness refers to the degree to which a program is successful in achieving intended results. However, that is not the only way in which the term is being used in practice. Sometimes the term 'effectiveness' is related to program processes, for example effectiveness of support. Likewise, the terms 'outcomes' and 'impact' could refer to the same concept as well as different concepts, despite evaluation scholars' attempts to distinguish between the two (e.g. Bamberger et al. 2006; Markiewicz and Patrick 2016). The range of value terms used to define, or signpost dimensions of value, serves to reinforce the contextual nature of evaluation in the public sector, and also the importance of clarifying meaning for the report reader.

Beyond formative and summative purposes, criteria covered a range of dimensions as appropriate to the context of specific evaluands. Without content knowledge it was not possible for the researcher to assess the completeness of criteria (and nor was this the intended focus of the document review). However, as noted earlier, there were clear indications that criteria were narrowly defined or incomplete. For the researcher, this was most noticeable in relation to evaluations where indigenous people were a primary target group.

6.4.2 Justifying criteria

Armed with information about what criteria have been included in an evaluation, the next step was to consider the relevance of these to the task at hand. Scholars, including Nunns (2017) and Hurteau et al. (2009), suggest that the relevance or validity of criteria, and therefore evaluative reasoning, is strengthened if the criteria are justified, for example by reference to a source.

In the review of evaluation reports, the task of linking sources to the development of criteria was challenging. This was especially so when criteria were not explicitly stated or clearly defined. In many cases the researcher had to first identify abstract value terms, and then read the reports for clues as to their meaning. These clues could be located in different parts of the report, for instance in the introduction section where background information might refer to need or expected performance, or in the findings section, either in the way headings were structured, or in the description of program performance. Once a definition for an implicit criterion was located, the search then began for sources that had informed how criteria were being defined. The task of linking sources to criteria was akin to doing a jigsaw puzzle. Rarely were they located in the same section of a report and explicitly linked. In summary, it was difficult to determine if criteria were justified because this information was not easily locatable in reports.

In one third of the reports, criteria were based on government sources alone. This finding bore similarities to those found in the survey data, whereby some participants' responses indicated a tendency to rely primarily on government sources to inform criteria development. This finding indicates evidence of a managerial approach to evaluation in which quality or success is determined primarily through an understanding of the program goals or objectives, or the decision-maker's perspective (House 1980a).

6.4.3 Toward comprehensiveness: identifying all relevant values

As noted, the document review identified little evidence of broader input into the evaluation design, or criteria development. Few reports provided evidence of engagement with all legitimate interest groups. One example of this was the 27 reports that included a specific focus on indigenous peoples. In only two reports (both conducted for New Zealand-based agencies) was there evidence of cultural expertise in defining values. In all other evaluations where indigenous peoples were a focus, the primary lens through which success or non-success was framed was that of government-based sources. More generally, stakeholder input, beyond government officials, was not evident in the majority of reports.

The concern here is that when quality or success is defined through primarily government sources, a program's quality or worth is determined by a narrow view of success rather than on a comprehensive set of criteria that incorporate all

relevant values. House refers to this as a 'managerial' approach to evaluation (House 1980a). One indication of a managerial approach in the document review was the focus on meeting obligations, noted in 27 reports, or on program outcomes (32 reports). In contrast, 13 reports included reference to fairness or equity-related values such as 'undertaking inclusive, culturally sensitive practice' (AU18). Only five reports included reference to ethic of care-related values, for example 'mental wellbeing' (AU09) and 'personalised service, sense of community' (AU22). It is possible that broader values perspectives have been incorporated into earlier work on program development, and thus inform program goals and objectives. However, without explicit reference to this information, the way in which success or good is conceptualised is not transparent, and the evaluation may not stand up to broader scrutiny.

6.5 Chapter summary

This component of the study has described how criteria are managed in Australian and New Zealand program evaluation reports. The document review includes 47 reports that were conducted or commissioned by government agencies with a focus on social services, environment or agriculture and primary industries. The research makes a significant contribution to the field of evaluation as the first study to include an in-depth examination of how criteria are managed in published program evaluation reports.

Although evaluative judgements were evident in all the reports, the way in which quality or success was defined was not always clear. Further, evaluative judgements in some reports were based on what appeared to be an incomplete set of criteria. This was especially noted in reports where indigenous peoples were a key target group for the evaluand. Criteria could also be narrowly defined through primarily government-based sources. This was the case for one third of the reports in the review. There was little evidence of criteria being explicitly justified in reports. In some cases, information from relevant sources indicated a criterion should be included but it was absent from the evaluation.

7 Synthesis of findings from phase one

7.1 Purpose of this chapter

In this chapter the findings from the online survey and the review of published evaluation reports are considered together, alongside the literature, to provide an evidence-based perspective on how criteria are currently managed in Australian and New Zealand program evaluation practice. The findings also inform a theoretical framework of key characteristics of defensible criteria for program evaluations.

7.2 Approach to synthesis

This section provides an overview of the process used to synthesise the findings from the online survey and document review. The synthesis process drew on Miles and Huberman (1994) who described a variety of techniques for recognising similarities and differences in data. The researcher adopted a three-step analysis process that involved (i) data reduction, (ii) data display and (iii) conclusion drawing and theory development. Each step is further described as follows.

First, data reduction began by re-reading the findings and interpretation sections from the two components of the study undertaken in phase one. A summary of key findings, in the form of short blocks of text, was then created. Second, a data display matrix (Table 25) was created. The first two columns of the matrix comprise an initial 'top down' frame with three categories: (i) organising and defining criteria; (ii) justifying criteria and (iii) including all relevant criteria (comprehensiveness). The blocks of text were assigned to one of these categories. The researcher then adopted an approach similar to the constant comparative method, whereby text from the survey data findings was compared with text from the document review findings. If text related to a similar context, it was assigned to a row in the matrix where the data could be considered alongside each other. Findings from the survey are included in the second column and the document review in the third column. Key ideas from the literature were included in the fourth column.

The third step, once the matrix had been populated, involved standing back and 'eyeballing' the data, with a view to identifying patterns and contrasting perspectives. While the data display matrix was developed within a couple of hours, the process of standing back and considering the data took many weeks. The

value of the display was that the data could easily be eyeballed; the text in each cell provided a short-hand way to return to the data with fresh eyes.

A problem that researchers can face when synthesising findings from multiple studies is that of determining their methodological comparability. On the one hand, both components of phase one were conducted by the same researcher and focused on the management of criteria in Australian and New Zealand program evaluations. On the other hand, key design features mean that the findings cannot be directly compared. For example, it is not possible to know if the authors of the 47 published evaluation reports have characteristics that are similar to the respondents who completed the online survey. Authors of the evaluation reports may or may not be affiliated to a professional evaluation association; and they may not have the same level of experience as those who responded to the survey. These sampling differences may account for some of the differences noticed in the two components of the study.

Table 25: Summary of key findings from phase one

Theme	Online survey (n=137)	Review of reports (n=47)	Excerpts from literature
Organising and describing criteria	45% of respondents say they evaluate against explicit criteria.	Six reports (4 Australian, 2 New Zealand) include explicit criteria.	Explicit criteria support evaluative reasoning / quality evaluations (Nunns et al. 2015). Stake and Schwandt (2006, p.407-408) refer to this as a 'quality-as-measured' approach.
	Almost all respondents (96%) say they use key evaluation questions (KEQs) to guide their evaluations.	The most prevalent approach to managing criteria is through KEQs, where criteria are implicit in abstract value terms. This was the strategy used in 21 reports. In another 16 reports, criteria were implicit in the evaluation purpose, objectives, terms of reference or methodology. Use of indicators was evident in four reports.	Stake and Schwandt (2006) and Bamberger et al. (2006) argue criteria can be too narrowly defined when explicit. They have observed evaluators focusing on easy-to-articulate indicators that do not fully represent quality. They advocate for an emergent and implicit quality-as-experienced approach to establishing criteria (Stake, 2004; Stake and Schwandt, 2006). In such an approach, quality is described in the form of narratives. Scriven (1994a) suggests indicators should be avoided if at all possible. They can be tenuous in their validity and open to manipulation.
		Explicit criteria are specified in more detail, using a hierarchical approach. In reports with implicit criteria there was ad hoc attention to defining value terms. Evaluators use a hierarchical approach to move from abstract to more specific levels of detail when defining explicit criteria. This strategy is also evident in reports where criteria are implicit, but the approach is more ad hoc. In 12 reports, criteria are implicit in a KEQ then defined as quantitative indicators.	Sadler (1985) and Scriven (1990) refer to the use of hierarchical approaches to move from the abstract to more specific levels of detail in terms of defining criteria.

Theme	Online survey (n=137)	Review of reports (n=47)	Excerpts from literature
Justifying criteria	96% of respondents consider program objectives to be an important source; 98% always/often use this source. The exploratory factor analysis clustered sources into two factors, of which one focuses on government-based sources. Stakeholders affected by a program are considered important by 95% of respondents and always/ often used by 86% of respondents. 95% of respondents consider literature on 'best practice' to be an important source; 88% always/often use this source. Needs assessment considered important by 77% of respondents but always/often used by only 57%. Australian evaluators 20% more likely to consider policy staff an important or very important source.	More than half (26) reports included reference to program objectives. Government-based documents are the only sources mentioned in one third (16) of the reports. Evidence of stakeholder input into development of criteria was found in less than one third (17) of reports. Relevant literature was evident in 18 reports. Five reports used literature to support the need for a program. External sources (e.g. literature) help to justify government-based sources; and sometimes tempering government expectations about success/quality. Linking sources to criteria was challenging, especially when criteria were not explicit.	Theorists suggest evaluations are most commonly framed around a program's objectives or goals (Mark et al. 2000; Rossi et al. 2004). Criteria should not be justified through sole reliance on program goals, as these may be politically motivated or based on a compromise of different interests rather than oriented around the actual needs of impacted stakeholders (Scriven 1972). Scriven (2013) prioritises need assessment. House and Howe (2000) suggests criteria be oriented around normative ethical perspectives that focus on whether the program is 'right', 'fair' or 'just'. Standard U2 (Program Evaluation Standards) includes an expectation that evaluators attend to all stakeholders with a legitimate interest in a program. Nunns (2017) and Hurteau et al. (2009) suggest that the relevance or validity of criteria, and therefore evaluative reasoning, is strengthened if the criteria are justified, for example by reference to a source.
	Three evaluator stances evident in the qualitative data: • prioritise client-based sources. • seek to include multiple perspective stakeholders • focus explicitly on social justice or equity. Factor analysis: two factors: • sources closely connected to government	Evidence of two evaluator stances evident in the reports. One prioritises client-based sources. The second seeks to include perspectives of stakeholders affected by a program, cultural experts, ethics, program staff and literature relevant to the evaluand.	Shadish et al. (1991) identify two approaches to valuing. Descriptive valuing: evaluators describe stakeholders' values Prescriptive valuing: certain normative values are promoted by the evaluator. House and Howe (2000) advocate for pluralist approaches which deliberately seeks out multiple value perspectives.

Theme	Online survey (n=137)	Review of reports (n=47)	Excerpts from literature
	• sources external to government.		Schwandt (2018) suggests there are three ways that evaluators approach evaluation: (i) social trustee model; (ii) advocacy model; (iii) democratic professionalism model.
	Respondents draw on sources as appropriate to the context of the evaluand and evaluation purpose.		'Evaluands are social, political and moral constructions that embody the different (and often conflicting) interests and values of stakeholders' (Schwandt, 1997, p. 26).
	Respondents say they often or always (96%) draw on content experts to inform criteria, and this source is also considered an important source (94%).	Content experts evident as sources of criteria in only 5 reports. None of the evaluators present themselves as content experts or having intimate knowledge or experience of a subject.	Eisner (1976, 2004) suggest evaluators can reach holistic judgements based on their extensive and intimate knowledge and experience of a subject.
Comprehensiveness	Respondents identify three factors that may inhibit their ability to identify all relevant criteria: • lack of time to engage • lack of budget • political imperatives.		Evaluation is informed by priorities and interests of the government of the day. The political context determines what type of evidence is valued and therefore what evaluation methods are regarded as valid and trustworthy. (Nunns et al. 2015)
	88% of respondents consider cultural expertise to be important; 75% always/often use cultural expertise as a source. Use of cultural expertise and views about its importance were higher for New Zealand respondents. In the qualitative responses, 14 respondents explicitly identified an equity focus as important.	Half (27) the reports had a focus on indigenous peoples. Two New Zealand (and no Australian) reports include reference to the use of indigenous cultural experts. There is evidence in several reports that criteria for cultural aspects of the program have been excluded. Little evidence in the sample of reports of an explicit social justice focus (1 report)	A number of scholars consider it important to ensure there is a focus on those most disadvantaged in society, e.g. Greene 1997; House and Howe 2000; Mertens 2009.

7.3 Current approach to managing criteria

This section considers the findings from phase one as presented in Table 25. The findings are discussed in three sections: (i) fully defining criteria; (ii) justifying criteria through authoritative sources and (iii) comprehensive criteria.

7.3.1 Fully defining criteria

The literature identifies two broad strategies to managing the values against which judgements are made. Those arguing for explicit criteria, e.g. Davidson (2005, 2014a) suggest that this strategy can better support defensible evaluative reasoning. When explicit, criteria are fully defined or described. At a high level, criteria are abstract value terms that are then further defined in more detail (Sadler 1985). A second approach, as evident in the work of Stake (2004) and Stake and Schwandt (2006) argue that explicit criteria can be too limiting,. Their preferred approach to establishing value premises is emergent and implicit, and quality or success are represented in the form of narratives.

The review of published evaluation reports identified no examples of criteria being described through narratives. A more prevalent approach was to use key evaluation questions to orient the inquiry and to indicate at the outset the dimensions of quality or success against which judgements would be (and were) made. However, the meaning of abstract value terms embedded in evaluation questions, or elsewhere, was not always defined or clear even at the conclusion of the report.

Being explicit does not guarantee that criteria are defensible. However, it does serve to reduce ambiguity. The evaluator is likely to be in a stronger position to establish the foundation for an evaluative argument if they begin by defining the value premises, even if these emerge through the evaluation. Likewise, if criteria are explicit, those interested in the evaluation are provided with essential information to then understand and assess the evaluative claims being made.

7.3.2 Justifying criteria through authoritative sources

Scholars justify criteria through reference to different types of sources. For example, Scriven (2013) prioritises the needs of consumers, while House and Howe (2000) suggest criteria be oriented around normative ethical perspectives that focus on whether the program is ‘right’, ‘fair’ or ‘just’. The program evaluation

standards include an expectation that evaluators attend to all stakeholders with a legitimate interest in a program (Yarbrough et al. 2011).

The findings from the online survey and review of reports provide evidence that confirms the views of scholars such as Mark et al. (2000) and Rossi et al. (2004) that evaluations are most commonly framed around a program's objectives or goals. In some of the reports reviewed, the inclusion of a non-government-based source, for example literature, did serve to support the inclusion of a government source, usually the program goals. However, in other reports, a source served to temper government stakeholders' expectations, for example when there was information that program outcomes, or the timeframe of expected outcomes, were unrealistic.

Few reports included sources that referenced the actual needs of a stakeholder group or explicitly oriented criteria round normative ethical principles such as equity or justice. Also, few reports included a reference to consulting stakeholders beyond government officials, content or cultural experts. A managerial client-based approach to valuing was also clearly evident in the online survey findings.

Nunns (2017) and Hurteau et al. (2009) suggest that the relevance or validity of criteria, and therefore evaluative reasoning, is strengthened if the criteria are justified, for example by reference to a source. In the current study, sources were evident but rarely identified in the same section of the report as the criteria. Mere reference to a source is not sufficient to justify criteria. Where criteria are defined through reference to government-based sources only, the risk is that quality or success is defined through too narrow a lens, potentially ignoring or misrepresenting the value perspectives of other stakeholders with a valid interest in a program. Where this occurs, evaluative conclusions have the potential to be indefensible and result in harm. Beyond noting the limitations of current practice, the findings from the current study provided few clues about how best to approach justification of criteria.

7.3.3 Comprehensive criteria

The findings from phase one confirm Shadish et al's (1991) hunch that most evaluators adopt a descriptive approach to valuing. One approach to descriptive valuing is to take the program's objectives and the client's perspectives on quality to define criteria. Evidence from the current study indicates that needs

assessments, cultural input and perspectives on justice and equity are considered important by some evaluators. However, evaluators face several challenges when it comes to accessing these sources. Some of these challenges relate to external factors, such as the political context or resourcing constraints that make it difficult to access wider perspectives. Other challenges relate to the lack of specific evaluation tools that can help evaluators identify and select all relevant values.

Whatever the reasons for limiting the use of sources, the findings from the review of reports highlight that criteria can be too narrowly defined, and in some cases, omitted. An example of this was the 27 reports that included a focus on indigenous peoples. Only three included a reference to the use of cultural experts and there was evidence that criteria for cultural aspects of programs had been excluded in some reports.

7.4 Theoretical framework of characteristics of defensible criteria

Drawing on the literature findings from phase one, a theoretical framework of the characteristics of defensible criteria is further refined from those outlined in section 2.4.3.

Characteristic	Description of characteristic
Description	Criteria are fully described. Describing criteria is not a requirement of defensibility, but without the description, criteria are invisible. Explicit criteria support the reader to recognise how criteria are being defined in the context of this evaluation.
Justification	Mere reference to a source is insufficient. The evaluator must include reasons why sources should be considered trustworthy or authoritative.
Comprehensiveness	The evaluator needs to provide evidence that all relevant value perspectives have been considered and where relevant, included. This means including value perspectives beyond the client and program objectives.

7.5 Chapter summary

In this chapter, findings from a survey of evaluation practitioners and a review of published evaluation reports were considered together to understand how criteria are managed in Australian and New Zealand program evaluation practice. In turn, the findings were considered alongside literature on approaches to valuing, as they

relate to (i) describing criteria, (ii) justifying criteria and (iii) comprehensiveness of criteria.

The findings identify a misalignment between theory about evaluative reasoning – what scholars think ought to occur – and practice, what evaluators say they do and what is observed in program evaluation reports. This was evident in several ways. First, while scholars refer to systematically and explicitly managing value, this approach was rarely evident in practice. Rather, ad hoc attention is given to fully describing criteria.

Second, scholars argue that criteria or values be justified through reference to different types of authoritative sources. However, the findings from this study indicate that some evaluators commonly frame evaluations around a program's objectives or goals, without further warranting these sources.

Third, many scholars (e.g. Greene 1997; House and Howe 2000; Mertens 2009) note the importance of incorporating the value perspectives of program impactees. The findings from this study indicate that criteria often have a narrow focus that excludes the perspectives of stakeholders with a legitimate stake in a program.

Understanding how evaluators currently manage the development of criteria is important as it provides a basis from which to consider how practice might be improved, to better support evaluative reasoning. In phase two of the study a theory-informed conceptual model is developed, with a view to supporting practitioners to systematically identify criteria for program evaluations.

PART THREE: TOWARDS SYSTEMATIC DEVELOPMENT OF CRITERIA

Part three comprises phase two of the study. It documents the development of a conceptual framework for identifying all relevant dimensions of value for an evaluand (Chapter 8) and the process used to translate the conceptual framework into a tool for evaluation practice (Chapter 9). Field testing the tool directly addresses RQ#3: What is an evidence-informed conceptual framework for defensible criteria that can be operationalised for evaluation practice in the public sector? (Chapter 10).

8 Normative ethics and criteria development

8.1 Purpose of this chapter

This chapter picks up the opportunity presented at the end of Chapter 7, to develop and evidence-based and theory-informed conceptual framework for identifying evaluation criteria. The first section outlines a rationale for a conceptual framework that draws on both prescriptive and descriptive valuing approaches. In the second section, the conceptual framework is introduced. A summary of the chapter follows.

8.2 Rationale for a conceptual framework for identifying defensible criteria

In Chapter 2, two characteristics – authoritative sources and comprehensiveness – were identified as critical to the development of defensible criteria. The importance of these characteristics was further confirmed in the review of evaluation reports (Chapter 6). With regard to the first characteristic, Shadish et al. (1991) argued that prescriptive theories, located in moral philosophy, provide such authority as they deal directly with conceptions of 'good' and 'bad'. Still, they suggested evaluators steer away from this approach to valuing on the basis that the evaluator could not justify including one prescriptive theory and ignoring others. Such a position, they argued, was not consistent with democratic political systems that foster pluralism of values.

What Shadish et al. (1991) fail to discuss in their dismissal of prescriptive valuing is that the alternative - descriptive valuing – can also derive value from normative sources. That is, when stakeholders espouse views about what is important in a program, they may implicitly draw on normative perspectives, that is, what they think the program ought to do or not do.

Rather than dismiss prescriptive valuing completely, an alternative approach might involve drawing on both prescriptive and descriptive valuing approaches. First, prescriptive valuing has the advantage of drawing explicitly from ethical values. As noted in Chapter 2, there are three competing normative perspectives through which concepts such as 'good' or 'bad' can be assessed. In western philosophy, normative ethical theories align to one of three – consequentialist, deontological or

Mathea Roorda #579498

virtue-based perspectives. Many scholars acknowledge the relevance of normative ethics to identifying dimensions of value, yet there is little evidence of explicit engagement with normative perspectives in Australian and New Zealand evaluation practice. Second, descriptive valuing has the advantage of understanding value from the perspective of all relevant stakeholders in a program. Many scholars espouse the importance of involving stakeholders in evaluation decisions (e.g. Brandon 1998; Brandon et al. 2010; Greene et al. 2006; Mark and Shotland 1985; Patton 2008; Wehipeihana, Davidson, and Mckegg 2010). Evaluation practitioners also consider stakeholders' perspectives to be important (as identified in Chapter 5) yet the review of reports (as discussed in Chapter 6) suggested stakeholder engagement was limited in many evaluations.

In the next section, aspects of both prescriptive and descriptive valuing are brought together to inform a conceptual framework for systematic development of criteria.

8.3 Introducing the conceptual framework

The framework presented in this section incorporates a prescriptive approach to valuing, drawing on the three normative perspectives introduced in section 2.5, as well as descriptive valuing, that is, describing dimensions of value for legitimate interest groups by drawing on authoritative sources as relevant to the context of an evaluand and evaluation.

In addition to the findings from phase one, the conceptual framework has been informed by two resources. The first is a matrix developed by Mepham, Kaiser, Thorstensen, Tomkins and Millar (2006) that was originally designed to support the analysis of the ethical issues raised by the development and use of emerging biotechnologies. Mepham et al's (2006) matrix applies a set of three principles, derived from two normative perspectives, to a set of selected groups with a legitimate interest in a program. The framework introduced here differs from Mepham et al. (2006) in that it adopts Newman and Brown's (1996) classification of normative perspectives into five principles, as illustrated in Table 26.

Table 26: Comparison of principles used in two normative frameworks

Ethical perspective	Newman and Brown's (1996) principles	Mepham et al's (2006) principles
Consequentialist	Consequences (maximising the good)	Wellbeing (maximising the good)
Deontological	Duty (meeting one's obligations)	Autonomy (treating others as 'ends in themselves')
	Social justice (being fair, just)	Fairness (respect for justice, as advocated by Rawls)
	Rights (doing the right action)	
Virtue-based	Ethic of care (taking care of relationships and context)	

The reason for including Newman and Brown's (1996) principles is that they locate the three major normative ethical perspectives in a way that makes them more easily identifiable in evaluation practice. They also include virtue-based ethics, which is missing from Mepham et al's (2006) matrix. As indicated in findings from the review of evaluation reports (see section 6.4.3), many programs address issues related to consequences (e.g. maximising outcomes), duty (e.g. meeting contractual obligations) or social justice and rights. Ethic of care-related values were much less evident across the evaluation reports reviewed. The researcher's hunch is that virtue-based perspectives may also relevant to many evaluations but less likely to be considered or included in the way criteria are described or defined.

The descriptors of the five principles included in the conceptual framework are drawn from Newman and Brown (1996). There are overlaps between the principles, but also key differences that are intended to differentiate between them, as outlined in Table 27.

Table 27: Definition and focus on five normative principles

Principles	Defined as:	Focus is on:
Consequences	Maximum benefits; action that achieves the best outcomes	Outcomes, effects, value for money, where the focus is on return on investment
Duty	Meeting obligations	Contractual obligations, commitments (including informal commitments)
Rights	Minimum protection; doing the right action	Minimum protection, rights of individuals, groups, entities
Equity	Being fair, just, equitable	Equity, fair deal for those most at risk/disadvantaged
Ethics of care	Being virtuous, e.g. trustworthy, loyal	Meeting the needs of others; attention to context and strengthening relationships.

The five principles form the columns of the conceptual framework, as depicted in Figure 17. Thus, the task of identifying criteria occurs through a normative ethical lens. The framework also recognises that there is more than one way to conceptualise 'good'. In this respect, the framework embraces a pluralist approach to identifying what is important or of value. Specifically, the framework gives equal footing to each principle.

All interest groups with a unique stake, or vested interest, in a program comprise the rows of the matrix. Attention to the value perspectives of all legitimate interest groups is consistent with U2 'attention to stakeholders' and 'U4 'explicit values' of the Program Evaluation Standards (Yarbrough et al. 2011) and the second principle: ethic of care of the Evaluation Standards for Aotearoa New Zealand (Superu and Aotearoa New Zealand Evaluation Association 2015). As with the principles, the framework gives equal footing to each interest group.

Depending on the program, interest groups might also include non-humans, as in the example of a land management program, where depending on the program, interest groups might include animals or tracts of land. Together, the two dimensions of the framework aim for comprehensiveness of value perspectives.

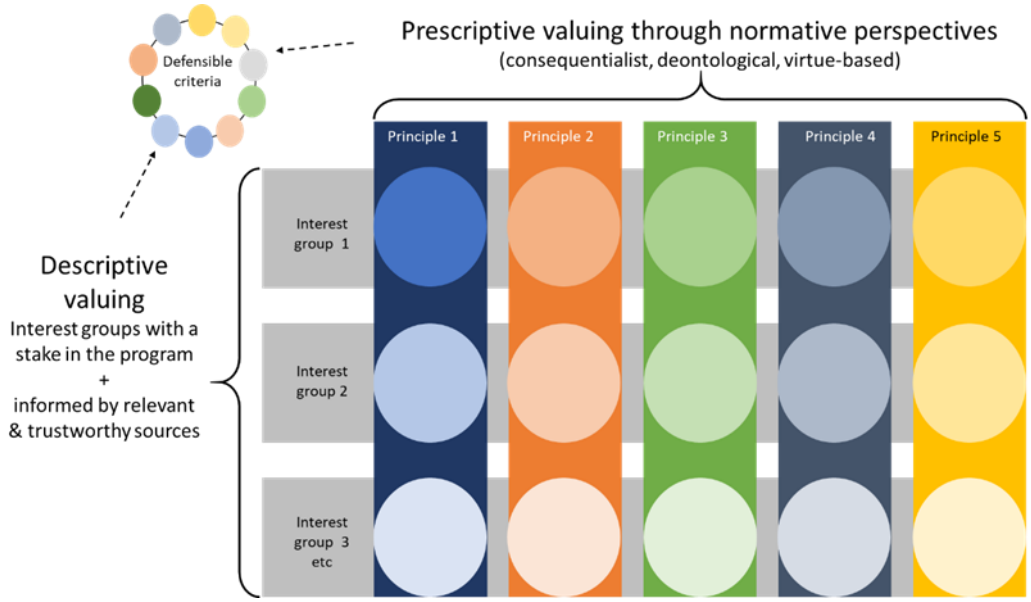


Figure 17: Conceptual framework for identifying dimensions of value

A systematic approach to identifying criteria using the framework would include firstly identifying dimensions of value for each of the interest groups. By example, if Interest group 1 is a government agency, and the first principle is concerned with consequences, the evaluator would seek to understand the action that will achieve the best outcomes from the perspective of that interest group. The same information would be sought if Interest group 2 is a forest. The evaluator would draw on authoritative sources, such as experts, needs assessment or literature, to describe or define dimensions of value for each interest group, through the lens of each normative principle. Up to five statements describing dimensions of value would be developed for each interest group. Depending on the context, some normative principles might not be relevant for an interest group, and thus, the 'cell' of the matrix would remain empty. In some cases, different interest groups might share a similar value perspective. One of the potential advantages of the framework is the ability to systematically examine dimensions of value for each interest group and to identify where there is alignment of values and where there are differences.

The task of identifying dimensions of value might be undertaken as a desk-based exercise, with the evaluator drawing on a range of sources such as relevant literature, program documentation and key stakeholder interviews. Alternatively, dimensions of value might be identified with stakeholders in a facilitated workshop, or in conjunction with a desk-based approach.

Once dimensions of value have been identified for all legitimate interest groups, a second step involves analysing the information and dimensions of merit which may be grouped into a set of criteria. For example, analysis of dimensions of value may result in a high-level criterion of 'effectiveness', with three or four sub-dimensions, each of which can be further defined based on information from the matrix.

Alongside these two steps, the evaluator must consider the trustworthiness of both the sources and the information that results from the process of identifying dimensions of merit. In an evaluation, this requires the evaluator to do more than just state the criterion against which a judgement is made. It is not sufficient to take stakeholders' (or the evaluator's) view of the criterion at face value. They must provide some confidence that supports the inclusion of a criterion and the way it has been defined. In practice this might include providing references to key sources or linking criteria to relevant background information about the evaluand. Providing this information serves to support the defensibility of criteria. The importance of this step is illustrated through Taylor's (1961) imaginary dialogue in which he demonstrates the importance of justifying the value premises in evaluative reasoning (see section 2.4.3).

The conceptual framework is intended to address several needs identified in phase one of the study. First, it supports criteria to be explicitly and transparently described or defined. The framework takes a grounded approach, with criteria of merit emerging from a consideration of all legitimate interest groups, as they connect with each normative perspective. Second, it prompts the evaluator to identify reasons for including different value perspectives, rather than take them at face value. Most importantly, the framework is intended to address the issue of criteria being too narrowly defined and to support the development of a comprehensive set of criteria.

8.4 Role of the framework in developing defensible criteria

In Section 2.3.3 the researcher suggested that each criterion needs its own supporting argument to support evaluative reasoning. The conceptual framework supports the development of an argument about criteria in two ways which are illustrated in Figure 18. First, it requires the evaluator to draw on relevant information or evidence (Rex, Thomas and Engel 2010) to support a claim about value. Such evidence might draw on sources such as needs assessment, which

measures gaps 'between the current (what is) state of affairs of a group or organisation and the desired (what should be) state in regard to variables of interest' (Altschuld, Kumar and Mathison 2005, p. 277), or content expertise (Stake 2001). Second, the conceptual framework requires the evaluator to systematically consider relevant dimensions of value.

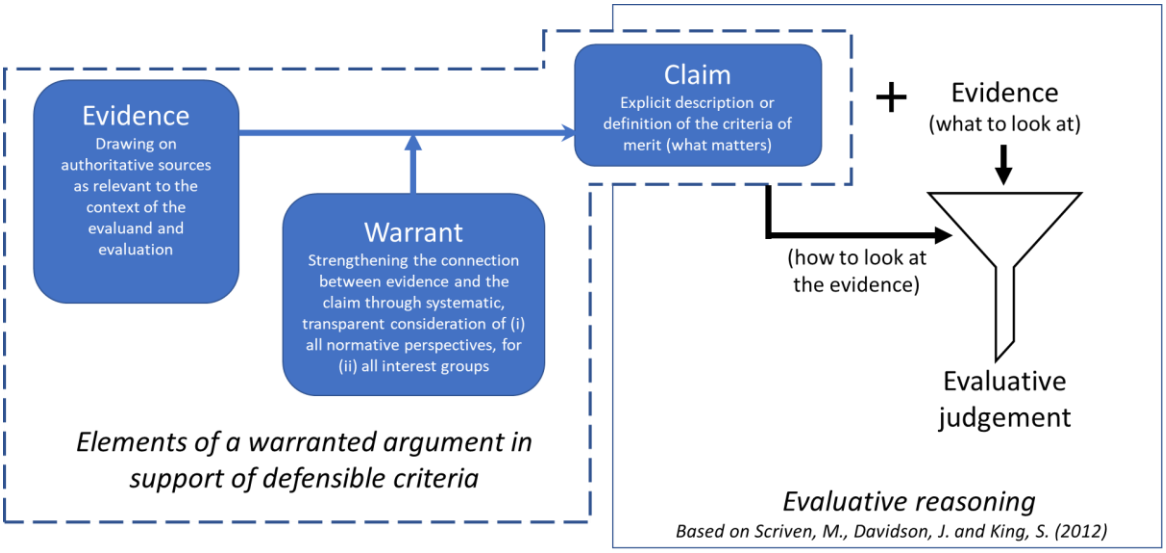


Figure 18: Description of an argument to develop defensible criteria

By identifying dimensions of value for all relevant interest groups, through the lens of all relevant normative perspectives, the evaluator reasons from authoritative sources to a claim about criteria of merit. In this respect, the conceptual framework draws on both ethical and methodological warrants (Rex et al. 2010). The ethical warrant comes from consideration of three normative perspectives and five ethical principles, while the methodological warrant draws on stakeholder-based or participatory approaches to evaluation design. Figure 18 illustrates the link between warranted criteria and the broader evaluative reasoning process as depicted by Scriven, Davidson, and King (2012).

8.5 Chapter summary

This chapter has introduced a conceptual framework for identifying dimensions of value. The framework draws from both prescriptive and descriptive valuing traditions. Set out as a matrix, the framework comprises five ethical principles that derive from three normative ethical perspectives – consequentialist, deontological and virtue-based. These perspectives and corresponding principles make up the

columns of the matrix. The rows comprise the legitimate groups with a vested interest in a program. The framework supports the development of an argument about criteria of merit by drawing on both ethical and methodological warrants.

In Chapter 9 a criteria matrix handbook is developed, with a view to turning the conceptual framework into an evaluation-specific, practice-based tool.

9 Development of a criteria matrix handbook

9.1 Purpose of this chapter

To have utility in practice, the conceptual framework (as described in Chapter 8) needs to be implementable by evaluation practitioners. In four sections, this chapter describes the development of a criteria matrix handbook intended to translate the conceptual framework into a practical process. First, the rationale for developing the criteria matrix handbook is outlined. Second, an overview of the handbook is provided. Third, the process for developing and refining the handbook is described. This drew on an approach used by the Evaluation Checklists Project. Findings from each stage of the review are documented. The fourth, and final, section is a reflection on the review process.

9.2 Rationale for developing a criteria matrix handbook

Early drafts of the conceptual framework were presented to evaluators at three evaluation conferences in 2016, namely the ANZEA conference in Wellington, the European Evaluation Society (EES) conference in Maastricht and the AEA conference in Chicago. The framework was also presented at an AES lunchtime seminar in Wellington in 2016 and at three webinars organised for students enrolled in a Practice of Evaluation subject at the Melbourne Graduate School of Education (May and November 2018 and May 2019). In the discussion following each of these conference, seminar and webinar sessions it was evident that evaluators were familiar with applied ethics and ethical conduct of evaluation, but less familiar with normative ethics. Use of the word 'ethics' confused participants, so in later seminars and webinars, it was replaced by the term 'normative perspectives'. Moreover, unless participants had studied philosophy, few had encountered terms such as consequentialism, deontology and virtue-based ethics. Once they understood the concepts and had been taken through the framework and an example of using it to identify criteria, there was a sense that the framework could be useful for evaluation practice. This was evident from the discussions that followed each presentation, and email requests for a copy of the PowerPoint slides.

It was also clear from these forums that the conceptual framework was not sufficient on its own. Rather, a detailed resource or set of resources would be required to help evaluators engage with the concepts. While keen to develop the

Mathea Roorda #579498

conceptual framework into a tool for practice, the researcher was also concerned that the tool would not be used in a mechanistic way, as can sometimes occur when people are given step-by-step instructions. Some effort was therefore given to developing a resource, the 'criteria matrix handbook' – hereon referred to as 'the handbook' that would provide clear guidance to evaluation practitioners so they could identify dimensions of value and then use this information to develop a justified set of criteria of merit.

9.3 Overview of the criteria matrix handbook

The purpose of the handbook was to provide evaluators with the information required to identify and develop criteria based on normative ethical perspectives and through use of a criteria matrix. The matrix is based on the conceptual framework as outlined in Chapter 8.

The handbook comprises two sections. The first section includes an introduction to evaluation criteria and the conceptual framework. It outlines why systematic development of criteria is important for reaching a defensible evaluative judgement and how consideration of normative ethical perspectives might help to achieve that aim. Some guidance for using the framework in different circumstances is provided, including at different stages of an evaluation and using it as a desk-based exercise or in a workshop setting. Next, evaluators are introduced to the different elements of the conceptual framework; that is the interest groups and the five normative principles.

The second section of the handbook sets out the steps for using the framework. These include customising the interest groups and normative principles for a specific evaluation, then identifying dimensions of value. A completed matrix for a seasonal labour initiative is included in the handbook as an example of the type of statements that might populate each cell. A list of potential sources is also included; a similar list to that provided to the survey respondents in phase one. The next step describes taking the information from the cells in the matrix, that is, the dimensions of value, and developing criteria of merit. Again, an example of a criterion is provided, based on an evaluation of a seasonal labour initiative. A subsequent step is to justify the criteria against which judgements will be made. A rationale for including this justification links criteria to evaluative reasoning.

A copy of the final version of the handbook, as provided to the field testers, is included in Appendix F.

9.4 Process for developing and refining the handbook

The development and refinement of the handbook was guided by the validation process of the Evaluation Checklist Project. Located at the Western Michigan University Evaluation Centre, the Evaluation Checklist Project (Checklist Project) is a repository of evaluation checklists that aim to distil and clarify:

> relevant elements of practitioner experience, theory, principles, and research to support evaluators in their work (Western Michigan University Evaluation Centre, 2016, p. 1).

While not conceptualised as a checklist as such, the handbook has the same aim; that is, to distil theory into an operational task. Accordingly, the approach used to develop and revise the information required to use the normative framework was submitted to a similar process to that used in the Checklist Project. To differentiate between the conceptual framework and the resource developed to support its use, the handbook was given a working title 'criteria matrix handbook'. Within the handbook, the term 'criteria matrix' is used to refer to the normative framework.

9.4.1 The Checklist Project validation process

The 'validation process'[45] used by the Checklist Project comprises two consecutive steps: expert feedback followed by field testing.

- The first step, expert feedback, has four sub-steps, as follows:

  1. Expert review of the draft document

2. Revision of the document based on expert feedback

3. Additional expert review if needed

4. Preparation of the document for field testing.

- The second step, field testing, aims to ensure checklists are useable and have tangible benefits when used in evaluation practice. Field testing

[45] This is the term used by the Checklist Project to describe their process for developing checklists.

includes recruiting field testers and revising the document based on field test results.

To be accepted for inclusion in the Evaluation Checklist Project collection, checklists have to meet six criteria (Western Michigan University Evaluation Centre, 2016, p. 1). Each criterion includes sub-criteria, which serve several purposes, including defining explicitly key value terms such as 'appropriateness' and 'clarity'. Being explicit about the meaning of such terms is especially important when there are multiple people as it supports more consistent feedback. The sub-criteria also act as prompts to support a more nuanced and fuller assessment. The entire set of criteria and sub-criteria are included here to show how each criterion is defined.

Appropriateness of evaluation content

- The checklist addresses one or more specific evaluation tasks (e.g. a discrete task or an activity that cuts across multiple tasks).
- The checklist clarifies or simplifies complex content to guide performance of evaluation tasks.
- Content is based on credible sources, including the author's experience.
- Content is consistent with the Program Evaluation Standards (Yarbrough et al. 2011) and the AEA's Guiding Principles for Evaluators and Statement on Cultural Competence in Evaluation.
- Content does not overtly favour one evaluation approach over others, unless the checklist is intended to support application of a particular evaluation approach.

Clarity of purpose

- A succinct title clearly identifies what the checklist is about.
- A brief introduction orients the user to the checklist's purpose, including the following:
    - The circumstances in which it should be used

- How it should be used (including caveats about how it should not be used, if needed)
- Intended users

Completeness and relevance

- All essential aspects of the evaluation task(s) are addressed.
- All content is pertinent to what users need to do to complete the task(s).

Organisation

- Content is presented in a logical order, whether conceptually or sequentially.
- Content is organised in sections labelled with concise, descriptive headings.
- Complex steps or components are broken down into multiple smaller parts.

Clarity of writing

- Content is focused on what users should do, rather than questions for them to ponder.
- Everyday language is used, rather than jargon or highly technical terms.
- Verbs are direct and action oriented.
- Terms are precise.
- Terms are used consistently.
- Definitions are provided where terms are used but might not be obviously known.
- Sentences are concise.

References and sources

- Sources used to develop the checklist's content are cited.
- Additional resources are listed for users who wish to learn more about the topic.
- A preferred citation for the checklist is included (at the end or beginning of the checklist).
- The researcher's contact information is included.

These six criteria guided the development of content and organisation of the handbook in two sections. The first included an introduction to criteria and to the criteria matrix; the second described the steps for using the matrix to develop dimensions of value and then construct and justify criteria of merit. In terms of appropriateness of the handbook content, the aim was to provide sufficient information for the normative principles and justification process to be clearly articulated, but not to prescribe the process too rigidly. Behind this thinking was a view that the matrix should be applicable to all program evaluation regardless of approach. For example, evaluators conducting a participatory-based evaluation might use the criteria matrix in a workshop with stakeholders, while evaluators using non-participatory approaches might do a desk-based exercise, drawing on existing documentation and other secondary sources to complete the matrix.

9.4.2 Adaptation of the validation process for the criteria matrix handbook

The approach used by the Checklist Project was adapted for this study in two ways. The first adaptation was to include a seventh criterion on 'content validity'. The rationale for including a seventh criterion is discussed in section 9.4.3.

The second adaptation was to separate the experts' review into two steps. In step one, the review focused on the accuracy, completeness and relevance of the information provided about the normative perspectives and principles. In step two, the review had a broader focus on the entire handbook.

Table 28 sets out an overview of the criteria used at different stages of the design and review of the criteria matrix handbook.

Table 28: Criteria for developing and reviewing the handbook

		Validation process used in this study	
Criterion	**Development of the handbook**	**Step 1: Review of normative ethical perspectives and principles**	**Step 2: Review of the criteria matrix handbook**
Accuracy	x	x	
Appropriateness of content	x		
Clarity of purpose	x		x
Completeness and relevance	x	x	
Organisation of content	x		x
Clarity of writing	x		x
References and sources	x		

The approach used in Steps 1 – 2 along with findings for each respective step, is outlined in the following sections.

9.4.3 Step 1: Content validity of the ethical principles

The researcher sought to provide an accurate and relevant description of the normative perspectives and principles included in the criteria matrix. Therefore, the first step of the validation process was an expert review which focused on content validity of a set of statements about the normative principles that were included in a section of the handbook entitled 'What are the normative principles?'

Checking for content validity was particularly important because the definitions included about each of the normative ethical perspectives were reduced to a few brief sentences in the handbook. The aim was to provide sufficient information for evaluators to understand each principle, but not overwhelm them with an overly complex and academic description. Hence, this phase of the review focused on (a) identifying whether the descriptions for the normative principles were accurate, and (b), drawing on the Checklist Project criterion of relevance, identifying whether the information was pertinent to what evaluators needed to know to undertake the task of identifying dimensions of value.

A snowball approach was used to identify evaluation scholars with known expertise in ethics. A description follows about how this approach worked in practice. The first scholar contacted was a New Zealand based evaluator, a former

chair of the AES Ethics Committee, and a facilitator of numerous professional development seminars on ethics. They readily agreed to participate as one of the expert reviewers. The second scholar contacted was a US based evaluation and philosophy scholar. They responded to say they were not able to participate as an expert reviewer, citing work commitments, but were able to provide contact details for several other scholars. A subsequent email to these scholars resulted in a second reviewer agreeing to participate in the study. This scholar was a former chair of the AEA Evaluation Ethics' Committee and an author of numerous texts and papers on ethics. The researcher contacted four other evaluation scholars. One person responded, stating they were too busy to participate, and three others did not respond.

The two expert reviewers who agreed to participate were sent a copy of the draft handbook and a review feedback form (Appendix D). They were asked to review the information on each normative perspective and the five principles, as outlined in the draft handbook. In addition to feedback on the accuracy and relevance of the information, the expert reviewers were invited to outline suggested changes to text where appropriate. The expert reviewers did not receive any gratuity. Both agreed to be named in the acknowledgements section of the thesis.

The feedback from the two scholars was collated into one table and the researcher considered how to respond to each comment. The researcher's actions were noted in a separate column, to facilitate a transparent review process. Changes were then incorporated into the document in preparation for review of the whole criteria matrix handbook.

9.4.4 Step 1: Findings from the expert review

Although the two ethics' scholars, referred to hereafter as Ethics Scholar 1 and Ethics Scholar 2, did not describe the same concerns about the normative perspectives and principles, each pointed to areas that required further clarification. Their comments are listed in full in Table 29, along with the researcher's actions regarding changes to the handbook.

The ethics scholars' feedback indicated the normative perspectives and principles were, in general, accurately defined. Much of their feedback centred around providing additional information to further clarify the principles, and to ensure they were not too narrowly defined. Both experts questioned whether it was

possible to disentangle the principles included in the deontological perspective, and in particular, rights and duty or rights and justice. The researcher's response was to keep all five principles in the framework and to suggest to evaluators that they might want to customise it as relevant for their evaluands and evaluations.

Ethics Scholar 1 also raised an issue about whose dimensions of value are included in a cell. They noted an interest group such as an administrator might have expectations about the outcomes they were seeking for another group (for example, the target group). Would this information be recorded under outcomes for the target group? This was indeed how the researcher envisioned the framework being used. Using the example provided by Ethics Scholar 1, the views of the administrator could inform an understanding of the outcomes expected by the target group. As such, the rows in the matrix were dimensions of value for an interest group, not the views of a specific stakeholder group. The difference was subtle but important. Dimensions of value for a target group (or any other interest group) could be informed by a number of sources such as literature, stakeholders representing the target group, legislation, as well as the administrator. Critically, a final step for using the framework requires evaluators to justify criteria. Using the above example, the evaluator would need to provide a reason for including the administrator's perspective on outcomes. Articulating this information, both in the handbook and to the field testers in the briefing session, was challenging, and in retrospect, is an aspect of the process of identifying dimensions of value that requires additional clarification in a future iteration of any resources developed.

Table 29: Expert feedback on the normative perspectives and principles

* The reviewers were asked: A: The definition included is accurate? If not, what changes would you include? B: The information provided is pertinent to what evaluators need to know to undertake the task (identifying dimensions of value). If not, what changes would you include?

Principle	Ethics Scholar 1	Ethics Scholar 2	Researcher's actions in response to the ethics scholars' feedback
Consequences	• The definition is accurate. • A question: If an interest group that was NOT the target of the program (e.g. administrators) was concerned about outcomes of the program for the target group, would that concern be mentioned in the non-target's matrix cell for maximising outcomes? Or would that cell only be used for outcomes that directly affect the non-target group?	• No comment on whether the definition is accurate. • The comment on the limitations of a consequentialist approach doesn't seem to go far enough. One issue here is that utilitarian approaches may be used to justify unjust policies or practices, citing the greatest good of the greatest number. There is also a question about whether consequences of a policy or program can be easily identified as the sum of consequences for a range of people, who may assess the value of costs or benefits differently.	Ethics Scholar 1: Clarify in the handbook and briefing with field testers that different sources may inform dimensions of value for a specific interest group. Ethics Scholar 2: Revise the section on limitations as suggested.
Duty	• The definition refers to 'expectations, commitments, and contractual obligations.' I suggest deleting 'expectations'. One can have expectations based on one's hopes and wishes that do not necessarily translate into duties that another party must abide by. For example, the fact that I might expect a job training program to provide me with follow-up services after I obtain a position does not mean that the program has necessarily committed itself to doing that. Perhaps a	• No response. • The description of this ethical stance emphasises the duty of a manager to carry out a program as intended and as required through contracts or other means. It would seem to me, however, that managers still have other obligations or duties to consider, such as an obligation to follow rules of due process and respect human rights. For an evaluation, therefore, you may want to interpret duty obligations two ways: as formal	Ethics Scholar 1: Remove 'expectations' from definition. Ethics Scholar 2: Broaden definition of 'duty' to reflect formal and implicit obligations.

Principle	Ethics Scholar 1	Ethics Scholar 2	Researcher's actions in response to the ethics scholars' feedback
	distinction between explicit and implicit expectations is warranted here. • The information provided is pertinent.	duties applicable within the programme, and as more general and implicit societal obligations. • There is an overlap between obligations and other elements of the matrix which may be impossible to eliminate but could be acknowledged.	
Rights	A and B – Disentangling this principle from Duty is a challenge. If someone has a right, then that means someone else has a corresponding duty. It's worth noting that you use the word 'obligations' in discussing both Duty and Rights. Unfortunately, I'm not sure how to address this issue within the context of the criteria matrix. What would be lost if you combined them?	I would like to see it clarified whether respecting the autonomy of people affected by a policy or program is a matter of rights or of the ethic of care. I would also like to see it noted that justice is often cited as an element of ethics, involving respect for rights.	Ethics Scholar 1: Revise text as suggested. Note that where there is a right, this will likely mean another interest group has a duty. Remove 'obligations' from text. Ethics Scholar 1 & 2: Keep 'rights', 'equity' and 'duty' as they reflect different values but include a statement that evaluators may want to customise the matrix.
Equity	A – I would add the following to the end of the second sentence: '…conditions that have led to these outcomes and how the program took these conditions into account.' The end of the second sentence might benefit from the following addition: '…a sense of entitlement to outcomes.' Or are you implying a sense of entitlement to outcomes	As noted in the statement, equity requires consideration of conditions that led to the status quo. You might add that equity considerations include recognition that some policies or activities appear equitable on the surface but may have inequitable consequences (for example, rules intended to promote equity may perpetuate existing pay disparities between groups).	Ethics Scholar 1: Include additional text as suggested. Ethics Scholar 2: Comment discussed with reviewer and agreed to leave description as is.

Principle	Ethics Scholar 1	Ethics Scholar 2	Researcher's actions in response to the ethics scholars' feedback
	and other things? If so, what might those other things be? B – I think my comments under A have implications for B.		
Ethic of care	A – The statement 'It includes a focus on 'care' and 'caring'…' begs the question a bit. What, exactly, does it mean to display care and caring when interacting with someone? The reference to 'strength of relationships' is vague, and not treating people 'as a means to an end' is basically a negative definition (i.e., this is what caring is not). Of course, this isn't an easy issue to deal with, given your comment that the contextual nature of care 'can be ambiguous.' But I think additional explication is needed here. B – Of the five principles, this is the one whose legitimacy is most likely to be questioned by stakeholders within the context of an evaluation. For example, it might be seen as violating equity in certain cases. If a diligent, hard-working student needs a particular grade in my class to keep his/her academic scholarship and stay in school, an ethic of care might lead me to consider awarding that grade even if his/her performance doesn't quite reach the required level. But one could argue that doing this	Descriptions of ethics often refer to principles of beneficence and doing no harm (non-maleficence). It's not clear where they fit, but I think this area of an ethic of care comes closest.	Ethics Scholar 1: Text revised to remove negative definition. Ethics Scholar 2: Comment discussed with reviewer and agreed to leave description as is.

Principle	Ethics Scholar 1	Ethics Scholar 2	Researcher's actions in response to the ethics scholars' feedback
	would be unfair to other students whose grades are close to the next level but not elevated. The issue here is not primarily one of ambiguity, which is the limitation you reference in your description.		
Other comments	Nice job! And good luck.	In general, the descriptions of different elements of the matrix are consistent with my own understanding of the literature and are reasonably presented. I like the use of the term duty instead of deontological, for example, as it presents the idea in plainer English (deontological being grounded in the Greek word for duty or obligation), and consequences instead of teleology.	

9.4.5 Step 2: Review of the handbook

Step 2 of the validation process occurred once feedback from the ethics scholars had been incorporated into the draft criteria matrix handbook. As noted earlier, the researcher sought to develop a clear, well-organised resource to support evaluators in their engagement with the criteria matrix. Thus, the purpose of the review process was to verify it was fit for this purpose.

Two rounds of review were undertaken, as illustrated in Figure 19. For the first round, a list of potential reviewers was developed by the researcher based on her evaluation networks in Australia and New Zealand. Of the nine evaluators approached, four responded stating they were not able to participate, either due to work commitments or because they were on annual leave during the timeframe the review was being undertaken. In total, five evaluators, three from New Zealand and two from Australia, agreed to participate as reviewers, and were sent the draft handbook along with a feedback form to complete. The feedback form included three criteria used by the Evaluation Checklist Project: clarity of purpose, organisation of content and clarity of writing.

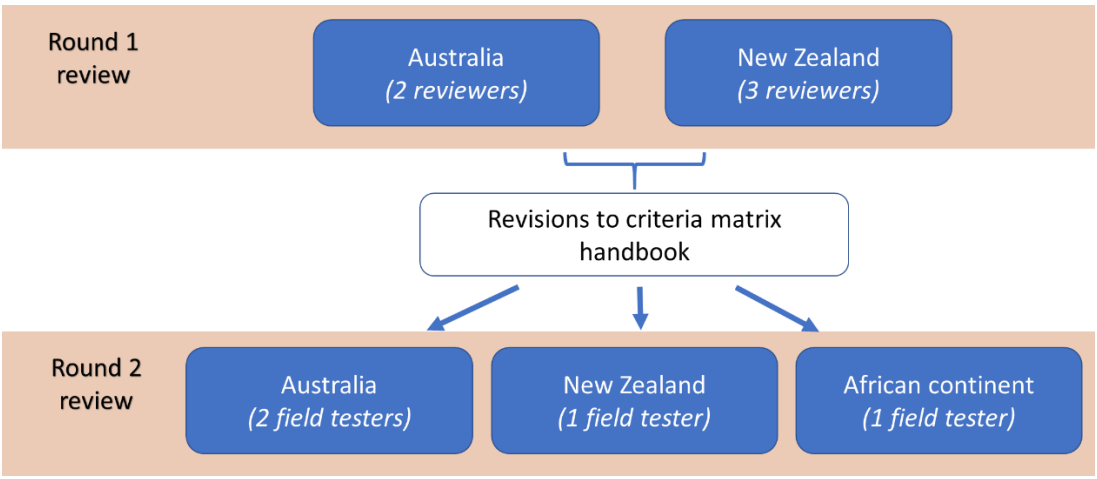


Figure 19: Review of the criteria matrix handbook

The researcher collated reviewers' feedback into a table, along with the number of reviewers providing similar type comments (see Table 30). As with the analysis of the ethics scholars' review, the researcher's actions in response to the feedback were recorded in a separate column of the table. Revisions were then made to the criteria matrix handbook.

The second round of review occurred alongside the field testing of the criteria matrix. Four field testers were invited to use the handbook in whatever way they wished. They were asked to provide the researcher with documentation produced as part of using the conceptual framework and other information about the evaluation context. The handbook also included a feedback form for field testers to complete. This form focused on identifying aspects of the handbook that were helpful or confusing, and information they thought needed to be added, modified or removed.

Only the findings from the review are included in this section. Further information about the process used to identify field testers, and results from their engagement with the criteria matrix, are discussed in Chapter 10.

9.4.6 Step 2: Findings from round one review

Five reviewers provided feedback against three criteria: clarity of purpose, organisation of content and clarity of writing. The number of comments that were of a similar nature are indicated in brackets in Table 30.

Table 30: Reviewers' feedback on the handbook

Criterion	Reviewers' comments (N)	Researcher's actions
Clarity of purpose	Title too succinct. Include information about it being a handbook (1). Add in something about ethics (1)	Title revised to 'Developing defensible criteria: A handbook for evaluators'
	Add a 'purpose of this document' section to the handbook (1)	New statement included at the start of the document describing its purpose
	Needs more information about justification of criteria (2).	Additional information included.
	Include a section on the problem the criteria matrix is attempting to address (1)	New section included: 'Why do we need such a tool'?
Organisation of content	Content is in logical order: Yes (5)	None required
	Concise, descriptive headings included: Yes (5)	None required
	Steps for developing defensible criteria require further detail on process (2).	Additional information provided on each step, plus inclusion of an example.
	Explain the difference between a criteria matrix and a rubric (1).	Separate appendix on evaluative rubrics included.
Clarity of writing	Add a definition of 'criteria of merit' (CoM) (1)	Definition of 'criteria of merit' included

Criterion	Reviewers' comments (N)	Researcher's actions
	Language is very technical/academic (1) Glossary needs to explain words like 'deontological' (1) Explanation of 'ethical standing' is difficult to understand (1)	Additional terms added to glossary Text revised
	Section 1.3 should start with 'The interest groups are…' (1)	Section on interest groups revised.

The researcher made several revisions to the handbook in response to the reviewers' feedback. A subsequent version of the handbook was then given to the four field testers to use. As well as engaging with the conceptual framework, the field testers were asked to reflect on the handbook using the same criteria as those used by the reviewers. An analysis of the reviewers' and field testers' feedback enabled the researcher to consider whether concerns raised by the reviewers had been sufficiently addressed in the revised iteration of the handbook which was used by the field testers. The analysis of feedback against each checklist criterion follows.

9.4.7 Step 2: Findings from round two review

This section includes the analysis of reviewers' and field testers' feedback on the criteria matrix handbook. The analysis is structured in three sections: (i) clarity of purpose; (ii) organisation of content; (iii) clarity of writing; (iv) assessment of the criteria matrix handbook. An analysis of the reviewers' and field testers' feedback enabled the researcher to consider whether concerns raised by the reviewers had been sufficiently addressed in the revised iteration of the handbook which was used by the field testers

9.4.7.1 *Clarity of purpose*

Reviewers: The reviewers' feedback identified a need to include additional information in the handbook to help readers understand the purpose of the criteria matrix tool. This included revising the title to more clearly articulate the purpose of the criteria matrix and providing more information in the introduction section of the handbook about the need for such a tool.

Field testers: The revised handbook provided sufficient information for two of the four field testers. As one of these field testers noted:

> After reading through the first five pages of the tool I felt I had a solid grasp of the normative principles and felt I was comfortable with / able to proceed in beginning the development of my own criteria matrix.

One of these field testers decided to amalgamate two of the principles into one category. They were reassured by the information included in the handbook which gave them permission to customise the matrix for their evaluation:

> 'Equity and rights were too similar to be separate for this project. The project is about disability access and inclusion and 'equity' as a term seemed more relevant'.

For a third field tester, the information appeared clear until they started to develop their own criteria matrix. Despite going back to the section 'What are the normative principles?' they struggled to understand the difference between the principles of equity, fairness and ethic of care. The feedback from the reviewers reinforced comments made by the ethic scholars in which they noted a number of interrelationships between principles.

The fourth field tester was unclear about the concept of interest groups, noting: 'I was thinking about these people as the audience'. This resulted in them viewing each row of the matrix from the perspective of one stakeholder group, rather than value as it related to an interest group (potentially informed by different stakeholder groups). Their lack of clarity about interest groups meant that, after spending 90 minutes on the task, they did not complete the criteria matrix as they were not sure they were using it as intended.

All the field testers noted the importance of including examples in the handbook. One commented that the example provided had helped them feel confident they were interpreting the principles in the same way:

> Until you see an example of the type of wording (required) it's a bit hard to do. It's also good to say that not every cell needs to be filled in. It's good to know it may not be relevant (for a particular program).

Two field testers suggested that more examples, using different types of programs, could be included in the handbook:

> I would add examples of what each of the principles could mean for different programs.
>
> If possible, I would also add additional criteria matrix examples from a range of content areas. For my purposes it would have been great to have an education-related example, so I could cross-check to make sure I was in the realm of doing something relevant.

Moving from dimensions of merit to developing criteria proved to be challenging for one field tester:

> When I got to the stage of converting the matrix (information) into criteria, my brain forgot everything I knew about criteria! In the example provided I wasn't sure what type of analysis enabled a user to get from the matrix to the criteria.

Of the three field testers who progressed to developing criteria, only two provided further detail about justification of criteria. All three suggested that a future iteration of handbook could include additional information and examples to explain this part of the process. As one field tester commented:

> Is there a particular form you'd suggest I use, or some sort of table to make (justification) explicit?

The same field tester found the list of sources included in the handbook a useful guide and noted that previously they had not thought about justifying criteria beyond empirical research.

9.4.7.2 *Organisation of content*

Reviewers: The reviewers' feedback indicated that the handbook was presented in a logical order, with the content labelled with concise, descriptive headings. One reviewer suggested including additional information about the process of moving from the criteria matrix, once dimensions of value had been identified, to that of constructing and defining criteria. In the next iteration of the handbook the researcher included an example of dimensions of value completed for a seasonal labour initiative (Appendix F, Box 3), followed by an example of a criterion developed from the criteria matrix information.

Field testers: None of the field testers had specific feedback about the organisation of the handbook; nor did they raise any concerns. It should be noted that all four field testers participated in a one-hour briefing that included a PowerPoint presentation that introduced the tool and then took them through each step of the process.

9.4.7.3 *Clarity of writing*

One of the key challenges faced by the researcher in developing the handbook was that of using plain English text to describe quite complex ideas from philosophy. The glossary provided a space for specialised academic terms to be defined but not laboured over in the main text. Other ideas had to be explained in enough detail so that evaluators could engage with the task meaningfully, for example providing clear information about each normative principle and enough information to distinguish between them.

Reviewers: The reviewers had mixed views about the extent to which explanations were provided in everyday language. Three reviewers commented that the explanations were described simply, with one noting: 'I thought some quite difficult concepts were described simply'. Two reviewers thought the handbook included some technical jargon that needed to be unpacked. One reviewer was particularly critical about the language used in the handbook, noting it was:

> quite often very technical / academic. Think of a proud Pacific woman – CEO [chief executive officer] of an organisation – read this paper 'aloud' to her.

The reviewer provided constructive advice about including definitions in the glossary.

Field testers: Three of the four field testers found the information in the handbook, including the instructions for using the criteria matrix, clear. As one field tester noted:

> (It) is clearly written, easy to understand and is beautiful! It walks the user through an introduction to criteria and the normative principles in a way that I found to be clear and easy to understand.

A fourth field tester did not comment explicitly about clarity of writing.

9.5 Adoption of the Checklist Project process: a reflection

As noted, the criteria matrix handbook was created to support evaluators to engage with a conceptual framework for systematically identifying dimensions of value, from which a set of defensible criteria could be developed. Both the development of the handbook and the review process were guided by criteria obtained from the Checklist Project, with each step of the process resulting in the handbook being further refined. In this section the researcher reflects on the use of the criteria and process for the current task of turning the conceptual framework into an operational task.

First, the criteria used by the researcher provided a useful guide for writing and structuring the content of the criteria matrix handbook. In this respect, the criteria were used literally as a checklist to ensure each point was covered off.

Next, two ethics experts reviewed content related to the ethical principles. This was followed by a review of the handbook, by five reviewers in round one and four field testers in round two. The feedback from the ethics scholars was particularly valuable in helping to clarify and validate the statements about each of the ethical principles. They provided practical suggestions about wording changes and suggested additional information that they considered ought to be included. Their comments indicate that engagement with each of the normative principles requires a deeper understanding of the concepts than is able to be conveyed in a short statement. Examples of this can be found in the scholars' reflection of the relationship between the principles, for example where a dimension of value framed from an ethic of care perspective might violate a principle of equity. A potential challenge with the handbook is that evaluators who have a limited knowledge of normative ethics engage with the principles in a superficial way. The implications of this may need to be acknowledged in further iterations of the handbook or considered alongside a broader strategy for educating evaluators about normative ethics.

More generally, the ethics scholars' feedback indicated there was value in considering normative perspectives when developing criteria of merit. The scholars engaged with the criteria matrix and did not raise any concerns about its use for the intended task of surfacing dimensions of value. In retrospect this

finding, albeit implicit, served to reinforce the value of engaging evaluators with ethics expertise in the review process.

Moving to the next phase of the field testing, the reviewers provided feedback on three aspects of the handbook: clarity of purpose, organisation of content and clarity of writing. Changes were made in response to their feedback. However, the experiences of those who subsequently used the handbook to develop criteria indicated that it required additional information, particularly with respect to moving from dimensions of value to criteria of merit. This request for additional supporting information presented a challenge for the researcher, in that the aim of the handbook was to support evaluators to engage with normative ethics while not prescribing the process by which they did so. As suggested by the field testers, future iterations of the handbook might include case examples that document how evaluators engage with normative principles in different ways to develop criteria of merit. More broadly, the difficulties evaluators had with understanding the principles, for example differentiating between equity, fairness and ethic of care, highlights the lack of knowledge many evaluators have about these important concepts. The findings also highlight the challenge in presenting quite complex information in language that is accessible to practitioners.

9.6 Chapter summary

This chapter has documented the development and review of a criteria matrix handbook intended to support evaluation practitioners to engage with normative ethical principles. The review was guided by the validation process used by the Western Michigan University Evaluation Centre Evaluation Checklists Project. The process was useful both for developing and refining the handbook. The findings from the review process highlight the challenges associated with developing resources for concepts that are complex and require evaluators to engage with a subject area where they have no, or little previous knowledge.

10 Practitioner engagement with normative ethics

10.1 Purpose of this chapter

This chapter explores the applicability of the conceptual normative ethics framework for program evaluation practice, and specifically whether it supports the development of defensible criteria (defined as explicit, fully described criteria that include all relevant dimensions of value and based on authoritative sources).

10.2 Research design

This section provides information about the process used to recruit field testers, followed by a description of data collected as part of the field testing.

10.2.1.1 *Field testers*

A snowball approach was used to identify potential field testers in New Zealand and Australia. The researcher did not place any restrictions on evaluator expertise as the intention was that the conceptual framework should be accessible to new evaluators as well as those more experienced. In total 22 Australian and New Zealand evaluators were directly contacted via email with an invitation to participate as a field tester, and to also send the request on to their networks. This occurred in at least two cases. In New Zealand, a manager of a public sector evaluation team shared the invite at a meeting with cross-sector research and evaluation managers. In Australia, an evaluation lecturer at the University of Melbourne shared information about the field-testing opportunity with Master of Evaluation students.

Seven evaluators expressed an interest in field testing the conceptual framework and were taken through individual briefing sessions, each lasting about one hour. A benefit of this process was one of informed consent; evaluators had the opportunity to explore the framework and the study and consider the time commitment required before agreeing to participate as field testers. The briefing sessions were developed around a PowerPoint presentation that was delivered either face to face or online, depending on the location of the field testers. The session was split into three sections: (i) background to the study; (ii) information about using the conceptual framework to identify dimensions of merit and then

develop and justify a set of criteria; and (iii) information about field testing the handbook.

Following the briefing sessions, three evaluators declined to participate, citing work commitments. The remaining four agreed to field test the criteria matrix process. Three agreed that their case examples could be written up as part of the study. The fourth requested at the outset that their case example not be published, citing the political sensitivity of the program as the primary reason. This field tester was subsequently not able to fully field test the criteria matrix due to work commitments.

Information from all four field testers informs the review of the criteria matrix handbook, as documented in Chapter 9. In this respect the field testers had a dual role: (i) field testing the conceptual framework using an evaluation and (ii) based on their experience using the handbook as a resource, providing a review of the handbook to identify information that was helpful or confusing and to suggest changes, where relevant.

10.2.1.2 *Field testing*

In this part of the study – where the focus was on practical engagement with a theory-informed tool – the conceptual framework is referred to as a criteria matrix. This is also the term used in the handbook to describe the conceptual framework (Appendix F).

The three field testers were asked to engage with the criteria matrix to identify and develop criteria for an evaluation they were currently working on, using the criteria matrix handbook as a guide. How they engaged with the criteria matrix was up to them. It was anticipated that participation in the field testing could take each field tester four to five hours over the course of several weeks.

All three field testers agreed to their information being included in the study as case examples once identifying information had been removed. The case examples were written up in a standard format and sent back to each field tester to review, prior to being finalised for inclusion in the thesis.

The field testers were also asked to participate in a follow up interview. Prior to the interview field testers were asked to send the researcher a brief description of their evaluand, their completed matrix and set of criteria. This information

provided a starting point for the follow-up interview. The researcher was also able to follow up on information that was not clear from the documentation provided.

The interview comprised four open-ended questions. The first question required field testers to describe how they used the matrix, including whether it was used in a workshop setting or completed as a desk exercise. If others were involved in the process, the field testers were asked to describe the discussion that occurred around specifying the principles and interest groups, and the discussion that occurred around identifying dimensions of value. The intention of this question was to understand how evaluators engaged with the matrix, with a view to exploring similarities and differences across the three field testers.

A second question asked about how the field testers usually identified criteria and how, if at all, they attended to the task of justifying criteria. This question was intended to provide context for understanding how these field testers engaged with the criteria framework. The researcher's assumption was that practitioners who were not familiar with developing explicit criteria might struggle with the task more than those who routinely developed criteria for their evaluations.

Next, the field testers were asked a broad question about what added benefit, if any, the criteria matrix might have for program evaluation practice. The question was purposefully phrased in neutral language to give the field testers the opportunity to say the matrix did not add anything useful to their practice, if that was the case. The fourth, and final, question provided an opportunity for the field testers to add any additional thoughts. This question was simply stated as: 'Do you have any other comments to add?'

10.3 Field testers' case examples

This section presents the three case examples of criteria developed using the conceptual framework. Each case example includes information provided by the field tester and the researcher's descriptive analysis of the follow-up interview. The case examples are structured as follows:

- the evaluand context
- the field tester's usual approach to management of criteria

- how the field tester used the framework, including the completed normative framework and resulting criterion / criteria and sources, and how (or if) they attended to the task of justifying criteria
- the field tester's reflection on using a normative framework to develop criteria.

10.3.1 Case example A: Criteria for an evaluation of a disability action plan

10.3.1.1 *Context*

The evaluand is a local council's action plan for disability inclusion. The agency is a newly amalgamated council. Its Disability Inclusion Action Plan ('action plan') is an internal document that has been developed as a legislative requirement. The action plan is intended to set out the council's commitment to improve access to services, facilities and employment opportunities for residents, their families and carers living with disability and experiencing marginalisation. The plan provides strategies for positively shaping council practices over a four-year period. It focuses on four areas:

- developing positive community attitudes and behaviours
- creating liveable communities
- supporting access to meaningful employment
- improving access to mainstream services through better systems and processes.

The plan notes that these four areas were nominated by people with disability within the local government's jurisdiction as being of primary importance in creating an inclusive community.

10.3.1.2 *Usual management of criteria*

The concept of developing explicit criteria is not new for this practitioner. Prior to engaging with the normative framework, for example, she had already developed a set of criteria for the action plan (Table 31). Her process for doing so had involved holding discussions and working meetings with the client project team. They had agreed that the criteria should relate directly to the four focus areas outlined in the action plan. Thus, from the information gathered in these meetings with the client a set of four criteria had been developed.

Table 31: Criteria developed for a disability inclusion action plan

Criterion 1: Effectiveness of encouraging positive attitudes and behaviours

Dimension	
Staff capacity	Staff capacity utilised to identify, understand and deliver inclusive practices
Partnerships	Delivery of partnerships which promote and raise awareness of inclusive practices.
Community attitudes and behaviour	Increase in positive attitudes and behaviour within the general community towards people with disability.

Criterion 2: Effectiveness of the creation of liveable communities

Dimension	
Access to public places and facilities	Effective provision of good access to public places and facilities for people with disabilities.
Access to transport	Effective provision of good public transport access for people with disabilities.
Social inclusion and participation	Effective encouragement of people with disabilities to be involved in council planning processes.

Criterion 3: Effectiveness of the creation of meaningful employment

Dimension	
Workforce participation	Increase the participation of people with disability in the workforce in (location)
Good organisational capacity	Improve the council's organisational capacity to support a diverse workforce within the council.

Criterion 4: Effective systems and processes

Dimension	
Service delivery	Increase access to council's information and services for people with disability
Staff procedures and practices	Develop and implement staff procedures and practices that support the implementation of access and inclusion

10.3.1.3 *How was the criteria framework used?*

The evaluator developed a normative criteria matrix that included five interest groups: program staff, staff with disabilities, general community, community members with disability and their families, and council partners.

She then considered the normative principles and decided that the concepts of fairness and equity were too similar for this project, so she merged them into one principle called ‘equity’. She then completed the matrix as a desk-based exercise, working systematically through the matrix and populating the cells using information that had been discussed in project meetings and the dimensions that had already been developed into criteria. In this respect she used the criteria framework to check that the values of different interest groups had been considered with respect to four value lenses (maximising outcomes, duty, equity and ethic of care). The completed criteria matrix is included in Table 32: Dimensions of value for a disability inclusion action plan.

The evaluation practitioner justified the existing criteria through explicit reference to the information provided by the project team, and the action plan, and by noting that both had included input from people with disabilities. The criteria are further backed by legislation, specifically the Australian Disability Inclusion Act 2014.

Mathea Roorda #579498

Table 32: Dimensions of value for a disability inclusion action plan

	Maximising outcomes / maximum benefits	Duty / meeting obligations	Equity / being fair or equitable	Ethic of care / relationships and context
Council program staff	Full integration of the plan into the council's internal systems and processes.	Commitment by council staff and councillors to maintain, roll out and update the plan.	Equal opportunity for people with disabilities to participate in employment and community life.	Council encourages and supports people with disabilities to input into the plan.
Council staff with disabilities	Full and equal access for staff with disabilities to work at X Council.	Employees abide by employment contracts.	Equal right to participate in employment and community life.	Employees provide feedback to managers regarding new placements.
General community	Increased awareness and positive attitudes towards people with disability.		Community supports equal access provided for all local community opportunities for people with disabilities.	Supporting X Council to introduce new measures, policies and procedures for marginalised people.
Community members with disability and their families	Full and equal access to employment opportunities and community facilities and services for residents, their families and carers.			Families and carers of people with disabilities provide honest information to council about additional support if required.
Council partners	Partners work closely with the council to deliver effective services for people with disability and their families and carers.	Set up good working relationships and MOUs between the council and partners.		

10.3.1.4 *Reflecting on use of the criteria framework*

The evaluator viewed her engagement with the framework as a 'useful intellectual exercise'. Because of the nature of the evaluand and the work she had already done, she did not use the framework to identify dimensions of value. Rather, she said it provided a systematic way of thinking through the range of rights and responsibilities of all the key interest groups, which in turn provided a good way to cross check the existing criteria. She said it was useful to think about the range of perspectives from all interest groups rather than just the council. In summary, she considered it to be a useful exercise to cross check that the sub-dimensions were exhaustive. Further, she commented: 'it all made good sense – there was nothing that didn't fit neatly in (to the framework)'.

10.3.2 Case example B: Criteria for an evaluation of a curriculum framework

10.3.2.1 *Context*

For the past three years an international educational organisation has delivered a program for students in their middle years (typically Grades 6–10). Schools that deliver the program agree to become X[46] schools. The organisation provides schools with a curriculum framework and requires, through authorisation processes and other requirements, that schools deliver the framework. The evaluand is a revised curriculum for the program.

10.3.2.2 *Usual management of criteria*

This field tester is an experienced evaluation practitioner accustomed to using explicit criteria in her evaluations. Her usual approach in educational evaluations is to source criteria from empirical research, specifically focusing on what makes a good educational intervention, the stated outcomes, as documented and provided by the client, and the contract signed by the provider and schools.

10.3.2.3 *How was the criteria matrix used?*

She had planned to work through the criteria framework with the client. However, due to unforeseen circumstances she had to complete the matrix as a desk-based exercise. The framework included four interest groups: students, teachers, school

[46] 'X' is used here to protect the identity of the program

Mathea Roorda #579498

leaders and the program provider, and five principles: maximising outcomes, rights, fairness, duty and ethic of care, as illustrated in Table 33.

The field tester drew on the sources she would normally use, that is, empirical research on other educational programs relevant to the evaluand and the expected program outcomes. The 'ethic of care' principle prompted her to revisit findings from an earlier evaluation of the program in which it had been evident that students and teachers had experienced stress associated with participating in the program. This resulted in the development of a 'quality of life' criterion, as outlined in Table 34. The field tester thought she could reference research on wellbeing and education to further justify the inclusion of this criterion.

Table 33: Dimensions of value for a curriculum framework

	Maximising outcomes	Rights (minimum protection)	Fairness (being fair, just equitable)	Duty (meeting obligations)	Ethic of care (being virtuous)
Students	Strong/positive/enhanced academic outcomes in program Better prepared for diploma program. Become lifelong learners, active community members, globally minded Strong/positive/enhanced student wellbeing	Receive basic/adequate education	Equal opportunities for all kinds of students, i.e. the program can't be a curriculum that only some students do well at / succeed in	Students attend school	Students support other students to succeed in the program
Teachers	Curriculum planning and implementation is easier and more flexible for teachers. Gains for students are worth the effort.	Teachers are guaranteed that they receive sufficient training and resources (both from the program and from their school leaders) to be able to implement the curriculum Teachers receive sufficient leadership support to implement the curriculum	Curriculum needs to be 'implementable' by teachers at all different types of school, i.e. the program can't be a curriculum that only some schools (e.g. private, well-resourced schools) can deliver. If schools are permitted to become (X) schools, it needs to be something they can actually deliver	Deliver the curriculum as designed. Teachers have agreed to work at schools that deliver X, therefore they have a duty to deliver the curriculum	Need to ensure students are prepared for program / X education, i.e. that they have strong relationships with their teachers who look after and care for students' mental and emotional wellbeing as well as their academic progress. Need to make sure students are ok and have good quality of life while being a (X) student. Teachers support each other to succeed in program. There is a more collaborative / team environment among those who work at the school.

	Maximising outcomes	Rights (minimum protection)	Fairness (being fair, just equitable)	Duty (meeting obligations)	Ethic of care (being virtuous)
School leaders	Curriculum planning and implementation is easier and more flexible for coordinators and school leaders. Gains for students are worth the effort.		Cost and resourcing is not a barrier to implementing the program Curriculum is actually implementable at all schools permitted to become [X] schools	Structures and resources are in place so that teachers can implement program as designed.	Need to make sure teachers are prepared for and can have a solid and rewarding work/life balance. Need to have school leaders who are concerned for and care about their mental and emotional wellbeing as well as the quality of their teaching. There is a more collaborative / team environment among those who work at the school
Program owner	Improved teaching and learning at schools using the program			Duty to provide schools with adequate supports / training / resources and materials so they can reasonably understand and do what they are being asked to do (i.e. deliver the curriculum)	Organisation cares for the mental and emotional wellbeing of all staff and students at all schools using the curriculum. Students and schools need to be able to 'do the program' while also having a decent quality of life.

Table 34: Criteria for a curriculum framework evaluation

Criteria and sub-criteria	Description
Curriculum effectiveness. Defined as:	
Better for students	Next chapter (of the program) is 'better for students' in that X students experience stronger academic (preparedness for X) and stronger non-academic outcomes (lifelong learners, active community members, global mindedness).
Easier and more flexible for schools	Next chapter (of the program) is 'easier for teachers' in that curriculum planning and design is easier and more flexible. Teachers and school leaders form collaborative working relationships and deliver program as a whole school team.
Equity. Defined as:	
Equity	Next chapter curriculum is one that allows all students to succeed, regardless of background characteristics. Next chapter curriculum is one that all X schools are able to implement, regardless of school type or resourcing.
Implementation effectiveness. Defined as:	
Program support	Program materials provide sufficient information and practical support so that all types of teachers and schools know what they are expected to do as part of X: Next chapter.
School support	School leaders support teachers in program: Next chapter delivery. School leaders know the types of support they need to provide so that teachers can deliver X: Next chapter. Teachers receive sufficient training to be able to implement the curriculum.
Quality of life. Defined as:	
Relationships	Program interactions with schools about the program: Next chapter delivery are supportive and respectful. Teacher/school leader interactions with students under X: Next chapter are supportive and respectful. Teacher/school leader interactions with one another are supportive and respectful.
Balance	Design of the program: Next chapter curriculum respects student, teacher and school leaders' mental and emotional wellbeing. Students, teachers and school leaders say they can 'do (the program)' while also having a positive work/life or school/life balance.

10.3.2.4 *Reflecting on use of the criteria framework*

The field tester's reflection is that the criteria framework helped her to avoid 'just focusing on maximising outcomes'. Consideration of the other principles required explicitly thinking about what they might mean for this program.

> Without the tool I wouldn't have thought about ethical principles and in particular the quality of life criterion. It would not have entered into my thinking at all. Now I think it's a critical piece.

The focus on a principle of ethic of care prompted her to draw on a source that she might not otherwise have gone to; that is, findings from a previous evaluation. She then sought to further justify a criterion of 'quality of life' by returning to the empirical research and specifically the literature on wellbeing in schools and among students. In effect this additional source serves to further deepen(?) the inclusion of this criterion.

The field tester said she feels more confident about the criteria that have emerged from the process:

> It feels more systematic and it makes me feel like a better human, in that I'm not being driven by the client's thinking about what is good. I'm also required to think about fairness, duty and ethics of care.

Further, she noted that the information in the handbook enabled her to engage with quite complex concepts:

> There are people who have been thinking about what makes stuff good for centuries and centuries, and this was a clearly articulated way to summarise and get me to think through what is good in the context of this program, from all those perspectives.

10.3.3 Case example C: Criteria for an evaluation of an adolescent girls and young women program

10.3.3.1 *Context*

The evaluand is a program aimed at supporting adolescent girls and young women (AGYW). Girls and young women in the country where this program is located experience much higher rates of HIV infection than their male counterparts. Social isolation, economic disadvantage, discriminatory cultural norms, gender-based violence, school dropout and child marriage all contribute to AGYW's vulnerability to HIV. The practitioner's task is to identify criteria for the formative component of the evaluation. The key question guiding the evaluation is: 'How meritorious is the program?'

10.3.3.2 *How does the evaluator usually manage criteria?*

Criteria are a relatively new concept for this field tester. She first came across them as a student in a master's course she is currently doing on evaluation. She works in international development as an evaluator. For the current evaluation, she had intended to use the OECD/DAC criteria, and specifically the criterion of 'relevance' which is defined by the OECD/DAC as:

> The extent to which the aid activity is suited to the priorities and policies of the target group, recipient and donor. In evaluating the relevance of a program or a project, it is useful to consider the following questions: To what extent are the objectives of the program still valid? Are the activities and outputs of the program consistent with the overall goal and the attainment of its objectives? Are the activities and outputs of the program consistent with the intended impacts and effects?

In the absence of the criteria framework, she had intended to go to stakeholders and ask a general question: 'What are the most important things to judge the program on?'

10.3.3.3 *How was the criteria framework used?*

The organisation the field tester works for held a workshop for teachers, parents and program implementers, to introduce the program to the community. Although the program is aimed at adolescent girls and young women, this group of stakeholders was noticeably absent from the workshop.

Prior to the workshop the field tester had requested time to run a discussion session with participants. She planned to use the criteria matrix as her guide for asking a series of questions about the value of the program for different interest groups. The field tester is not of the same cultural group as those being targeted by the program. Thus, she had to rely on colleagues in the office to translate her questions into the local language and then back translate the information that was gathered during the discussion. Her five questions were oriented around each of the five principles in the criteria matrix:

- If the program achieves maximum outcomes, what would it look like for each (interest) group?

- Are there any duties or obligations that need to be made? If yes… if the program is to meet these obligations what would that be like (for each interest group)?
- Are there any rights that need to be upheld? If yes, for a good program, what are these rights? And for whom should they be upheld?
- Fairness for all – what would that look like (for each interest group)?
- What relationships need to be upheld?

Her reflection is that participants engaged primarily with the first principle of maximising outcomes, as illustrated in Table 35. This limitation may have been due to her inability to facilitate the session directly as she was reliant on other staff to communicate with participants.

The workshop discussion resulted in more information than the evaluation practitioner needed for the formative focus of the evaluation. For the purposes of her evaluation, she focused on information that pertained to a criterion of 'relevance' (see Table 36).

Table 35: Dimensions of value for an AGYW program

	Maximising outcomes	Fairness/rights	Duty	Ethic of care
AGYW (young women)	Tradition of early marriages (due to lack of financial support, parental coercion) is reduced AGYW have skills to engage in income generating activities (to be financially independent) AGYW have knowledge they need to enable them to avoid contracting HIV/AIDS AGYW have self-confidence to refrain from engaging in sexual activity to gain material things that their peers have Reduced incidence of HIV Reduced pregnancies Less sexual activity among young people AGYW develop leadership skills Increased number of girls in school and stay at school at higher levels Increased awareness of HIV/AIDS messages	Girls and boys treated the same (in work, not sure about home)	AGYW think independently AGYW resist peer pressure	AGYW value themselves AGYW empowered AGYW encouraged to have ambitions AGYW are a positive influencer of others
Boys		Boys have the opportunity to participate in the program.		
Implementers			Follow up on activities of AGYW (assist them in a timely way)	Create a conducive environment for learning

	Maximising outcomes	Fairness/rights	Duty	Ethic of care
			Implement the activities as promised Provide financial support for education	Create enabling environment for girls to report cases of abuse Ongoing engagement between teachers, NGO and other stakeholders Learner teacher friendly Transparency among all stakeholders
Community	Changed community perception of girls vs boys (achievers vs non-achievers) Abolish cultural practices that hinder education for girls Reduced incidence of gender-based violence Change stereotypes around gender and value Removal of unforeseen barriers that result in gender inequality	Leaders/chiefs back the project so that community members take it seriously Religious leaders do not officiate marriages of underage children	Engage AGYW and youth in the community development work Environment or platform that enables youth to reach out to them for support Chiefs uphold by-laws that penalise parents who marry off their children early Establish by-laws in education Community involvement in project activities Report cases of violence to the right authorities Introduce a variety of role models Improved environment to facilitate accessing HIV testing, accessing treatment and staying on treatment More extra curricula activities to engage youth	Community appreciates the project intentions Local leaders, implementers should all have access to the same information, at the same time.

	Maximising outcomes	Fairness/rights	Duty	Ethic of care
Parents	Abiding children to home rules		Allow daughters to participate in the activities Educate their children about HIV so they are sure they get the right information	

Table 36: Criterion of relevance for the AGYW program

Coverage	Program targets all AGYWs with specific interventions targeted to different age groups and interest groups like orphans, physically challenged, albinos, school dropouts, young mothers, and those living with HIV. Boys of the same age group 10–24 years should also be covered by the program.	
Empowerment	Program	Empowers AGYWs with knowledge and skills to negotiate right decisions; economic opportunities especially for school dropouts; and to enrol and stay in school especially for secondary education.
	Community	Chiefs, clergy, appreciate and effectively execute their role in protecting and supporting AGYWs
	Parents	Provide correct guidance to their children and report cases of abuse.
Environment	Program focuses on transforming community perceptions about gender values and harmful traditions like early marriage and gender-based violence. Promotes peer support networks that enable AGYWs to help each other and reach out to others as change agents. Promotes positive role models for youths (both AGYWs and males).	
Youth friendly health services	The program provides for provision of accessible, acceptable and integrated HIV and adolescent sexual and reproductive health services.	

10.3.3.4 *Reflection on use of the criteria framework*

The field tester began the interview by noting she is new to the concept of evaluation criteria and does not have much to compare the criteria framework to. That said, she believes the workshop yielded information that she would not have included if she had approached it from the OECD/DAC criteria. In particular, she commented that 'the principles helped to focus on relevance from a local context'.

10.4 Analysis of the case examples

This part of the field testing focused on evaluation practitioners' engagement with normative ethical perspectives and principles to develop criteria for evaluations they were currently working on. The field testers were provided with resources, including a criteria matrix handbook developed by the researcher and asked to engage with the information in a way that was appropriate for their evaluand and evaluation context.

At the outset it should be noted that analysis of this component of the study is based on the experiences and perspectives of three evaluation practitioners. The size of the sample makes it difficult to compare different evaluation practitioners' experiences with using the normative framework. That said, there are some

interesting observations that can be made about the use of the framework across these three cases. Within each case example there are also opportunities to examine how, and whether, the framework contributed to the development of criteria.

An initial observation is that the criteria framework was used for different purposes, to both retrospectively cross-check existing criteria (case example A) and to prospectively identify dimensions of value (case examples B and C). The field testers also engaged with the framework in two different ways: as a desk-based exercise (case examples A and B) and in a workshop setting with stakeholders (case example C). These differences indicate that evaluators can engage with normative perspectives in ways that make sense for a specific evaluand and evaluation context.

The findings indicate that principles drawn from the three normative perspectives (consequentialist, deontological and virtue based) can yield useful dimensions of value for different types of social programs in different contexts. In this study, the criteria matrix was used to identify criteria for evaluating a disability action plan, curriculum program and a sexual and reproductive health program in an international development setting.

Field testers were able to apply different principles to their evaluands to obtain dimensions of value. Two field testers (case example B and C) worked with the existing five principles as outlined in the handbook, while in case example A, the principles of fairness and equity were collapsed into one dimension of the criteria matrix.

One field tester (case B) initially struggled with the principle of ethic of care. However, after deeper consideration of its applicability to the evaluation, she became convinced that ethic of care was a critical dimension of quality for their evaluand. In this case, consideration of value from an ethic of care perspective resulted in a criterion that would not otherwise have been included. A second field tester (case C) was convinced the criteria matrix surfaced a broader range of value perspectives that would not have been identified through use of the OECD/DAC criteria. Thus, in two of the three cases, there was evidence that the criteria matrix achieved the aim of developing a more comprehensive set of criteria than would otherwise have occurred.

There were differences across the cases with field testers' management of justifying criteria. One field tester (case example A) had implicitly justified her criteria through reference to legislation and by noting that multiple stakeholders had provided input into the development of an action plan, even though this had been filtered through meetings with the client project team. A second field tester, working in the educational sector, normally sought justification of criteria through reference to empirical studies documented in peer-reviewed literature. The exercise made her more aware of the need to be explicit about justifying criteria and to also consider a broader range of sources, for example, previous evaluations of the program. A third field tester surfaced dimensions of value, developed a criterion but did not explicitly justify it. Had she done so, she might have argued that engagement with local community stakeholders had resulted in a much more nuanced criterion of relevance than might otherwise have been developed if they had relied on the OECD/DAC.

10.5 Chapter summary

While the testing of the criteria framework's applicability for program evaluation was on a modest scale, the findings indicate that consideration of value from diverse ethical perspectives and principles can be a useful exercise in surfacing dimensions of value or cross-checking existing criteria. Further, it has the potential to identify a broader range of perspectives on success or good than might otherwise occur. The field testing provides proof of concept for the criteria matrix framework, demonstrating it is possible to practically engage with normative ethical principles and combine both prescriptive and descriptive valuing traditions.

PART FOUR: DISCUSSION AND CONCLUSION

The final section of the study comprises Chapter 11, where the findings from all four components of the study are discussed, including implications for the field of evaluation and for practice. Limitations of the study are outlined and opportunities for future research identified. This is followed by the research conclusions.

11 Discussion

This chapter begins with a brief statement of the research problem and a broad overview of the study. The research findings are then discussed in the context of the literature. The contributions that this study makes to knowledge of evaluation are then discussed. Next, the strengths and limitations of the study are considered in terms of scope and methodology. The chapter ends with reflections, suggestions for future research and conclusions.

11.1 Research problem and overview of study

In the public sector context, evaluation is viewed as 'complementary, if not integral to program management' (McPhee 2006, p.2). In Australia and New Zealand, the expectation is that evaluators provide evaluative information that is robust and high quality. Evaluative reasoning – grounded in a logic for combining criteria with evidence to reach defensible evaluative judgements – has emerged in recent years as an important construct linked to evaluation quality (Nunns et al. 2015). But where evaluative reasoning is addressed by scholars it tends to be at a broad level, with limited focus on the individual elements required to reach a defensible evaluative conclusion. Of all the elements, the development of criteria is arguably the most fundamental (Fournier 1993) and also the least understood. Beyond a few resources, the identification and selection of criteria is a particularly underdeveloped area of evaluation practice.

This study has sought to address this gap in knowledge and in doing so, contribute to stretching and growing evaluation theory and practice (Chelimsky 2012). Its central concern has been to address the question: What is required to develop a defensible set of evaluation criteria for program evaluations? Informed by literature on valuing, evaluation criteria and normative ethical perspectives (chapter 2) the study was undertaken in two phases, with the second phase responding to what was learned in the first phase. In this respect, the study has exemplified research as a process which is 'emergent rather than tightly prefigured' (Creswell 2003, p.181).

11.2 Phase one: approach, summary and discussion of findings

In phase one of the study, the aim was to examine how criteria are managed in Australian and New Zealand program evaluation practice. Specifically, how do concepts such as program effectiveness, cultural responsiveness, accessibility and efficiency get their meaning? What sources do evaluators draw on, how are criteria portrayed in evaluation reports, and what strategies are used to justify them? The development of criteria was examined through a survey of 137 experienced evaluation practitioners (discussed in Chapter 5) and a document review of 47 published evaluation reports (discussed in Chapter 6). To the researcher's knowledge, this study is the first to provide an in-depth systematic examination of criteria development in program evaluation. Its contribution to the field of evaluation cannot be underestimated. Empirical studies such as this 'provide powerful insight and offer evidence from which prescriptive theoretical ideas can be developed and enhanced' (Christie 2012, p.2).

Drawing on existing literature on development of criteria and the findings from phase one, three characteristics were identified as important to the development of defensible criteria. In the context of this study, the term 'characteristic' is used to refer to the elements that are required for criteria to be considered defensible.

The first characteristic of defensible criteria is authoritative sources. A source is authoritative when it has a defensible warrant for its inclusion in the definition of good for the evaluand in its context. Can the evaluator provide defensible reasons to support the inclusion of certain sources? Criteria should not be justified through sole reliance on program goals, as these may be politically motivated or based on a compromise of different interests (Scriven 1972; 1982). Authoritative sources may be based on evidence, such as a needs assessment (Scriven 1972; 1986), knowledge in a subject area (Eisner 1976; 2004) or cultural expertise (Goodwin, Sauni, and Were 2015; Wehipeihana and McKegg 2018).

The second characteristic of defensible criteria is comprehensiveness. Do the criteria include all the important dimensions of value for an evaluand (House and Howe 2000)? Do they include the perspectives of those most disadvantaged in society and not just those of the client (Greene and Caracelli 1997; House and Howe 2000)? To answer these questions, the evaluator must be aware of the list of important and relevant values (Scriven 1990, 1995; Stake and Schwandt 2006;

Wehipeihana and McKegg 2018). They may approach this task through descriptive valuing (Shadish et al. 1991), that is ensuring the perspectives of all interest groups with a legitimate interest in a program are included. Such a stance is supported by Standard U2 of the Program Evaluation Standards (Yarbrough et al. 2011). They may also approach the task through prescriptive valuing (Shadish et al. 1991), that is, ensuring all normative perspectives, or ways of conceptualising 'good', 'success' and so forth, are considered.

A third characteristic, full description, is necessary but not sufficient to justify evaluation criteria. An evaluator is in a better position to defend the comprehensiveness and authoritativeness of criteria if abstract value terms are explicitly defined or described (Sadler 1985). However, criteria may be explicit but not justified, as demonstrated in the review of evaluation reports undertaken as part of this study (Chapter 6).

In summary, the three characteristics – authoritative sources, comprehensiveness and full description – help to build an argument in support of defensible criteria, which in turn serve to support defensible evaluative reasoning. The first two are closely connected to the concept of justification, while the third has a role in communicating criteria so that their defensibility can be explicitly defended.

In the second part of phase one, survey research provided empirical evidence that a critical element of evaluative reasoning is weak in Australian and New Zealand program evaluations. The findings from phase one identified a misalignment between theory about evaluative reasoning – what scholars think ought to occur – and practice, what evaluators say they do and what is observed in program evaluation reports.

Making evaluative claims is not all that evaluators do, but it is what sets evaluation apart from other types of inquiry (Davidson 2014a, 2005; Schwandt 2015, 2018; Scriven 2013a). Evaluative reasoning is grounded in a procedural logic (Fournier 1993) that requires a clear link between evaluative judgement, evidence, criteria and appropriate warrants (Arens 2005; Superu and Aotearoa New Zealand Evaluation Association 2015). Evaluative reasoning is the crux of evaluation quality (Nunns et al. 2015). Logically, a failure to justify the value premises against which judgements are made weakens the evaluative reasoning link, and risks evaluations being dismissed as based on personal opinions or preferences. To support

defensible evaluative reasoning, criteria must be based on authoritative sources and incorporate all relevant perspectives (that is, be comprehensive).

The findings from phase one demonstrate that evaluation practice in Australia and New Zealand exemplifies a variety of concerns expressed in evaluation literature. These include Julnes's (2012, p.4) comment that 'evaluators have often been unreflective, and even sloppy, in their approaches to valuing'; earlier research by Nunns et al. (2015); and concerns by Greene (2011) and Schwandt (2008) that evaluative reasoning is neglected by evaluators.

11.3 Phase two: approach, summary and discussion of findings

The findings from phase one led to a new research question being incorporated in the study: What is an evidence-informed conceptual framework for defensible criteria that can be operationalised for evaluation practice in the public sector? The rationale for a theory-to-practice focus is as follows. Theoretical perspectives exist on identifying and managing criteria (Alkin et al. 2011; Davidson 2005; Stake and Schwandt 2006) but there exists a chasm between what different scholars advocate as appropriate valuing strategies and what occurs in practice. While the term 'systematic' differentiates informal from professional evaluation (Stufflebeam and Shinkfield 2007) there is little evidence of deliberate theory-informed strategies being employed in program evaluations to identify and manage the values against which judgements are made.

Phase two explored how engagement with normative perspectives might support the development of a comprehensive set of criteria. A conceptual framework was developed, that drew on both prescriptive and descriptive traditions of valuing (discussed in Chapter 8). The conceptual framework makes several novel and significant contributions to the field of evaluation. It draws on a prescriptive valuing approach, providing a way for practitioners to engage with normative ethical perspectives which deal with conceptions of 'good', 'bad' and so forth. In doing so, it directly responds to a call for evaluation to be more ethically-oriented (Scriven 2016; Stame 2018). It also draws on a descriptive valuing approach, requiring the evaluator to identify and include the perspectives of all groups with a legitimate interest in a program, as recommended by the Program Evaluation Standards (Yarbrough et al. 2011).

With the framework conceptualised as a two-dimensional matrix, the evaluator is required to systematically consider the value perspectives of each interest group, through the lens of five ethical principles derived from three normative perspectives – consequentialist, deontological and virtue-based. Its aim is to support comprehensiveness, that is, to systematically identify all relevant dimensions of value for an evaluand and evaluation. The framework responds directly to those who have puzzled about how an evaluator would be able to ascertain whether a list of criteria is complete (Fournier 1993). Use of the framework does not preclude additional criteria emerging as more becomes known about the evaluand.

The framework endorses a pluralist approach to valuing (Greene 2001), both in terms of considering different normative perspectives, and in seeking an understanding of value that incorporates but is broader than that espoused by end clients or policy makers (House and Howe 1999). It provides a practical way to systematically ensure that criteria include the value perspectives of all interest groups including those that typically have no voice, as in the case of those most disadvantaged in society, animals or the environment.

The link between normative perspectives and conceptions of value is not new (House and Howe 2000; Shadish et al. 1991) but does not readily translate to evaluation practice. As demonstrated in this study, evaluation practitioners have limited knowledge of values. Concepts such as 'value for money' and 'efficiency' are part of mainstream managerial parlance, but unless practitioners have a background in philosophy, they do not necessarily make the connection between these commonly used terms and normative ethics. Yet normative ethics is integral to values theory and values are at the centre of evaluation.

With this in mind, a practice-based criteria matrix framework handbook was developed, reviewed and field tested, with a view to translating the conceptual framework into a tool for practice and thus bridge the gap between theory and practice. The handbook was reviewed by drawing on the Evaluation Checklists Project validation process (discussed in Chapter 9). Its applicability to program evaluation practice was then tested (and discussed in Chapter10).

Field-testing of a criteria matrix framework provides proof of concept for an evaluation-specific tool that can support practitioners to engage with normative

ethical perspectives. As such, the study disproves Shadish et al's. (1991) assertion that prescriptive valuing is too difficult to carry out in practice. Findings from the initial field-testing indicate that the framework can assist practitioners to identify relevant criteria they might otherwise have ignored, thus ensuring that evaluative judgements are based on a more comprehensive set of criteria . It can be used in different types of evaluations as both a prospective tool for identifying criteria, or a retrospective tool, for checking existing criteria.

As a prospective tool, the criteria matrix framework is unique in that it supports practitioners to identify dimensions of value using a transparent, grounded approach. Rather than starting an evaluation with a predetermined evaluation question, dimensions of value can be surfaced through consideration of each interest group, through the lens of different normative ethical perspectives. In such an approach, criteria focus the evaluative inquiry on what is truly important, informing both the development of evaluation questions and the way in which evaluative terms embedded in those questions are defined.

11.4 Contributions to knowledge

As noted in Chapter 2, the literature about evaluative reasoning is unequivocal about the role of criteria in reaching an evaluative conclusion. Criteria provide the value premises on which the evaluator reasons towards an evaluative judgement (ANZEA and SUPERU 2014; Martens 2018; Nunns et al. 2015; Yarbrough et al. 2011). Examining the literature on valuing in evaluation, a reader may be bewildered by the debates about how criteria ought to best be identified – either explicitly or implicitly, prescriptively or descriptively. Given the importance of criteria to evaluation, it is striking that these diverse views about criteria are not backed by empirical research. The notable exceptions are three studies that examine criteria within a broader focus on evaluative reasoning (Hurteau et al. 2009 and Nunns et al. 2015) and the logic of evaluation (Ozeki et al. 2019). This study adds to the body of empirical research on valuing with an in-depth examination of criteria in Australian and New Zealand program evaluation. The findings indicate that explicit criteria are not routinely included in evaluation practice. Further, the failure to fully describe, provide reference to authoritative sources, and ensure all relevant value perspectives are included, serves to weaken the evaluator's ability to reason to a defensible evaluative judgement. Understanding how evaluators currently develop criteria is important as it

provides a basis from which to consider areas where knowledge about this aspect of evaluation needs to be developed or further refined.

In addition to contributing empirical studies to the knowledge base, this study also contributes to theory on evaluative reasoning in practice by drawing on the findings of how criteria are developed in current program evaluation practice, and literature on values and valuing. Scholars have tended to focus on criteria at a meta-level, for example, Fournier (1993) described criteria as the first step in the logic of evaluation, and Nunns (2017) as one of the elements that constitute evaluative reasoning. This study has attempted to explicate the next level of detail of evaluative reasoning by identifying three characteristics that together serve to strengthen the value premises of an evaluative argument. In the language of argumentation, authoritative sources provide the evidence required to make a claim about the criteria against which an evaluative judgment is made; attendance to all relevant dimensions of value provides a type of methodological warrant; and full description serves to make the reasoning process explicit.

The conceptual framework developed as part of this study is an additional contribution to the knowledge base. The framework draws on expansive literature on value theory and the long tradition of stakeholder inclusion in evaluation (Greene et al. 2006) to deal head-on with the issue of narrowly-defined criteria that is noted in the literature and also observed in both empirical strands of this study. Whereas evaluation scholars tend to focus on one or other normative perspective or ignore reference to them altogether, the framework considers each perspective as providing a useful lens through which to consider what is important, for each group with a vested interest in a program. Systematic consideration of all relevant dimensions of value to inform criteria is consistent with democratic traditions in evaluation. It is hoped that evaluators explore the applicability of the framework in different contexts and for different purposes, that they assess its strengths and weaknesses, and share their experiences so that the conceptual framework can be improved or adapted.

The study took the conceptual framework one step further and developed it into a tool for practice. To support the transition, a handbook was developed and reviewed to support evaluation practitioners to engage with the framework. The handbook provides step-by-step instructions for identifying dimensions of value based on normative perspectives and then using these to develop a justified set of

criteria. As such, the handbook facilitates the process of identifying and developing criteria and contributes a new evaluation specific tool to evaluation practice.

Initial field testing has provided proof of concept that the tool can support evaluators to identify a comprehensive set of criteria against which to assess an evaluand. The three characteristics of defensible criteria are incorporated into the tool, and the tool itself supports the development of a warranted argument to make a defensible claim about criteria. Initial field testing of the tool indicates it can be used in different evaluand contexts and can be used to both retrospectively check existing criteria or to develop criteria at the design stage of an evaluation. However, further research is required to understand its applicability within different evaluation approaches, and in different program settings. This will no doubt result in improvements and potential hybrid approaches for identifying criteria of merit.

11.5 Limitations

This study includes several methodological limitations that need to be considered when interpreting the results.

11.5.1 Online survey

The list of sources included in Question 7 of the online survey were reviewed with input from experienced evaluators and others. It was therefore anticipated there would be a high degree of agreement about the meaning of the items included. However, items such as cultural expertise and needs assessment, for example, can be understood in different ways by evaluators in and across New Zealand and Australia. The implication is that respondents have rated items from their own perspective, and unless they included a comment, their understanding and therefore use and views about the importance of the source are not known. This is an acknowledged limitation of the online survey.

A second limitation concerns the inability to calculate the response rate for the online survey. The survey had 137 responses from evaluators who met the selection criteria and completed all the compulsory questions. Neither the AES nor ANZEA could provide information on the number of their members who match the same selection criteria as the survey. In addition, some New Zealand evaluators belong to both professional organisations, although it is not known how many. For these reasons it has not been possible to provide a response rate for the survey.

What is known is that those who responded to the survey tended to be experienced evaluators, and therefore knowledgeable about how evaluations are done in the public sectors within which they work. There were strong similarities between the responses of evaluators working in the Australian public sector space, and those working in New Zealand. This overall finding provides validity to the study.

In addition, the study examined the topic of criteria in current program evaluation practice from two perspective. Together, the findings from the survey and the document review provide a deeper and richer picture of the research problem.

11.5.2 Document review

The researcher's analysis and interpretation of the data on criteria is grounded in the information that is presented in the 47 reports included in the sample. It is possible that other sources informed criteria development that have not been documented in the reports and were therefore not captured in the analysis process. Nonetheless, it is important to note that the reports are a public document and presumably include all the information the evaluator deems pertinent to support their evaluative claims.

Another consideration is the amount and level of detail the researcher elected to code and analyse. The researcher's context, as a Pākehā evaluator, meant she was alert to criteria on cultural relevance and appropriateness in evaluations where indigenous peoples were a primary target group, and thus was relatively confident about identifying criteria as relevant to specific reports. In other evaluations where the researcher had less content knowledge, it was difficult to identify if criteria were absent. This limitation informed what sub-analyses were conducted. Other researchers, analysing the same set of reports, might identify additional gaps in criteria.

11.5.3 Criteria matrix

There are acknowledged limitations associated with the normative framework as currently conceptualised. First, it does not deal with prioritisation of values. Even if criteria are justified, there are likely to be contexts where there are competing value perspectives. Which conception of value takes priority? In such circumstances, the evaluator may need to undertake a separate prioritisation exercise. Second, the normative principles come out of western philosophical

thinking. These principles may not resonate with te au Māori, Aboriginal and Torres Strait Islander other worldviews. Both these limitations can also be viewed as opportunities for future research.

11.6 Final reflections and opportunities for future research

In this section the researcher returns to the central research question to consider: What is now known about developing defensible criteria for program evaluations, and where are there still gaps in knowledge?

In evaluative reasoning, the evaluator's aim is defensibility, not conclusiveness. Their role is to set out a convincing evaluative argument, one that is credible, plausible and probable (House 1977). In Scriven et al's (2012) depiction of evaluative reasoning, evidence (what data to look at) is combined with a claim about value (which determines how the evidence will be interpreted) to reach an evaluative judgment. Their depiction of evaluative reasoning is reproduced in Figure 20. The central argument of this thesis is that each element of evaluative reasoning also requires its own reasoning process. Thus, the dotted line in Figure 20 depicts a warranted argument about criteria. Specifically, evaluators need to defend their claims about value premises by providing reasons that are based on authoritative information or evidence. These reasons may need to be supported by appropriate warrants that link the evidence to the claim through appeal to some authority that is context dependent.

The study has identified authoritative sources as an important characteristic of defensible criteria, and thus essential to building an argument about value. As illustrated in Figure 20, authoritative sources provide the evidence required to make a claim about what matters. In practice, whatever the evaluation approach, and whatever tool or approach is used to develop criteria, evaluators should question the authority of sources, that is, their credibility, integrity or reliability in the context of a specific evaluand and evaluation.

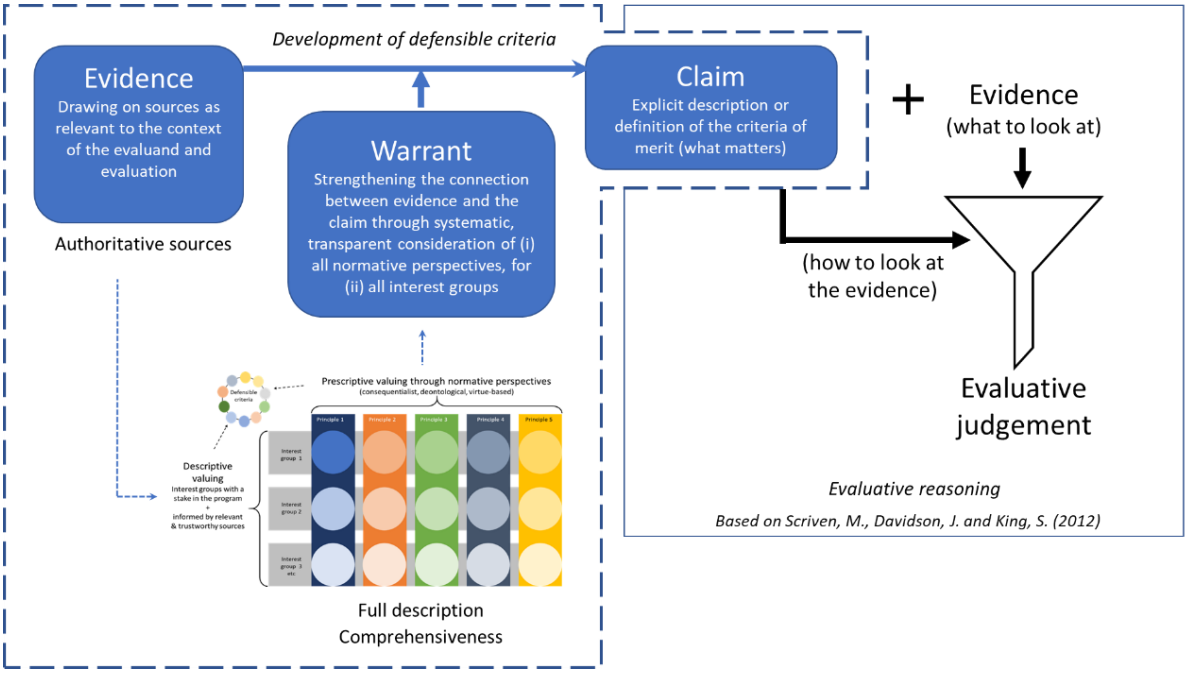


Figure 20: Revised conceptual framework of defensible criteria

The conceptual framework relies on authoritative sources to identify dimensions of value for each interest group. However, the study findings do not provide guidance about how to assess authoritativeness. Nor does the study address the issue of who should set the criteria. Should it be the evaluator or the client? Further research may provide a clearer understanding of what is required of authoritative sources in different program evaluation contexts.

Figure 20 locates the role of the criteria matrix framework in developing a defensible set of criteria. Specifically, the framework supports criteria to be fully described and comprehensive. In doing so, the criteria matrix framework has a warrant-making role, bridging evidence to the claim about criteria of merit. It supports the warrant component of an argument in two ways. First, systematic consideration of all relevant dimensions of value provides a type of methodological warrant (Booth, Colomb, and Williams 2003) as the framework provides transparent assurance that all relevant perspectives of value have been considered. The process of systematically considering dimensions of value (a) for all legitimate stakeholders and (b) for all relevant ethical principles guards against selectivity or value bias. It ensures that value is not defined by one interest group or through the lens of one normative perspective. Second, the tool provides an ethical warrant as

it deals directly with conceptions of value (what is 'good' or 'bad') that have been worked out over centuries of philosophical thinking (Shadish et al. 1991).

While the framework developed as part of this study is underpinned by normative ethical perspectives, it is but one framework. It is anticipated that further research will identify additional evaluation specific approaches that can support this element of evaluative reasoning. As such, the conceptual framework should be considered a 'stake in the ground' from which further research may identify alternative approaches to developing defensible criteria for program evaluations. In addition, the current framework will benefit from ongoing improvement as it is applied in different contexts and for different purposes.

The current study makes an important and significant contribution to the knowledge base, but there is still much more to explore with regard to the development of criteria. Several topics for future research have emerged from the present study, including areas where additional research would help refine or improve the criteria matrix. First, replications of the present study with a larger sample of field testers would help refine the tool and understand its applicability in practice. Second, research in different non-western cultural contexts could explore the applicability of the normative ethical principles included in the current tool or identify additional perspectives or principles as relevant to local cultural contexts. Third, replications of the field testing could be undertaken with evaluations of programs beyond the social sector. Fourth, there are opportunities to replicate the study in other countries beyond Australia and New Zealand.

Future research could also attend to the acknowledged limitations associated with the framework as currently conceptualised. Specifically, further theoretical discussion about prioritisation of different value perspectives is necessary. In addition, who should be involved in prioritising criteria, and how might that be applied to the criteria matrix?

An additional area of future research is to examine the information required to engage with the criteria matrix. For example, what value theory should be included in evaluation curricula, and what level of knowledge is required to engage with each of the normative ethical perspectives?

Finally, in phase one of the study, survey respondents identified needs assessment as an important or very important source for identifying criteria, yet not always

included in practice. Future research could usefully identify the reasons for this gap. Likewise, research could explore effective strategies for including cultural expertise as a source for identifying criteria in Australian and New Zealand program evaluations.

11.7 Conclusion

Defensible criteria are critical to defensible evaluative reasoning but are an under-researched topic. The study contributes to the field of evaluation by explicating three characteristics of defensible criteria; examining how criteria are managed in current Australian and New Zealand program evaluation practice, by developing a conceptual framework, an evaluation specific tool and supporting handbook for identifying and developing defensible criteria for program evaluations. As such, the study provides theory-informed practical guidance for evaluators committed to undertaking quality evaluations.

The study helps enhance the practical applicability of normative ethics to program evaluation. Several scholars have espoused the importance of normative ethics to the task of identifying criteria or dimensions of value but methods for engaging practitioners with these concepts have been absent in the literature. The present study begins the process of bringing values theory directly to the work of program evaluators.

The final comments in this thesis are addressed to those who commission or conduct program evaluations and who are justifiably concerned with the development of quality evidence to support effective decision-making. The first – obvious but important – point to note is that evidence is never value-free. Values exist in every scientific endeavour. But in evaluation, values are fundamental as they define how quality or goodness is to be understood in a specific context. In evaluation we refer to these values as criteria.

A second, related, point is that professional evaluators are also committed to developing quality evaluative information. That requires diligent attention to each element of the evaluative reasoning chain. With comprehensive, reasoned criteria, the evaluator is in a stronger position to deliver evaluative judgements to support defensible policy and program decisions.

References

ACT Government. 2010. “Evaluation Policy and Guidelines.”

Alkin, M. C., A. T. Vo, and C. A. Christie. 2012. “The Evaluator’s Role in Valuing: Who and With Whom.” *New Directions for Evaluation* 133:29–41.

Altschuld, J. W. and D. D. Kumar. 2005. “Needs Assessment.” Pp. 276–77 in *Encyclopedia of Evaluation*, edited by S. Mathison. Thousand Oaks, CA: SAGE Publications.

Anderson, C. W. 1979. “The Place of Principles in Policy Analysis.” *American Political Science Review* 73:711–23.

ANZEA and SUPERU. 2014. “Proposed Evaluation Standards for Aotearoa New Zealand: Exposure Draft.” (September).

Aotearoa New Zealand Evaluation Association. 2011. “Evaluator Competencies.”

Aotearoa New Zealand Evaluation Association. n.d. “Our History.” Retrieved April 6, 2018 (https://www.anzea.org.nz/what-we-do/).

Arens, S. 2005. “A Study of Evaluative Reasoning in Evaluative Studies Judged ‘Outstanding.’” Indiana University.

Australasian Evaluation Society. 2013. “Guidelines for the Ethical Conduct of Evaluations.”

Australian Evaluation Society. n.d. “Talking about Evaluation.” Retrieved November 30, 2019 (https://www.aes.asn.au/resources/talking-about-evaluation-1.html).

Bamberger, M., J. Rugh, and L. Mabry. 2006. *RealWorld Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Beyer, B. K. 1985. “Critical Thinking: What Is It.” *Social Education* 49(4):270–76.

Booth, W. C., G. G. Colomb, and J. M. Williams. 2003. *The Craft of Research*. 2nd ed. Chicago: University of Chicago Press.

Boston, J., A. Bradstock, and D. Eng. 2010. “Ethics and Public Policy.” Pp. 1–17 in *Public policy: Why ethics matters*, edited by J. Boston, A. Bradstock, and D. Eng.

Canberra: ANU E Press.

Braham, M. and M. van Hees. 2014. "Ordering the Right and the Good, Practically Speaking." *Journal of Value Inquiry* 48(1):95–105.

Brandon, P. R. 1998. "Stakeholder Participation for the Purpose of Helping Ensure Evaluation Validity: Bridging the Gap Between Collaborative and Non-Collaborative Evaluations." *American Journal of Evaluation* 19(3):325–37.

Brandon, P. R., N. L. Smith, C. Trenholm, and B. Devaney. 2010. "Evaluation Exemplar: The Critical Importance of Stakeholder Relations in a National, Experimental Abstinence Education Evaluation." *American Journal of Evaluation* 31(4):517–31.

Bromell, D. 2012. *Doing the Right Thing: Ethical Dilemmas in Public Policy Making*. Dunedin.

Cavanagh, G., D. Moberg, and M. Velasquez. 1981. "The Ethics of Organizational Politics." *Academy of Management* 6(3):363–74.

Chelimsky, E. 1987. "The Politics of Program Evaluation." *Society* Nov-Dec.

Chelimsky, E. 1995. "The Political Environment of Evaluation and What It Means for the Development of the Field: Evaluation for a New Century: A Global Perspective." *American Journal of Evaluation* 16(3):215–25.

Chelimsky, E. 1998. "The Role of Experience in Formulating Theories of Evaluation Practice." *American Journal of Evaluation* 19(1):35–55.

Chelimsky, E. 2012. "Balancing Theory and Practice in the Real World." in *Eastern Evaluation Research Society's Meeting*.

Childress, J. F. 2009. "Methods in Bioethics." *The Oxford Handbook of Bioethics* (April 2018). Retrieved April 10, 2018 (www.oxfordhandbooks.com).

Chouinard, J. A. and P. Milley. 2015. "From New Public Management to New Political Governance: Implications for Evaluation." *Canadian Journal of Program Evaluation* 30(1):1–22.

Christie, C. A. 2012. "Advancing Empirical Scholarship to Further Develop Evaluation Theory and Practice." *Canadian Journal of Program Evaluation*

26(1):1–18.

Christie, C. A. and M. Rose. 2003. “Learning About Evaluation Through Dialogue: Lessons From an Informal Discussion Group.” *American Journal of Evaluation* 24:235–43.

Comrey, A. L. and H. B. Lee. 1992. *A First Course in Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Cooksy, L. J. and V. J. Caracelli. 2009. “Metaevaluation in Practice: Selection and Application of Criteria.” *Journal of MultiDisciplinary Evaluation* 6(11):1–15.

Corbin, J. and A. Strauss. 2008. *Basics of Qualitative Research*. 3rd ed. Thousand Oaks, CA: SAGE Publications.

Coryn, C. L. et al. 2017. “A Decade of Research on Evaluation: A Systematic Review of Research on Evaluation Published between 2005 and 2014.” *American Journal of Evaluation* 38(3):329–47.

Crane, J. A. 1988. “Evaluation as Scientific Research.” *Evaluation Review* 12(5):467–82.

Crane, J. A. 1989. “The Problem of Valuation in Risk-Cost-Benefit Assessment of Public Policies.” Pp. 67–79 in *Technological Transformation: Contextual and Conceptual Implications*, edited by E. F. Byrne and J. C. Pitt. Kluwer Academic Publishers.

Creswell, J. W. 1994. *Research Design: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage Publications, Inc.

Creswell, J. W. 2003. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 2nd ed. Thousand Oaks, CA: Sage.

Creswell, J. W. 2015. *A Concise Introduction to Mixed Methods Research*. London: Sage Publications, Inc.

Creswell, J. W. and V. L. Plano Clark. 2007. *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage Publications, Inc.

Crotty, M. 1998. *The Foundations of Social Research: Meaning and Perspective in the Research Process*. Crows Nest, Australia: Allen & Unwin.

Dahler-Larsen, P. 2001. "From Programme Theory to Constructivism: On Tragic, Magic and Competing Programmes." *Evaluation* 7(3):331–49.

Dalkey, N. and O. Helmer. 1963. "An Experimental Application of the Delphi Method to the Use of Experts." *Management Science* 9(3):458–67.

Davidson, E. J. 2010. "Process Values and Deep Values in Evaluation." *Journal of Multidisciplinary Evaluation* 6(13):206–8.

Davidson, E. J. 2013. *Actionable Evaluation Basics: Getting Succinct Answers to the Most Important Questions*. Auckland, New Zealand: Real Evaluation.

Davidson, E. J. 2014a. *Evaluative Reasoning*. Florence: UNICEF Office of Research.

Davidson, E. J. 2014b. "How 'Beauty' Can Bring Truth and Justice to Life." Pp. 31–43 in *Revisiting Truth, Beauty, and Justice: Evaluating With Validity in the 21st Century*. Vol. 142, edited by J. C. Griffith and B. Montrose-Moorhead. New Directions for Evaluation.

Davidson, J. E. 2005. *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Department for Communities and Local Government. 2009. *Multi-Criteria Analysis: A Manual*. London: Department for Communities and Local Government.

Department of Finance. 1994. "Doing Evaluations: A Practical Guide."

Department of Finance and Administration. 2000. "The Outcomes & Outputs Framework: Guidance Document." (November):1–52. Retrieved April 12, 2011 (https://www.finance.gov.tt/wp-content/uploads/2013/11/pubA7AACE.pdf).

Department of Finance and Deregulation, Department of FInance and Deregulation. 2010. "Strategic Review of Indigenous Expenditure." Retrieved March 14, 2014 (https://www.finance.gov.au/sites/default/files/foi_10-27_strategic_review_indigenous_expenditure.pdf).

Department of Prime Minister and Cabinet. 2018. "Indigenous Advancement Strategy Evaluation Framework." Retrieved September 15, 2019 (https://www.niaa.gov.au/sites/default/files/publications/ias-evaluation-framework.pdf).

Eisner, E. W. 1976. “Educational Connoisseurship and Criticism: Their Form and Functions in Educational Evaluation.” *Journal of Aesthetic Education* 10(3/4):135–50.

Eisner, E. W. 2004. “The Roots of Connoisseurship and Criticism: A Personal Journey.” Pp. 196–202 in *Evaluation roots: Tracing theorists’ views and influences*, edited by M. C. Alkin. Thousand Oaks, CA: Sage.

Empowered Communities. 2015. “Empowered Communities: Empowered Peoples - Design Report.” Retrieved July 10, 2016 (https://empoweredcommunities.org.au/wp-content/uploads/2018/04/EC-Report.pdf).

European Commission. 2004. *Evaluating EU Activities: A Practical Guide for the Commission Services*. Belgium.

Fitzpatrick, J. L., J. R. Sanders, and B. R. Worthen. 2004. *Program Evaluation: Alternative Approaches and Practical Guidelines*. 3rd ed. Boston: Allyn and Bacon.

Fitzpatrick, J. L., J. R. Sanders, and B. R. Worthen. 2011. *Program Evaluation: Alternative Approaches and Practical Guidelines*. 4th ed. Upper Saddle River, N.J.: Pearson Education.

Fournier, D. M. 1993. “Reasoning In Evaluation: A Distinction Between General and Working Logic.” Syracuse University.

Fournier, D. M. 1995a. “Editor’s Notes.” *New Directions for Evaluation* 68(Winter):1–4.

Fournier, D. M. 1995b. “Establishing Evaluative Conclusions: A Distinction between General and Working Logic.” *New Directions for Evaluation* 68(Winter):15–32.

Gephart, W. J. 1981. “Values, Valuing, And Evaluation.”

Glaser, B. G. and A. L. Strauss. 1967. *The Discovery of Grounded Theory*. Chicago: Aldine.

Gluckman, P. 2011. *Towards Better Use of Evidence in Policy Formation: A Discussion Paper*. Auckland.

Gluckman, P. 2012. "Speech to Transit of Venus Forum: An Overview and Introduction." 1–2.

Gluckman, P. 2013a. "Speech to the Institute of Public Administration New Zealand (IPANZ) 'Communicating and Using Evidence in Policy Formation: The Use and Misuse of Science.'" 1–10.

Gluckman, P. 2013b. "The Role of Evidence in Policy Formation and Implementation." (September):1–39. Retrieved February 10, 2014 (http://www.pmcsa.org.nz/wp-content/uploads/The-role-of-evidence-in-policy-formation-and-implementation-report.pdf).

Goodwin, D., P. Sauni, and L. Were. 2015. "Cultural Fit: An Important Criterion for Effective Interventions and Evaluation Work." *Evaluation Matters—He Take Tō Te Aromatawai* 1:25–46.

Greene, J. C. 1997. "Evaluation as Advocacy." *American Journal of Evaluation* 18(1):25–35.

Greene, J. C. 2001. "Dialogue in Evaluation: A Relational Perspective." *Evaluation* 7(2):181–87.

Greene, J. C. 2002. "With a Splash of Soda, Please: Towards Active Engagement with Difference." *Evaluation* 8(2):259–66.

Greene, J. C. 2005. "A Value-Engaged Approach for Evaluating the Bunche-Da Vinci Learning Academy." *New Directions for Evaluation* 106(Summer):27–45.

Greene, J. C. 2009. "Summary of the Evaluation." Pp. 47–48 in *Evaluation in Action: Interviews with Expert Evaluators*, edited by J. Fitzpatrick, C. Christie, and M. M. Mark. Thousand Oaks, CA: Sage Publications, Inc.

Greene, J. C. 2011. "The Construct(Ion) of Validity as Argument." Pp. 81–91 in *Advancing Validity in Outcome Evaluation: Theory and Practice*. Vol. 130, edited by H. T. Chen, S. I. Donaldson, and M. M. Mark. New Directions for Evaluation.

Greene, J. C. and V. J. Caracelli. 1997. "Defining and Describing the Paradigm Issue in Mixed-Method Evaluation." Pp. 5–17 in *Advances in mixed-method evaluation: The Challenges and Benefits of Integrating Diverse Paradigms*. Vol. 74. San

Francisco: New Directions for Evaluation.

Greene, J. C., V. J. Caracelli, and W. F. Graham. 2008. "Toward a Conceptual Framework for Mixed-Method Evaluation Design." Pp. 121–48 in *The Mixed Methods Reader*, edited by V. L. Plano Clark and J. W. Creswell. Thousand Oaks, CA: Sage Publications, Inc.

Greene, J. C., L. Destefano, H. Burgon, and J. Hall. 2006. "An Educative, Values-Engaged Approach to Evaluating STEM Educational Programs." *New Directions for Evaluation* 109:53–71.

Gribble, M. 2018. *New Zealand Guide to Pavement Structural Design*. Wellington.

Guba, E. G. and Y. S. Lincoln. 1989. *Fourth Generation Evaluation*. Thousand Oaks, CA: Sage.

Hair, J. F., R. E. Anderson, R. L. Tatham, and W. C. Black. 1998. *Multivariate Data Analysis*. 5th ed. Upper Saddle River, N.J.: Prentice-Hall.

Hall, J. N., J. Ahn, and J. C. Greene. 2012. "Values Engagement in Evaluation: Ideas, Illustrations, and Implications." *American Journal of Evaluation* 33(2):195–207.

Harper, T. L. and S. M. Stein. 1992. "The Centrality of Normative Ethical Theory to Contemporary Planning Theory." *Journal of Planning Education and Research* 11(2):105–16.

Herbert, J. P. n.d. "School Evaluation in Australia." 22–25. Retrieved February 10, 2012 (http://www.directions.usp.ac.fj/collect/direct/index/assoc/D769844.dir/doc.pdf).

Hertting, N. and E. Vedung. 2012. "Purposes and Criteria in Network Governance Evaluation: How Far Does Standard Evaluation Vocabulary Takes Us?" *Evaluation* 18(1):27–46.

Hooker, B. 2002. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Oxford University Press.

House, E. R. 1977. *The Logic of Evaluative Argument*. UCLA: Center for the Study of Evaluation.

House, E. R. 1980a. *Evaluating with Validity*. Beverley Hills, CA: Sage.

House, E. R. 1980b. "The Role of Theories of Justice in Evaluation - Justice on Strike." *Educational Theory* 30(1):67–72.

House, E. R. 1994. "The Future Perfect of Evaluation." *Evaluation Practice* 15(3):239–47.

House, E. R. 1995. "Putting Things Together Coherently: Logic and Justice." *New Directions for Evaluation* 1995(68):33–48.

House, E. R. 2001. "Unfinished Business: Causes and Values." *American Journal of Evaluation* 22(3):309–15.

House, E. R. and K. R. Howe. 1999. *Values in Evaluation and Social Research*. Thousand Oaks, CA: Sage Publications, Inc.

House, E. R. and K. R. Howe. 2000. "Deliberative Democratic Evaluation." Pp. 3–12 in *Evaluation as a Democratic Process: Promoting Inclusion, Dialogue and Deliberation*. Vol. 85, edited by K. E. Ryan and L. DeStefano. New Directions for Evaluation.

Hurteau, M., S. Houle, and S. Mongiat. 2009. "How Legitimate and Justified Are Judgments in Program Evaluation?" *Evaluation* 15(3):307–19.

Johnson, R. B., A. J. Onwuegbuzie, and L. A. Turner. 2007. "Toward a Definition of Mixed Methods Research." *Journal of Mixed Methods Research* 1(2):112–33.

Julnes, G. 2012. "Managing Valuation." Pp. 3–15 in *Promoting valuation in the public interest: Informing policies for judging value in evaluation*. Vol. 133, edited by G. Julnes. New Directions for Evaluation.

Kallemeyn, L. M., L. M. Hall, N. Friche, and C. McReynolds. 2015. "Cross-Continental Reflections on Evaluation Practice: Methods, Use and Valuing." *American Journal of Evaluation* 36(3):339–59.

Kawakami, A. J., K. Aton, F. Cram, M. K. Lai, and L. Porima. 2008. "Improving the Practice of Evaluation through Indigenous Values and Methods." in *Fundamental issues in evaluation*, edited by N. L. Smith and P. R. Brandon. New York: The Guilford Press.

Kerr, S. 2012. “Kaupapa Māori Theory-Based Evaluation.” *Evaluation Journal of Australasia* 12(1):6–18.

King, J., K. McKegg, J. Oakden, and N. Wehipeihana. 2013. “Rubrics: A Method for Surfacing Values and Improving the Credibility of Evaluation.” *Journal of MultiDisciplinary Evaluation* 9(21):11–21.

Kirkhart, K. E. and M. C. Ruffolo. 1993. “Value Bases of Case Management Evaluation.” *Evaluation and Program Planning* 16:55–65.

Kristjansson, K. 2006. “Emulation and the Use of Role Models in Moral Education.” *Journal of Moral Education* 35(1):37–49.

Lachausse, R. G. 2011. “Using Stakeholder Values to Determine Evaluation Questions in Health Evaluation.” in *American Evaluation Association Conference*. Anaheim, CA.

Lincoln, Y. S. and E. G. Guba. 1979. “The Distinction Between Merit and Worth in Evaluation.” in *Annual Meeting of the Evaluation Network*. Cincinnati.

Lincoln, Y. S. and E. G. Guba. 1986. “Trustworthiness and Naturalistic Evaluation.” Pp. 73–84 in *Naturalistic Evaluation*, edited by D. D. Williams. New Directions for Evaluation.

Lunt, N. T. and L. G. Trotman. 2005. “A Stagecraft of New Zealand Evaluation.” *Evaluation Journal of Australasia* 5(1):3–10.

Mackay, K. 1998. *The Development of Australia’s Evaluation System*. ECD 4. Washington DC.

Magendanz, D. 2003. “Conflict and Complexity in Value Theory.” *The Journal of Value Inquiry* 37:443–53.

Mark, M. M., G. T. Henry, and G. Julnes. 2000. *Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Public and Nonprofit Policies and Programs*. San Francisco: Jossey-Bass.

Mark, M. M. and R. L. Shotland. 1985. “Stakeholder-Based Evaluation and Value Judgments.” *Evaluation Review* 9(5):605–626.

Markiewicz, A. and I. Patrick. 2016. *Developing Monitoring and Evaluation*

Frameworks. Thousand Oaks, CA: Sage Publications, Inc.

Martens, K. 2018. "Rubrics in Program Evaluation." *Evaluation Journal of Australasia* 18(1):21–44.

Martens, K. S. 2018. "How Program Evaluators Use and Learn to Use Rubrics to Make Evaluative Reasoning Explicit." *Evaluation and Program Planning* 69(14):25–32.

Mathison, S. 2005. *Encyclopedia of Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Mcphee, I. 2006. "Evaluation and Performance Audit: Close Cousins – or Distant Relatives?" in *Canberra Evaluation Forum*. Canberra: Australian National Audit Office.

Mepham, B., M. Kaiser, E. Thorstensen, S. Tomkins, and K. Millar. 2006. *Ethical Matrix Manual*. The Hague: LEI.

Mertens, D. M. 2009. *Transformative Research and Evaluation*. New York, NY: Guilford Press.

Mertens, D. M., K. L. Bledsoe, M. Sullivan, and A. Wilson. 2010. "Utilization of Mixed Methods for Transformative Purposes." Pp. 193–214 in *SAGE Handbook of Mixed Methods in Social & Behavioural Research*, edited by A. Tashakkori and C. Teddlie. Thousand Oaks, CA: Sage Publications, Inc.

Miles, M. B. and A. M. Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. 2nd ed. Thousand Oaks, CA: Sage.

Ministry of Social Development. 2004. "Ngā Ara Tohutohu Rangahau Māori: Guidelines for Research and Evaluation with Māori."

Moewaka Barnes, H. 2003. "Maori and Evaluation: Some Issues to Consider." in *Evaluating Policy and Practice: A New Zealand Reader*, edited by C. Davidson, N. Lunt, and K. McKegg. Wellington: Pearson Education.

Montrosse-Moorhead, B., A. P. Bellara, and A. J. Gambino. 2017. "Communicating About Evaluation: A Conceptual Model and Case Example." *Journal of MultiDisciplinary Evaluation* 13(29):16–30.

Newman, D. L. and R. D. Brown. 1996. *Applied Ethics for Program Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Noddings, N. 1984. *Caring, A Feminine Approach To Ethics & Moral Education*. Berkeley, CA: University of California Press.

Nunns, H. 2017. "The Practice of Evaluative Reasoning in the Aotearoa New Zealand Public Sector: A Thesis Presented in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy at Massey University , Wellington , New Zealand." Massey University.

Nunns, H., R. Peace, and K. Witten. 2015. "Evaluative Reasoning in Public-Sector Evaluation in Aotearoa New Zealand: How Are We Doing?" *Evaluation Matters—He Take Tō Te Aromatawai* 1:137–63.

O'Connor-Fleming, M. L., E. Parker, H. C. Higgins, and T. Gould. 2006. "A Framework For Evaluating Health Promotion Programs." *Health Promotion Journal of Australia* 17(1):61–66.

O'Flynn, J. 2007. "From New Public Management to Public Value: Paradigmatic Change and Managerial Implications." *The Australian Journal of Public Administration* 66(3):353–66.

Oakden, J. 2013a. "Evaluation Rubrics: How to Ensure Transparent and Clear Assessment That Respects Diverse Lines of Evidence." *BetterEvaluation*. Retrieved October 20, 2013 (www.betterevaluation.org).

Oakden, J. 2013b. "Guest Blog: Why Rubrics Are Useful in Evaluations." *BetterEvaluation*. Retrieved March 15, 2013 (https://www.betterevaluation.org/en/blog/guest-rubrics).

Okili, C. and S. D. Pawlowski. 2004. "The Delphi Method as a Research Tool: An Example, Design Considerations and Applications." *Information and Management* 42(1):15–29.

Ozeki, S., C. L. Coryn, and D. C. Schroter. 2019. "Evaluation Logic in Practice: Findings from Two Empirical Investigations of American Evaluation Association Members." *Evaluation and Program Planning* 76:1–7.

Patton, M. Q. 2002. *Qualitative Research & Evaluation Methods*. 3rd ed. Thousand

Oaks, CA: Sage Publications, Inc.

Patton, M. Q. 2008. *Utilization-Focused Evaluation*. 4th ed. Thousand Oaks, CA: Sage.

Peck, L. R. 2018. "The Big Evaluation Enterprises in the United States." in *The Evaluation Marketplace: Exploring the Evaluation Industry*. Vol. 160, edited by S. B. Nielsen, S. Lemire, and C. A. Christie. New Directions for Evaluation.

Perrett, R. W. and J. Patterson. 1991. "Virtue Ethics and Maori Ethics." *Philosophy East and West* 41(2):185–202.

Perrin, B. 2003. "Some Thoughts and Reflections from Europe, Australia and the United Kingdom: A Practical Take on the Art of Evaluation." Pp. 236–39 in *Evaluating policy and practice: A New Zealand reader*, edited by N. Lunt, C. Davidson, and K. McKegg. Auckland: Pearson Education.

Pihama, L., F. Cram, and S. Walker. 2002. "Creating Methodological Ppace: A Literature Review of Kaupapa Maori Research." *Canadian Journal of Native Education* 26:30–43.

Plano Clark, V. L. and M. Badiee. 2010. "Research Questions in Mixed Methods Research." Pp. 275–304 in *SAGE Handbook of Mixed Methods in Social and Behavioural Resaerch*, edited by A. Tashakkori and C. Teddlie. Thousand Oaks: Sage Publications, Inc.

Pollitt, C. 1993. *Managerialism and the Public Services: The Anglo-American Experience*. 2nd ed. Oxford: Blackwell.

Popkin, R. H. and A. Stroll. 1993. *Philosophy Made Simple*. 2nd ed. New York: Three Rivers Press.

Punch, K. F. 2005. *Introduction to Social Research: Quantitative and Qualitative Approaches*. Second. London: Sage Publications, Inc.

Putnam, H. 2002. *The Collapse of the Fact/Value Dichotomy and Other Essays*. Cambridge, MA: Harvard University Press.

Rescher, Nicholas. 1969. *Introduction to Value Theory*. Englewood Cliffs, N.J.: Prentice-Hall.

Rex, L. A., E. E. Thomas, and S. Engel. 2010. "Applying Toulmin: Teaching Logical

Reasoning and Argumentative Writing." *The English Journal* 99(6):56–62.

Roorda, M. and R. Peace. 2008. "Challenges to Implementing Good Practice Guidelines for Evaluation with Māori a Pākeha Perspective." *Social Policy Journal of New Zealand* (34).

Rossi, P. H., M. H. Lipsey, and H. E. Freeman. 2004. *Evaluation: A Systematic Approach*. 7th ed. Los Angeles, CA: Sage.

Ryan, B. 2003. "Death by Evaluation: Reflections on Monitoring and Evaluation in Australia and New Zealand." *Evaluation Journal of Australasia* 3(1):6–16.

Sadler, D. R. 1985. "The Origins and Functions of Evaluative Criteria." *Educational Theory* 35(3):285–97.

Sanderson, I. 2002. "Evaluation, Policy Learning and Evidence-Based Policy Making." *Public Administration* 80(1):1–22.

Sanderson, I. 2010. "Is It 'What Works' That Matters? Evaluation and Evidence-Based Policy-Making." *Research Papers in Education* 18(4):331–45.

Sapnas, K. G. and R. A. Zeller. 2002. "Minimizing Sample Size When Using Exploratory Factor Analysis for Measurement." *Journal of Nursing Measurement* 10(2):135–54.

Schwandt, T. A. 1989. "Recapturing Moral Discourse in Evaluation." *Educational Researcher* 18(8):11–17.

Schwandt, T. A. 1992. "Better Living Through Evaluation? Images of Progress Shaping Evaluation Practice." *Evaluation Practice* 13(2):135–44.

Schwandt, T. A. 1997. "Evaluation as Practical Hermeneutics." *Evaluation* 3(1):69–83.

Schwandt, T. A. 1997. "Reading the 'Problem of Evaluation' in Social Inquiry." *Qualitative Inquiry* 3(1):4–25.

Schwandt, T. A. 1998. "Constructivist, Interpretivist Approaches to Human Inquiry." in *The Landscape of Qualitative Research: Theories and Issues*, edited by N. K. Denzin and Y. S. Lincoln. Thousand Oaks, CA: Sage.

Schwandt, T. A. 2008. "Educating for Intelligent Belief in Evaluation." *American*

Journal of Evaluation 29(139):139–50.

Schwandt, T. A. 2015. *Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice*. Stanford, CA: Stanford University Press.

Schwandt, T. A. 2018a. "Acting Together in Determining Value : A Professional Ethical Responsibility of Evaluators." *Evaluation* 24(3):306–17.

Schwandt, T. A. 2018b. "Evaluative Thinking As A Collaborative Social Practice: The Case of Boundary Judgment Making." Pp. 125–37 in *Evaluative thinking*. Vol. 158, edited by A. T. Vo and T. Archibald.

Scriven, M. 1972. "Prose and Cons about Goal-Free Evaluation." *The Journal of Educational Evaluation* 3(4):1–4.

Scriven, M. 1986. "New Frontiers of Evaluation." *American Journal of Evaluation* 7(1):7–44.

Scriven, M. 1990. "The Evaluation of Hardware and Software." *Studies in Educational Evaluation* 16:3–40.

Scriven, M. 1991. *Evaluation Thesaurus*. 4th ed. Newbury Park: Sage Publications, Inc.

Scriven, M. 1994. "The Final Synthesis." *American Journal of Evaluation* 15(3):367–82.

Scriven, M. 1995. "The Logic of Evaluation and Evaluation Practice." *New Directions for Evaluation* 1995(68):49–70.

Scriven, M. 1996. "The Theory behind Practical Evaluation." *Evaluation* 2:393–404.

Scriven, M. 2002. "Evaluation : Future Tense." 22(3):301–7.

Scriven, M. 2007a. "The Logic and Methodology of Checklists." Retrieved February 19, 2015 (https://wmich.edu/sites/default/files/attachments/u350/2014/logic%26methodology_dec07.pdf).

Scriven, M. 2007b. "The Logic of Evaluation." Pp. 1–16 in *Dissensus and the Search for Common Ground, CD-ROM*, edited by H. V. Hansen, C. W. Tindale, J. A. Blair, R. H. Johnson, and D. M. Godden. Ontario: OSSA: Windsor.

Scriven, M. 2012. "The Logic of Valuing." Pp. 17–28 in *Promoting valuation in the public interest: Informing policies for judging value in evaluation*. Vol. 133, edited by G. Julnes. New Directions for Evaluation.

Scriven, M. 2013a. "Conceptual Revolutions in Evaluation: Past, Present, and Future." Pp. 167–79 in *Evaluation Roots*, edited by M. C. Alkin. Los Angeles, CA: Sage.

Scriven, M. 2013b. "Key Evaluation Checklist (KEC)." Retrieved March 19, 2014 (http://www.michaelscriven.info/images/KEC_3.22.2013.pdf).

Scriven, M. 2016. "The Last Frontier of Evaluation: Ethics." Pp. 11–48 in *Evaluation for an equitable society*, edited by S. I. Donaldson and R. Picciotto. Charlotte, NC: Information Age Publishing, Inc.

Scriven, M., E. J. Davidson, and S. King. 2012. "Biting the Evaluative Bullet." Pp. 1–8 in *Conference of the American Evaluation Association*. Minneapolis.

Shadish, W. R. 1994. "Need-Based Evaluation Theory: What Do You Need to Know to Do Good Evaluation?" *Evaluation Practice* 15(3):347–58.

Shadish, W. R., T. D. Cook, and L. C. Leviton. 1991. *Foundations of Program Evaluation: Theories of Practice*. Newbury Park, CA: Sage Publications, Inc.

Shadish, W. R. and J. K. Luellen. 2005. "History of Evaluation." Pp. 183–86 in *Encyclopedia of Evaluation*, edited by S. Mathieson. Thousand Oaks, CA: Sage Publications, Inc.

Sharp, C. 2003. "Development of Program Evaluation in Australia and the Australasian Evaluation Society - the Early Decades." *Evaluation Journal of Australasia* 3(2).

Shipman, S. 2012. "The Role of Context in Valuing Federal Programs." Pp. 53–63 in *Promoting valuation in the public interest: Informing policies for judging value in evaluation*. Vol. 133, edited by G. Julnes. New Directions for Evaluation.

Silverman, D. 2006. *Interpreting Qualitative Data*. London: Sage Publications, Inc.

Smith, J. and A. Sweetman. 2009. "Putting the Evidence in Evidence-Based Policy." Pp. 59–101 in *Strengthening Evidence Based Policy in the Australian Federation*, edited by Productivity Commission. Canberra: Productivity

Commission.

Smith, N. L. 1987. "Toward the Justification of Claims in Evaluation Research." *Evaluation and Program Planning* 10:309–14.

Sobaith, A. E., C. Ritche, and E. Jones. 2010. "Consulting the Oracle? Applications of Modified Delphi Technique to Qualitative Research in the Hospitality Industry." *International Journal of Contemporary Hospitality Management* 24(6):886–906.

Social Policy Evaluation and Research Unit and Social Policy Evaluation and Research Unit Social Policy Evaluation and Research Unit. 2018. "Bridging Cultural Perspectives." Retrieved September 26, 2019 (https://thehub.sia.govt.nz/resources/bridging-cultural-perspectives/).

Stake, R. et al. 1997. "The Evolving Syntheses of Program Value." *American Journal of Evaluation* 18(1):89–103.

Stake, R. E. 1975. "To Evaluate An Arts Program." in *Evaluating the arts in education: A responsive approach*, edited by R. E. Stake. Columbus, OH: Charles E. Merrill.

Stake, R. E. 2001. "Representing Quality in Evaluation." Pp. 3–11 in *Visions of Quality*, edited by A. P. Benson, D. M. Hinn, and C. Lloyd. Oxford: Elsevier Science.

Stake, R. E. 2004. *Standards-Based and Responsive Evaluation*. Thousand Oaks, CA: Sage.

Stake, R. E. and T. A. Schwandt. 2006. "On Discerning Quality in Evaluation." in *The Sage Handbook of Evaluation*, edited by I. F. Shaw, J. C. Greene, and M. M. Mark. Thousand Oaks, CA: Sage Publications, Inc.

Stame, N. 2018. "Strengthening the Ethical Expertise of Evaluators." *Evaluation* 24(4):438–51.

Stevahn, L., J. A. King, G. Ghere, and J. Minnema. 2005. "Establishing Essential Competencies for Program Evaluators." *American Journal of Evaluation* 26:43–59.

Strike, K. A. 1979. "The Role of Theories of Justice in Evaluation: Why a House Is

Not a Home." *Educational Theory* 29(1):1–9.

Stufflebeam, D. L. 2001. "Evaluation Values and Criteria Checklist." Retrieved February 15, 2018 (https://wmich.edu/sites/default/files/attachments/u350/2014/values_criteria.pdf).

Stufflebeam, D. L. 2013. "My Tribute to a Trail Blazer." in *The Future of Evaluation in Society: A Tribute to Michael Scriven*, edited by S. I. Donaldson. Charlotte, NC: Information Age Publishing, Inc.

Stufflebeam, D. L. and A. J. Shinkfield. 2007. *Evaluation Theory, Models, and Applications*. San Francisco, CA: Jossey-Bass.

Superu and Aotearoa New Zealand Evaluation Association. 2015. "Evaluation Standards for Aotearoa New Zealand." Retrieved February 10, 2016 (http://www.superu.govt.nz/sites/default/files/Superu_Evaluation_standards.pdf).

Taylor, P. 1961. *Normative Discourse*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

The Department of Prime Minister and Cabinet. 2008. "Cabinet Manual." Retrieved February 12, 2012 (http://www.cabinetmanual.cabinetoffice.govt.nz/%5Cnhttp://www.cabinetmanual.cabinetoffice.govt.nz/files/manual.pdf).

Thomas, P. and C. Palfrey. 1996. "Evaluation: Stakeholder-Focused Criteria." *Social Policy and Administration* 30(2)(2):125–42.

Torrie, R., M. Roorda, R. Peace, M. Dalgety, and R. Bailey. 2015. "Peeling Open The Kiwi: Reterritorializing (Pakeha/White) Evaluation in Aotearoa New Zealand." Pp. 131–49 in *Continuing the Journey to Reposition Culture and Cultural Context in Evaluation Theory and Practice*, edited by S. Hood, R. Hopson, and H. Frierson. Charlotte, NC: Information Age Publishing, Inc.

United States Government Accountability Office. 2011. "Government Auditing Standards." 1–234.

Vedung, E. 2010. "Four Waves of Evaluation Diffusion." *Evaluation* 16(3):263–77.

Vestman, O. K. and R. F. Conner. 2006. "The Relationship Between Evaluation and

Politics." Pp. 225–42 in *The Sage Handbook of Evaluation*, edited by I. Shaw, J. Greene, and M. Mark. London: Sage Publications, Inc.

Wehipeihana, N., E. J. Davidson, and K. Mckegg. 2010. "What Does It Take to Do Evaluation in Communities and Cultural Contexts Other Than Our Own?" *Journal of MultiDisciplinary Evaluation* 6(13):182–92.

Wehipeihana, N. and K. Mckegg. 2018. "Values and Culture in Evaluative Thinking: Insights from Aotearoa New Zealand." *New Directions for Evaluation* 158:93–107.

Weiss, C. H. 1998. "Purposes of Evaluation." *Evaluation* 20–45.

Weiss, C. H. 1999. "The Interface between Evaluation and Public Policy." *Evaluation* 5(4):468–86.

Western Michigan University Evaluation Centre. 2016. "Evaluation Checklists Project Charter." Retrieved September 10, 2017 (https://wmich.edu/evaluation/checklists/checklistsabout).

Williams, B., A. Onsman, and T. Brown. 2010. "Australian Paramedic Graduate Attributes: A Pilot Study Using Exploratory Factor Analysis." *Emergency Medicine Journal* 27:794–99.

Williams, B. and S. van't Hof. 2014. "Wicked Solutions: A Systems Approach to Complex Problems." Retrieved July 10, 2015 (http://www.bobwilliams.co.nz/ewExternalFiles/Wicked.pdf).

Yarbrough, D. B., L. M. Shulha, R. K. Hopson, and F. A. Caruthers. 2011. *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*. 3rd ed. Los Angeles: Sage Publications, Inc.

Young Cho, J. and E. Lee. 2014. "Reducing Confusion about Grounded Theory and Qualitative Content Analysis: Similarities and Differences." *The Qualitative Report* 19(32):1–20.

Appendix A: Online survey of evaluators

Developing evaluation criteria for public sector initiatives

About this survey

Hello/Tēnā koe

Thank you for your interest in participating in this survey about developing evaluation criteria for public sector initiatives. The survey is part of a wider PhD study that aims to better understand what is required of criteria to deliver valid and defensible evaluative conclusions. The focus of the study (and this survey) is on evaluations where it is expected the evaluator will make an assessment of a policy or program, e.g. how good it is, and whether it's good enough.

The questionnaire will take approximately 6-10 minutes to complete. You will be asked for your name and email address so you can be invited to participate in Round 2 of the survey. The research findings will be reported anonymously. You will be sent a summary of the findings when they are completed.

Participation in the survey is voluntary. There are no chocolates or other incentives to participate but you may find the survey a good opportunity to reflect on your evaluation practice. By completing the survey we assume that you have given your consent to participate. Please read the Information Sheet for full details. If you have any questions about the survey please contact Mathea Roorda: m.roorda@student.unimelb.edu.au.

First, some questions about you

1. I have been involved in evaluation for:

- ◯ Less than five years
- ◯ 5-9 years
- ◯ 10-14 years
- ◯ 15 or more years

2. I manage or conduct evaluations for the: (please tick all that apply)

- ☐ Australian Commonwealth Government
- ☐ Australian States/Territories
- ☐ New Zealand Central Government

3. I currently work in a: (please tick all that apply)

- ☐ Privately-owned company
- ☐ Government agency
- ☐ Not-for-profit organisation
- ☐ University

Other (please specify)

Developing evaluation criteria for public sector initiatives

4. My evaluation roles include: (please tick all that apply)

- ☐ Conducting evaluations
- ☐ Commissioning evaluations

Other (please specify)

5. Approximately how much of your evaluation work has a focus on public sector initiatives?

Very little		Half		All
○	○	○	○	○

Many public sector evaluations are expected to make assessments about how good a policy or program is, and whether it's good enough. This involves identifying the relevant dimensions of quality and value that the evaluation will focus on.

For example, in a government grant program aimed at assisting the forestry industry to achieve a sustainable future, the evaluation may focus on utilisation of new and innovative technologies and changes to environmental impact. Alternatively - or in addition - the evaluation may focus on development of commercial alliances, creation of new jobs, improved safety, efficiency and competitiveness of operations.

6. In the public sector evaluations I am involved with, assessments are generally made against: (tick all that apply)

- ☐ Key evaluation questions
- ☐ Criteria of merit
- ☐ Performance standards

Other (please specify)

Sources used to decide what to focus an evaluation on

Below is a list of potential sources evaluators can use to identify the dimensions that will distinguish (for example) a good or well functioning program/policy from one that is poor.

Please rate these items as follows. (A) Reflect on the evaluations you are involved with and identify the most common sources used to decide dimensions of quality and value. (B) Give your opinion about importance of these sources. Are there some you regard as more important than others?

Note: frequency of use may, or may not, correspond to level of importance. This is part of what the survey is designed to capture.

Developing evaluation criteria for public sector initiatives

7. When developing evaluation criteria, or deciding what is important in an evaluation:

	A) I use this source:	B) I think this source is:
Stakeholders affected by the policy/program		
Evaluation contract manager		
Policy documents		
Program staff		
Needs assessment		
Policy staff responsible for developing the initiative		
Ethical guidelines or standards		
Content expertise		
Evaluator/s		
Lobby groups		
Program or policy objectives		
Literature on 'best practice' for the intervention		
Evaluation commissioner/sponsor		
Government legislation, standards or regulations		
Cultural expertise		
Political Minister		
The government's strategies		

8. Have any sources been missed? If so, list them here.

1)
2)
3)
4)

Help us to understand your responses

9. In the evaluations you are involved with, what factors influence decisions about which sources are used?

Developing evaluation criteria for public sector initiatives

10. In your view, what makes a source more (or less) important when identifying dimensions of quality or value / evaluation criteria?

11. Do you have any additional comments about developing evaluation criteria or about this questionnaire?

Your contact details are required so you can participate in Round 2 of this...

12. So that we can contact you again, and send you the survey findings, please type your:

Name:

Email Address:

Appendix B: Information sheet for online survey

- Do you have five years or more evaluation experience?
- Do you conduct or manage public sector evaluations in Australia/Aotearoa New Zealand?

If you can answer **yes** to both questions, read on.

Dear AES/ANZEA member

You are invited to participate in a study on evaluation criteria being conducted by PhD student Mathea Roorda. The purpose of this study is to draw on the knowledge of evaluation managers and practitioners as well as the literature to better understand what is required of criteria to deliver evaluative conclusions that are credible and warranted.

The study has several phases using different methods. You are invited to take part in Phase 1 by participating in a Delphi study. The Delphi technique is based on the idea that participants' responses are summarised and returned to the group to consider again until a consensus is approached. You will be asked to participate in at least two rounds of feedback. Each round should take no more than 20 minutes of your time.

What will I have to do?

This first round includes an introduction to the research issue and a list of characteristics about evaluation criteria that have been sourced from evaluation guidelines and international evaluation literature. You will be asked to:

- rate each item in terms of how important you think it is to the development of 'good' evaluation criteria
- add any characteristics of 'good' criteria you think are missing.

Your tasks in round 2 will be similar to the first but will include a refined list based on an analysis of round 1. The aim is to reach a consensus about which characteristics are important for the development of criteria in defensible public

sector evaluations. Round 3 will proceed only if further clarification on 'good' criteria is needed.

If you agree to participate, you will be asked to provide your email address to Mathea so you can receive a link to the subsequent rounds. You can withdraw from the study at any time.

How do I participate in the study?

To participate, please click here [link to be inserted] or copy the link into your internet browser.

What will happen to the information I give you?

Mathea is the only person who will have access to the information gathered from the Delphi study. She will keep all your information on a password-protected computer. She will do everything she can to protect your anonymity and the confidentiality of your information. The regulations of the university mean that she will keep all the data for a minimum of five years. After that it will be destroyed.

What if I say yes, then I change my mind?

If at any time you wish to withdraw from this project, you can let Mathea know and she will not use your information.

Will I receive information about the research findings?

You will be sent a summary of the Delphi Study findings when they are completed.

Where can I get further information?

If you have any questions or require more information about the study itself, you may contact Mathea Roorda by email: **m.roorda@student.unimelb.edu.au**. You can also click here [link] for more information.

The Melbourne University Human Research Ethics Committee has approved the project (Ethics ID 1340793.1). If you have any questions with regards to the ethical conduct of this study, you may contact the Executive Officer, Human Research Ethics, The University of Melbourne, on ph: +61 3 8344 2073; fax +61 3 9347 6739.

Please keep this information sheet for your records.

Thank you for your time and consideration.

Sincerely,

Associate Professor Janet Clinton (Supervisor)

Email: **jclinton@unimelb.edu.au**

Appendix C: Coding of criteria[47]

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU02	1	How effectively has NIM been administered and implemented?	Effectiveness of administration and implementation	Defined in sub questions: what have been the resource implications of implementing the program? Have suitable individuals and groups been targeted by NIM? Have people been able to transfer into and out of NIM appropriately (e.g. choosing to transfer from IM under NTER to VIM)? What has been the effect of the introduction of NIM on service providers?	Evaluation
AU02	2	Has the measure been implemented in a non-discriminatory manner?	Non-discriminatory implementation	Definition not clear. Report refers to 'no evidence that Centrelink staff are prejudicing or discriminating'.	Evaluation
AU02	3	What are the short, medium and longer-term impacts of income management on individuals, families and communities? How do these effects differ for the various streams of the program?	Impact on individuals, families and communities	Defined as indicators such as alcohol sales, school attendance, rate of hospitalisation of children, child protection notifications and substantiations, and crime and arrest rates. Findings note: disproportionate impacts on indigenous people (fewer apply for exemptions and have a lower success).	Descriptive, informs judgement

[47] Appendix C includes reference to organisations' and program's names and initials. These have not been anonymised as they appear in publicly available reports.

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU03	4	1) Is the current structure and focus of the EKYSN appropriate?	Appropriateness of structure and focus	'Appropriateness of current structure and focus' is a heading in the findings. Implicitly defined by sub-headings and includes: governance structures (defined further in 4.1.1); program infrastructure (not defined); service implementation (not defined), performance monitoring and reporting (defined). No mention of cultural appropriateness yet this is explicit in Coronial inquiry.	Evaluation
AU03	5	2) How has EKYSN impacted on justice, health and wellbeing outcomes?	Impact on justice, health outcomes	Justice outcomes defined as: 'Measurable reduction in the rates of offending or recidivism by Indigenous Australians'. Health outcomes: defined through indicators (qual and quan) e.g. levels of marijuana, tobacco and alcohol used, physical, mental and sexual health (p. 96).	Descriptive, informs judgement
AU03	6	3) Is the EKYSN meeting its aims and objectives?	Meeting aims and objectives	Aims: expand diversity and availability of existing youth services; identify service delivery gaps; implement new youth activities; facilitate community capacity building to support ongoing activities; engage and support young people to transition from childhood into productive adulthood in EK communities by significantly increasing access to and engagement in meaningful structured activities.	Evaluation

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU09	7	Are there indications that the PPL scheme is facilitating women's labour force participation, particularly by allowing them to take time out of the workforce following a birth and facilitating workforce re-entry? Does the DAPP lead to an increase in the time that fathers and partners take off work following the birth of a child?	Impact on women's labour force participation. fathers/partners taking increased time off work	'Mothers' includes groups of special policy interest (single mothers, Aboriginal or Torres St Is mothers, mothers from culturally and linguistically diverse backgrounds, mothers employed on casual contracts or self-employed).	Descriptive, informs judgement
AU09	8	Are there indications the PPL scheme is enhancing the health and wellbeing of babies and mothers, by enabling mothers to spend longer at home with their newborn children?	Outcomes: Health and wellbeing of babies and mothers	Mothers health defined as physical and mental wellbeing; reducing stress.	Descriptive focus informing judgement
AU09	9	Are there indications that the PPL scheme is encouraging gender equality and improving the balance of family and work life in Australian families?	Gender equity and work life balance	Not explicitly defined. Implicitly, findings focused on sharing childcare, housework.	Descriptive focus informing judgement
AU09	10	Does the DAPP lead to increased opportunities for fathers and partners to bond with their newborn? …lead to fathers and partners taking a greater share of caring responsibilities for their newborn? …lead to fathers and partners providing greater support for mothers following the birth of a child?	Effectiveness	Defined as fathers/partners bonding with newborn, taking greater responsibility for care of child; supporting mothers.	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU10	11	What have been the effects of the CDCT on program participants, their families and the broader community?	Effect on participants, families, community	Defined in sub questions as: reductions in consumption of alcohol, illegal drug use, gambling; crime, violence and harm related to these behaviours; increase in perceptions of safety in trial locations; any other positive impacts (e.g. increase in self-reported wellbeing, reduction in financial stress); any other unintended adverse consequences (e.g. feelings of shame, social exclusion).	Descriptive focus informing judgement
AU12	12	Examine the effectiveness of the ACBS program in teaching protective behaviours to children in the Tamworth region.	Effectiveness in teaching protective behaviours to children	Defined explicitly as 'Effectiveness in increasing children's knowledge of protective behaviours' (p. 10).	Evaluation objective
AU12	13	What are the strengths and limitations of the way the ACBS program is implemented?	No criteria evident	n/a	Descriptive focus informing judgement
AU12	14	Does the ACBS program integrate with the school curriculum and culture?	Integration	Defined explicitly as integration 'effectively with the broader educational environment of the region' (p. 10).	Descriptive focus informing judgement
AU12	15	Discuss the extent to which the program, as implemented in the Tamworth region, supports the aims of FaHCSIA's FSP.	Alignment with aims of FSP	Defined explicitly as 'alignment with the objectives of FaHCSIA's Family Support Program' (p.10). There are six FSP objectives. Evaluation notes that not all the objectives are relevant to this program (e.g. one on contributing to social and economic participation is not relevant).	Evaluation objective
AU13	16	Assess current performance of the Reconnect Program, at service level, in relation to the agreed outcomes as	Performance as outlined in agreement	Five key performance areas: 1) average cost per case; 2) contact with clients within one working day of referral; 3) targeting priority groups; 4)	Evaluation objective

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
		outlined in Reconnect Funding Agreements.		achievement of project goals; 5) completion of agreed service development. (includes rubric, p. 13).	
AU13	17	Assess appropriateness of current operation in relation to set criteria.	Appropriateness of operation against set criteria	Defined as the appropriateness of the age range, specialist services, geographical delivery areas (as outlined in Diagram 3, p. 17).	Evaluation objective
AU13	18	Assess interim outcomes nationally within the Reconnect program that are contributing to the wider strategic objectives for homelessness.	Outcomes (defined as reduction in homeliness)	Outcomes relating to employment, education and training; stability in housing; engagement with financial management services, counselling and case management; family functioning after domestic violence; receiving legal services; no exits into homelessness.	Evaluation objective
AU14	19	Assess the impact of the Package on access to services.	Impact on access to services	Defined as 'impact (of the Package) on access to services' (under 'what the eval involved'). 'Package' refers specifically to the early intervention and Autism Advisor Service components (two components of a broader package). Australian Govt Helping Children with Autism Package (2008). Aims to 'improve access to best practice early intervention, education and support for families of children diagnosed with autism. 'Access to services' is not clearly defined. Report findings provide clue about how impact is defined (Part A: Outcomes for children and families; Part B: Awareness and registration; Part C: Early intervention (access, equity, affordability, quality and best practice): Part E: Package as a Service Model.	Evaluation objective
AU15	20	What is the need for the YCBP?	Need	Defined as young people under age 25 who are carers who are providing care for someone with a disability or long-term health condition or caring	Descriptive focus informing judgement

Report ID	Statement No.	Statement indicating criteria or value construct	Value construct: Value term	Value construct: Definition/s found in report	Question type
				for elderly (aged 65 and over) who wish to continue their education.	
AU15	21	To what extent is the YCBP reaching the desired cohort of young carers?	Reach	Defined as the number, nature and characteristics of eligible and ineligible applicants. 'Nature' is not defined.	Evaluation question
AU15	22	How effective are the advertising, application and assessment processes in selecting the desired cohort of young carers?	Effectiveness of advertising, application, assessment processes	Defined in question, i.e. effectiveness refers to targeting young carers, the application and assessment processes. (p. 31)	Evaluation question
AU15	23	How appropriate is the $3000 bursary…in assisting recipients?	Appropriateness of bursary	Implicit in findings which examined how the bursary was used, bursary attrition rates and recipient perceptions of the bursary value and instalments.	Evaluation question
AU15	24	To what extent did the YCBP help recipients continue their education? Help recipients continue their caring responsibilities? Help recipients to be socially connected? Reduce the pressure experienced by recipients to work part time?	Outcomes: Contribution of program (on recipients being able to continue education, caring responsibilities etc)	Defined in sub-questions. Longer term outcomes include: educational attainment, feelings of reduced stress and recognition, change in support accessed, sense of social inclusion.	Evaluation question
AU15	25	How cost effective is the YCBP?	Cost effectiveness	Defined as administration costs per recipient.	Evaluation question

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU18	26	Examine the development and implementation of projects and how they align with the standards outlined by the National Association of Services Against Sexual Violence.	Alignment with NASASV standards	Six standards outlined: 1) using coherent conceptual approaches to project design; 2) demonstrating use of a theory of change; 3) undertaking inclusive, relevant and culturally sensitive practice; 4) undertaking comprehensive project development and delivery; 5) using effective eval strategies; 6) supporting thorough training and professional development of educators.	Descriptive focus informing judgement
AU18	27	Examine participants' knowledge, values, attitudes and beliefs around respectful dating relationships and change that might be attributed to participation in a RR program.	Change in participants' knowledge, values etc	Defined as pre-post measures of change.	Descriptive focus informing judgement
AU20	28	How, and to what extent, have the centres achieved their objectives to support the parents/carers of children with ASD to improve their capacity to participate in the community; and manage the needs of their child with an ASD; are parents, centre staff and the ASD sector satisfied with centre services?	Achievement	Defined as achievement of ASELCC Strategy objectives (p. 7).	Evaluation question
AU20	29	How, and to what extent, has each of the pilots influenced the local and wider ASD sectors; demonstrated strategies to increase access to early childhood education; delivery of best practice; collaboration with research bodies; building workforce capacity.	Influence on local/wider ASD sectors; impact on access to early childhood education, best practice; collaboration, workforce capacity	Aspects of impact are detailed in the question: 'impact on the local and wider ASD sectors; demonstrated strategies to increase access to early childhood education; delivery of best practice; collaboration with research'.	Evaluation question

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU20	30	How suitable are different service delivery models for replication elsewhere in Australia?	Replicability: Defined as suitability for replication elsewhere	Suitability of different service delivery models for replication elsewhere in Australia (p. 33). Implicitly it includes cost implications (p. 2).	Evaluation question
AU20	31	How appropriate is the program model in relation to government policy, and evidence of best practice?	Appropriateness of model (defined as aligned with govt policy, best practice)	No definition of the relevant government policy is included.	Evaluation question
AU20	32	How well does the program model align with the emerging service models and evidence of best practice?	Alignment (aligned with emerging service models, best practice)	Duplication with above question about delivery of best practice. No definition included for this criterion.	Evaluation question
AU20	33	What are the strengths and weaknesses of the program model/s and how could it be improved?	No criteria evident	n/a	Descriptive focus informing judgement
AU19	34	For people with disability who previously received supports, to what extent has the NDIS contributed to changes in the patterns and uses of supports?	Changes in patterns and uses of supports	Not defined. Findings indicate it is possibly about more flexibility around services they access (demand).	Evaluation question
AU19	35	What has been the impact of the NDIS on the overall provision and quality of disability services?	Impact on provision and quality of disability services	Not defined.	Descriptive focus informing judgement
AU19	36	What has been the impact of the NDIS on the disability sector, including the relevant govt agency in each jurisdiction and advocacy organisations?	Impact on disability sector	Not defined except implicitly as impact on disability sector, including relevant govt agency in each jurisdiction and advocacy organisations.	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU19	37	Consider impacts on workforce (skills, retention rates, capacity, satisfaction, workforce culture, composition and proportion of occupation types).	Impact on workforce	Defined as impacts on workforce (skills, retention rates, capacity, satisfaction, workforce culture, composition and proportion of occupation types).	Descriptive focus informing judgement
AU19	38	Consider impacts on supply and diversity of disability supports (particularly sustainability, ability to respond to choice and control, and service capacity).	Impact on supply and diversity of disability supports	Defined as impacts on supply and diversity of disability supports (particularly sustainability, ability to respond to choice and control, and service capacity).	Descriptive focus informing judgement
AU19	39	To what extent has the NDIS enabled people with disability, their families and carers, to have increased choice and control over their supports?	Choice and control over supports	Defined as people with disability, their families and carers, having increased choice and control over supports.	Evaluation question
AU19	40	To what extent did people have increased choice and control over the development and implementation of their plan?	Choice and control over planning	Defined as increased choice and control over the development and implementation of their plan.	Evaluation question
AU19	41	To what extent have people with disability, their families and carers been able to manage their funding on their own, customise creative sets of options for themselves, or find suitable brokers, depending on their preferences?	Ability to self-manage, self-determine	Defined as: able to manage their funding on their own, customise creative sets of options for themselves, or find suitable brokers, depending on their preferences.	Evaluation question
AU19	42	To what extent has an NDIS contributed to changes in wellbeing and quality of life for people with disability, their families and carers?	Outcomes related to wellbeing and quality of life	Defined as: changes in wellbeing and quality of life for people with disability, their families and carers.	Evaluation question
AU19	43	To what extent has an NDIS enabled people with disability, their families and carers to achieve their goals?	Achievement of goals (for participants)	Defined as disabled people, families and carers achieving their goals.	Evaluation question

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU19	44	To what extent has the NDIS helped people with disability, their families and carers during major life transitions such as starting preschool or school, leaving school, starting tertiary education, starting work, leaving home, leaving state care, leaving the workforce, and entering the aged care system?	Support during major life transitions	Defined as: helped people with disability, their families and carers during major life transitions such as starting preschool or school, leaving school, starting tertiary education, starting work, leaving home, leaving state care, leaving the workforce, and entering the aged care system.	Evaluation question
AU19	45	Have there been any other changes, including unintended changes (anticipated and unanticipated, positive and negative) in the experiences of people with disability, their families, and carers as a result of the scheme?	No criteria evident	n/a	Descriptive focus informing judgement
AU21	46	What are the short, medium and (where possible) longer-term impacts of income management on individuals, their families (particularly their children) and communities? Consider unintended consequences, positive and negative.	Impact of income management	Defined as: people on welfare payments being financially stable, meeting priority needs such as food, housing, clothing and utilities. Program objectives are to reduce immediate hardship and deprivation by directing welfare payments to the priority needs of recipients, their partner, children and any other dependents; help affected welfare payment recipients to budget so they can meet their priority needs; reduce the amount of discretionary income available for alcohol, gambling, tobacco and pornography; reduce the likelihood that welfare payment recipients will be subject to harassment and abuse in relation to their welfare payments; encourage socially responsible behaviour, particularly in the care and education of children. Short term outcomes	Descriptive focus informing judgement

Report ID	Statement No.	Statement indicating criteria or value construct	Value construct: Value term	Value construct: Definition/s found in report	Question type
				defined as within first year; medium first three years; long term in next 4-10 years.	
AU21	47	How do these effects differ for the various measures of the project? Have there been changes in spending patterns, food, alcohol, gambling, and pornography and tobacco consumption?	Changes (spending patterns…)	See above	Descriptive focus informing judgement
AU21	48	Has place-based income management contributed to changes to financial management, child wellbeing, alcohol abuse, housing and homelessness, violence and child neglect?	Changes to financial management, child wellbeing and other (expected) changes	See above	Evaluation question
AU21	49	What impact has the Matched Savings Payment had on customers' ability to manage their money, including savings?	Impact on money management	See above	Descriptive focus informing judgement
AU21	50	Do the three measures achieve appropriate outcomes (based on the aims of each measure and of income management) for their participants?	Achievement of appropriate outcomes for participants	See above	Evaluation question
AU21	51	Are there synergies or complementarities between income management and other place-based measures?	Complementarity between program and other measures	See above	Descriptive focus
AU21	52	Has the outcome of income management differed across different groups, for example, women, Indigenous people and people from culturally and	Outcomes for different groups	See above	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
		linguistically diverse backgrounds? Consider also – if sufficient data is available — location, age, educational status, work status, type of payment, length of time on income support, and family composition.			
AU21	53	Is there a stigma attached to income management and/or the BasicsCard (in the view of people on income management and merchants)?	Impact	Stigma resulting from participation in program.	Code-checker thinks Descriptive focus.
AU21	54	How effectively has place-based income management been administered and implemented? What are the regional/jurisdictional variations (if any)?	Effectiveness of administration and implementation in different areas	Defined as: 'whether it was delivered as intended to the eligible population (including access to necessary services' (p. 16, May 2014 report).	Evaluation question
AU21	55	How efficiently has place-based income management been administered and implemented?	Efficiency of administration and implementation	Not defined.	Evaluation question
AU21	56	How do the characteristics of PBIM customers compare with the eligibility criteria for placement on PBIM?	Alignment of participants with eligibility criteria	?	Descriptive focus informing judgement
AU21	57	What has been the effect of the introduction of place-based income management on service providers?	Effect	Effect of program on service providers.	Descriptive focus informing judgement
AU21	58	What is the level of take-up of Financial Management Program Services?	No criteria evident	n/a	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU21	59	What is the level of take-up of other relevant support services (e.g. Communities for Children)?	No criteria evident	n/a	Descriptive focus informing judgement
AU21	60	What is the profile of people on the different income management measures?	No criteria evident	n/a	Descriptive focus informing judgement
AU21	61	Have there been any initial process 'teething issues' that need to be addressed?	No criteria evident	n/a	Descriptive focus informing judgement
AU21	62	What are the views of participants in the place-based income management model and their families on the implementation of the project?	No criteria evident	n/a	Descriptive focus informing judgement
AU21	63	Has the measure been implemented in a non-discriminatory manner?	Non-discriminatory implementation	Defined as the 'prejudicial treatment of an individual on the basis of their actual or perceived membership of a particular group' (p. 179, Earlier baseline report. Specifically it's about being treated in a discriminatory way when considered for referral to PBIM.	Evaluation question
AU22	64	Gather information about the effectiveness of the Find and Connect Service in improving outcomes for Forgotten Australians and Former Child Migrants.	Effectiveness in improving outcomes	Defined in findings (i.e. not prescribed but emergent): working well from both client and service provider perspectives includes personalised service, sense of community generated by and within services; non-affiliation of service providers with past 'care' providers; use of social groups' dedicated service response (Exec summary).	Evaluation objective

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU22	65	Identify areas of the service that may need to be improved.	No criteria evident	n/a	Descriptive focus informing program development
AU22	66	Assess the extent to which the services were meeting the service delivery design principles agreed by the National F&C Consultative Forum in Dec 2013.	Alignment with design principles	Defined in the design principles (1) clients drive the design of services and activities; 2) system for feedback loops is provided; 3) the way services are offered (further defined, p. 70); 4) services offer flexibility in access; 5) services are respectful of the different histories and needs of FAs and FCMs; 6) services employ specialist skilled staff to deliver services.	Evaluation objective
AU24	67	What is the ability of the current service delivery system to assist older employees with a disability to retire from ADE?	Ability	Defined as whether ADEs encompass the skills and capabilities to support older employees with disabilities to transition to retirement, and what their role could be. Includes policy related issues, resourcing available in system, systems and tools to support retirement planning (in the findings).	Descriptive focus informing judgement
AU24	68	Can ADE effectively prepare employees for the transition to retirement?	Effectiveness in preparing employees for retirement	Defined in findings (includes effective case management); also includes no. of participants who have progressed toward retirement and feel optimistic about life after retirement; identified need for life-long person centred planning and other things that need to be in place.	Descriptive focus informing judgement
AU24	69	Is there potential for wider application of the model and a permanent transition to retirement program?	Wider applicability	Findings / exec summary identify what is required to successfully assist person into retirement.	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU25	70	How effective were the CDFIs in increasing the ability of disadvantaged individuals to gain access to appropriate credit?	Effectiveness	Exec summary: creating products that were attractive to people who otherwise would face financial exclusion; access to financial education, budgeting advice, money mentoring; amount of loans/no. of loans; repayment of loans. Also defined through Results based accountability framework (RBA) developed b Mark Friedman - draws attn to four areas of performance results: quantity, quality, effort and effect. Questions were designed around this framework.	Evaluation question
AU25	71	Was the demand for appropriate financial products and services best met by CDFIs?	Fit (between CDFI and participants' need for financial products)	Defined in findings and discussion (alignment between need / supported by literature and program's ability to meet that need), documenting what they do, e.g. act in interests of clients, recommending alternatives to loans where more approp., services provided such as budget or debt renegotiation.	Evaluation objective
AU25	72	What mechanisms and barriers did the CDFIs face in gaining access to capital?	No criteria evident	n/a	Descriptive focus informing program development
AU26	73	What was the level of community support before it was introduced and what is the level of community support now?	Change in community support	Defined as comparator (pre-post intervention) in question.	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU26	74	How has IM been implemented in the APY Lands? Is it as expected? Have there been any unanticipated consequences?	Implemented as expected	Defined implicitly in findings: development of info products to facilitate the introduction.	Descriptive focus informing judgement
AU26	75	What are people's perceptions of how IM is working in the APY Lands?	No criteria evident	Defined in sub questions: any groups who have particularly benefited in the early stages of the introduction to IM? Any groups where it has not been beneficial? Contribution to changes in spending patterns b directing welfare payments to priority needs? (e.g. food, alcohol, gambling, pornography and tobacco consumption); contribution to individual and community wellbeing and improved outcomes relating to financial management, child wellbeing, substance abuse, housing and homelessness (including rental arrears)? contributed to reduced levels of 'humbugging' (harassment or abuse in relation to their welfare payments).	Descriptive focus informing judgement
AU27	76	Assess the extent to which trial participants have and exercise greater control and/or choice.	Control and choice	Defined as having greater control and/or choice (in comparison with control group).	Evaluation objective
AU27	77	Identify whether (and how) the flexibility of a Career Account benefits trial participants through employment outcomes.	Employment outcomes	Benefit defined as more young job seekers with a mental illness in employment (overall goal).	Descriptive focus informing judgement
AU27	78	Identify any challenges or concerns raised by trial participants and DES providers.	No criteria evident	n/a	Descriptive focus informing program development

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU27	79	Draw out possible implications for design of support for employment outcomes for young people with mental health conditions in the future, which do not significantly increase the risk of inappropriate use of Govt funding.	No criteria evident, implicitly about sustainability	About sustainability with particular focus on appropriate use of govt funding.	Evaluation objective
AU28	80	Appropriateness: are the underlying rationales for the programs still valid? Are the programs aligned with identified need and broader govt policy? Do the programs represent the most effective way of achieving govt objectives?	Appropriateness, defined as alignment with identified need and government policy	Defined explicitly on p. 19. 'Appropriateness refers to whether the HSS and CCS programs are designed and delivered in a suitable way, targeting a community need and leading to improved outcomes for program clients.	Evaluation question
AU28	81	Effectiveness: How well do the programs meet their objectives? How well do administrative arrangements support the achievement of program objectives? How effective are the links between both programs, as well as between the programs and other settlement and mainstream services?	Effectiveness in meeting program objectives; Effectiveness of administrative arrangements; Effectiveness of links between programs and links with other services.	Defined explicitly as 'How well does the program meet its objectives? How well do contractual arrangements support the achievement of program objectives? How well do HSS services meet client needs? How effective are the links between the HSS program and other settlement and mainstream programs' (p. 36) Each of the program objectives includes additional sub-questions.	Evaluation question
AU28	82	Efficiency: How efficient are the programs from the perspective of service delivery, program management and program administration?	Efficiency of service delivery, program management & administration	Defined explicitly. Evaluators consider efficiency from 'three perspectives - namely, the delivery of services, the management of providers, the administration of the program.' (p. 85). Each is further defined, e.g. for delivery of services, 'we focused on program processes to manage expenditure on services, the relationship between program inputs and outputs. In addition, we considered whether there are alternative models	Evaluation question

Report ID	Statement No.	Statement indicating criteria or value construct	Value term	Value construct Definition/s found in report	Question type
				that would allow govt to deliver the program more efficiently'.	
AU29	83	Do all eligible job seekers receive timely and appropriate service?	Timeliness and appropriateness	Defined as indicators and as whether the removal of caps enabled services to become demand driven and the extent to which participants were satisfied with the quality and intensity of assistance.	Evaluation question
AU29	84	Are levels of access to services similar across different groups of job seekers?	Equitable accessibility	Defined as indicators, e.g. the use of services (as a proportion of target pop) by young and mature job seekers, school leavers, women and men, indigenous job seekers and job seekers from diverse cultural and linguistic backgrounds, in diff geo locations, and with different primary disabilities.	Descriptive focus informing judgement
AU29	85	Does DES lead to increased skills development and skills transfer and sustainable employment outcomes for participants?	Outcomes (defined as skills development, skills transfer and sustainable employment)	Defined as indicators, e.g. participants and employers reporting high levels of satisfaction with the program (see indicators).	Evaluation question
AU29	86	Is the DES service model more effective and efficient than the previous model?	Effectiveness and efficiency	Defined as more timely and efficient (further defined in indicators such as proportion of referrals that result in service commencement within 4 weeks of referral; mean duration between first job placement and 26 week employment outcome;	Evaluation question

				Value construct	
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
				mean duration to exit as independent worker; mean no. of employers that contribute to 26 week employment outcomes; mean cost per 26 week employment outcome; extent to which providers agree that admin load is lower under DES).	
AU50	87	How well has the CERF Program performed against its objective to disburse $100 million over four years to expand and strengthen public good environmental research?	Performance against objectives	'Performance' defined in text following question (p. 3) i.e. 'The evaluation assessed how governance, institutional arrangements and research priorities supported the delivery of Program objectives'. Also defined on p. 20: 'we expect govt to invest in the 'right' environmental research (that is, research that will have real impact on policies and decisions that matter). Also, p. 20: research that is 'done in the 'right' way (that is, high quality research output achieved efficiently'.	Evaluation question
AU50	88	How well has the CERF Program performed in terms of developing research capacity?	Performance in terms of developing research capacity	Different definitions in report! Defined as 'more environmental researchers' (p. 13); 'engagement and support of world-class researchers' (p. 1). Defined in program logic, page 4 as 'improved multidisciplinary, applied research involving users; improved knowledge and skills of researchers, more innovative research approaches developed and applied'. On p. ii there is reference to 'it has improved the capacity of researchers and their organisations to work in an interdisciplinary way to address environmental issues'.	Evaluation question
AU50	89	How well has the CERF Program performed in terms of building research knowledge?	Performance in terms of building research knowledge	Defined in text following the question: '…the ways in which knowledge, information and other products were synthesised and translated to	Evaluation question

Report ID	Statement No.	Statement indicating criteria or value construct	Value construct: Value term	Value construct: Definition/s found in report	Question type
				improve environmental decision making and management'. (p. 4)	
AU50	90	How well has the CERF Program performed in terms of use of its outputs by decision and policymakers?	Performance in terms of use of research outputs	Defined in text following the question: '…engagement of end users in the development of priorities, projects and related activities, and how research outputs were communicated to and shaped for end users'. (p. 4)	Evaluation question
AU51	91	How effective has the standard been in reducing actual (not simulated) conditioning energy use (heating and cooling) relative to houses constructed to earlier energy efficiency standards, specified as between 3.5 stars in the 2003-2005 releases of the BCA?	Effectiveness in reducing energy use	Defined as the amount of actual (not simulated) heating energy and cooling energy used, and associated costs (comparing old standard with new standard).	Evaluation question
AU51	92	What are the actual benefits and associated costs of the 5-star standard relative to the 3.5–4 star standard, in terms of construction costs, avoidable energy costs, heating and cooling appliance costs, and total lifetime costs?	Benefits and costs of the new standard relative to the old one	Costs defined in financial terms: construction costs, avoidable energy costs, appliance costs, and total lifetime cost. Benefits defined as improved homeowners' comfort, temperature difference between external environment and conditioned space, reduced greenhouse gas emissions, reduced energy bills (comparison of houses using pre-post standards). Benefits: Greenhouse gas savings is based on a measurement (CO2-e) and the factors used in the 'National Greenhouse Accounts Factors - July 2012) 'Dept of Climate Change and Energy Efficiency 2012). Costs defined as: (a) Construction costs based on 2011 prices, related to those elements of the building directly related to the star rating (window area, window orientation, glazing material, floor, wall and roof materials, wall area, insulation); b) energy costs saved by consumer.	Descriptive focus informing judgement

				Value construct	
Report ID	Statement No.	Statement indicating criteria or value construct	Value term	Definition/s found in report	Question type
				Energy costs defined as 'standard energy tariff rates in each city).	
AU52	93	Effectiveness of NERP	Effectiveness	How well the program has met the program objectives (p. 8). Impact of program. (p. 14)	Evaluation objective
AU52	94	Efficiency of NERP	Efficiency	'How administrative costs have been kept to a minimum, the attraction of additional resources (whether cash or in-kind) and how well research hubs, relevant institutions and the department have delivered the program and directed resources to the conduct, collaboration and communication of research.' (p. 8)	Evaluation objective
AU63	95	Evaluate the contribution of Commonwealth environmental watering to the objectives of the MDBA's Environmental Watering Plan.	Contribution of program to objectives of EWP	Methodological – multiple sources of data used to understand past flow events, flow-ecology relationships.	Evaluation objective
AU63	96	Evaluate the ecological outcomes of Commonwealth environmental watering at each of the seven selected sites.	Outcomes - ecological	Ecosystem diversity, hydrology, stream metabolism and water quality, vegetation diversity, fish, generic diversity (each is defined on p. 5). (All are environmental indicators).	Evaluation objective
AU80	97	To what extent have the objectives and outcomes of the CFF program(s) been achieved?	Achievement of program objectives, outcomes	The short term objectives of CFF are to identify, trial and encourage the adoption of practices to reduce greenhouse gas emissions, adapt to a more variable climate and increase soil carbon sequestration. The KPIs refer to - no. of research projects producing high quality research outcomes with the potential to inform new practices/technologies; no. of projects delivering trials and demonstrations of practices and technologies; no. of projects communicating info about relevant policies, practices and technologies	Evaluation question

			Value construct		
Report ID	Statement No.	Statement indicating criteria or value construct	Value term	Definition/s found in report	Question type
				to farmers, land owners and the broader agriculture industry.	
AU80	98	How effective and efficient was the implementation, administration and service delivery of CFF programs and projects?	Effectiveness and efficiency of implementation, administration, service delivery	Effectiveness and efficiency not explicitly defined. Implicitly, it includes alignment with the Commonwealth Grant Rules and Guidelines - this includes applying input, output and outcome measures (p. 30); meeting planned milestones and funding KPIs (p. 89). There is also reference to activities being completed 'within required timeframes and budget' (p. 15).	Evaluation question
AU81	99	Examination of all 108 projects funded under the TCFA IDP to assess compliance against grand deeds.	Compliance against grant deeds	Defined as compliance against deed requirements including exercise of delegations and approvals, acquittal of funds and supporting documentation, monitoring progress and reporting, identifying performance improvement opportunities in line with grant requirements.	Evaluation objective
AU81	100	Analysis of overall outcomes from the TCFA IDP.	Outcomes	No clear definition, which is noted in the evaluation. Evaluator states they did not attempt to develop a definition.	Evaluation objective
AU83	101	Evaluate the performance of the RFCS service providers against the following criteria: corporate governance, financial management, service delivery.	Performance in relation to corporate governance, financial management, service delivery	Defined as performance against corporate governance, financial management, service delivery. Each criterion is defined in 1.6 (Assessment Methodology).	Evaluation objective
AU83	102	Evaluate the service providers' performance and compliance with their 2011-15 Deeds of Grant and the RFCS Program Resource Manual.	Performance and compliance in relation to grant deeds and program resource manual	Not defined.	Evaluation objective

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
AU84	103	Assess the overall economic feasibility of all operations (by the APLC, states and landholders) currently underway to control the 2010-11 locust plague.	Economic feasibility of current locust plague operations	Benefits from the control campaigns by all parties concerned are compared with costs specifically incurred in executing the control operations for the current outbreak. And on p. 1: 'Other non-market costs and benefits that may result from control of locust outbreaks such as costs to human health from chemical use or benefits to other industries and the public - are discussed by not estimated'.	Evaluation objective
NZ01	104	In this trial a group of children and families receiving the Early Start programme would be contrasted with an equivalent group of children and families not receiving the service on a series of measures that reflected the goals of the Early Start service.	Comparison of goal-based outcomes (coded as outcomes)	Measures that reflect the goals of the ES service. Measures spanned: health outcomes; pre-school education and welfare use; parenting; child abuse and neglect; and child behaviour.	Purpose
NZ03	105	Are improvements in child behaviour and parenting practice maintained over the longer follow-up period?	Maintenance of post-intervention gains	Maintenance/stability of improvements in child behaviour/parenting practice (as defined by scales such as Eyeberg Child Behaviour Inventory).	Descriptive focus informing judgement
NZ03	106	How much drop-off in efficacy is evident between the post-course and 30-month follow-up measures?	Efficacy	Drop off in efficacy.	Descriptive focus informing judgement
NZ03	107	Do sub-group differences identified in the Pilot Study persist?	Persistence of sub-group differences	Sub-group = Māori and non-Māori. The study also examines between-site variations.	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
NZ05	108	Is the transition and integration back into the community working well?	Outcomes related to transition and integration into the community	Note clear. P. 3 refers to 'Support provided to young people after they leave TPW'. P. 4 refers to 'successful transition' meaning 'they had not reoffended in over 12 months or had committed only a single offence of a minor nature'.	Evaluation question
NZ05	109	Is the program being targeted correctly?	Correctly targeted (coded as targeted)	The MAC programme seeks to target the 40 most serious and persistent young male offenders each year.	Evaluation question
NZ05	110	Is the program being delivered consistently?	Consistent delivery	Not clear. It could be the following: 'investment into each young person has to be consistent across the group, and that dealing with a young person who is not coping with the MAC programme diverts that energy away from the others.' (p. 29). Note: no assessment of this question in the report.	Evaluation question
NZ05	111	Do the program components work well?	Working well	Core components defined on p. 11 but no definition of what 'working well' for each of the components means, or what 'working well' as a whole (i.e. integrated components) means.	Evaluation question
NZ05	112	Are people engaging positively with the program?	Positive engagement	Maintain a positive standard of behaviour throughout the programme; no negative behaviour such as disrespect for staff (p. 33 findings).	Evaluation question
NZ06	113	The aim of the present study is to examine the impact of New Zealand's TPUs on educational outcomes.	Impact of program on educational outcomes	Outcomes concerned with in this study are enrolment in a school by the young woman after giving birth and eventual attainment of NCEA Level 1 and 2 qualifications	Aim
NZ25	114	Identify changes occurring within communities – either behaviours or factors that positively influence	Outcomes - Defined as change in behaviours	See below - not sure how this question is different to below. May relate to behaviour change, as opposed to broader impact.	Evaluation objective

			Value construct		
Report ID	Statement No.	Statement indicating criteria or value construct	Value term	Definition/s found in report	Question type
		behaviours associated with family violence.	associated with family violence		
NZ25	115	Understand the impact of community-led initiatives to prevent family violence beyond the life of the funded projects.	Impact	Definition of 'impact' emerged from data. Impacts have been grouped thematically according to: • increased awareness and message infiltration; • behaviour and attitude change; • community responsiveness and ownership; • interagency collaboration; and, • statutory intervention. It's not clear how this is different to the previous question. The findings focus on impacts and do not differentiate between 'changes occurring within communities and impact'.	Evaluation objective
NZ25	116	Identify critical success factors.	No criteria evident	n/a	Descriptive focus informing judgement
NZ25	117	Identify barriers to the on-going success of the local initiatives.	No criteria evident	n/a	Descriptive focus informing judgement
NZ38	118	1. To identify the effectiveness of the FGC process for facilitating positive outcomes for children and young people and their whānau.	Effectiveness (defined as facilitating positive outcomes for children etc)	Implicitly: young people and whānau's satisfaction with process/positive feedback about immediate outcomes in terms of safety, health, behaviour and education.	Evaluation purpose
NZ38	119	2. To identify best practice for implementing FGCs.	No criteria evident	n/a	Descriptive focus informing

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
					program development
NZ38	120	3. To identify culturally appropriate practices for implementing FGCs.	No criteria evident	Implicitly in findings: use of Māori tikanga at FGCs, cultural support and supervision for staff, conducting whakapapa searches to engage with whānau and hapū. Having a Hui-a-Whānau or whānau hui prior to the FGC was thought to work well with Māori as it gives whānau more time to discuss issues and come up with solutions for the FGC plan. There were also good examples of strong relationships with local hapū/Iwi resulting in good working relationships with the community.	Descriptive focus informing program development
NZ38	121	4. To further develop the theoretical underpinnings of how the FGC process effects change for children/young people and their whānau.	No criteria evident	n/a	Descriptive focus informing knowledge development
NZ39	122	To what extent can self defence courses assist participants to: · recognise sexual and family violence and abuse; · understand that it is not okay (and not their fault); · learn clear strategies for recognising, resisting and responding to specific violent/abusive events, disclosing abuse and seeking help, supporting other people who are victims of violence and abuse; and · feel empowered and have their self-esteem enhanced.	Effectiveness (in teaching participants to recognise sexual abuse, family violence etc)	Defined in terms of question, with a focus on individual outcomes. Includes a specific focus on Māori women and girls.	Evaluation question

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
NZ39	123	Investigate the extent to which short term/intermediary outcomes, if achieved, can reduce vulnerabilities to child abuse, sexual and family violence and re-victimisation, and stranger danger.	Longer term effectiveness, defined as achievement of intended outcomes	Effectiveness in terms of girls and women retaining the knowledge, skills and attitudes to reduce their likelihood of becoming victims of abuse or sexual violence. (p. vii).	Evaluation purpose
NZ40	124	1. Are all the parts implemented, and implementation milestones achieved?	Achievement of implementation milestones	Implementation.	Evaluation question
NZ40	125	2. To what extent do the CAP components work well in themselves – reflecting formal intentions, and as experienced/perceived by users?	Working well (as expected and as experienced)	Individual components work well.	Evaluation question
NZ40	126	3. Do all the components work together as intended, and as an effective whole, as perceived by key stakeholders?	Effectiveness, defined as components working together as intended	All components work together.	Descriptive focus informing judgement
NZ40	127	4. How well, or to what extent, were problems or barriers to implementation circumnavigated or overcome?	Effectiveness (defined as ability to manage implementation issues)	Management of problems or barriers.	
NZ41	128	Did the project meet its delivery targets, and if not why not?	Achievement of delivery targets	Target definition changed. Originally it was based on estimate of likely candidates in project area. Then it became: 'Reaching the intended target group (victims at the highest risk from repeat domestic violence offenders).'	Evaluation question
NZ41	129	Were the homes safety audited, and what did the safety audits find?	No criteria evident	n/a	Descriptive focus

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
					informing judgement
NZ41	130	What were the changes made to the houses, and what equipment was provided?	No criteria evident	n/a	Descriptive focus informing judgement
NZ41	131	Was there evidence that victims were safer and/or felt safer as a result of the intervention?	Impact of program (defined as people feeling/being safer)	Reduction in high anxiety and stress & other symptoms of distress caused by domestic violence. Evidence that victims report reduced anxiety, improved sleeping, confidence, stability and concentration. Also considers children's wellbeing including reduced fear, improved sleeping, behaviour and performance at school.	Descriptive focus informing judgement
NZ41	132	What worked well/not well?	No criteria evident	n/a	Descriptive focus informing judgement
NZ42	133	Impact of program on mortality, early childhood education and health service engagement utilisation, CYF contact and maltreatment marker injuries.	Impact on various indicators	Impact on various indicators related to child mortality, early childhood education, health service engagement utilisation, CYF maltreatment marker injuries.	Evaluation aim
NZ44	134	Outcomes: young people's participation in education, training or work-based learning, budgeting and parenting activities, and achievement of NCEA qualifications.	Outcomes (defined as participation in education, training etc)	Young people's participation in education, training or work-based learning, budgeting and parenting activities, and achievement of NCEA qualifications.	Evaluation objective
NZ44	135	Impact of Youth Service on young people's benefit receipt, using a comparison group drawn from young	Impact on benefit receipt	Young people's benefit receipt, using a comparison group drawn from young people who started a main benefit before the introduction of the service	Evaluation objective

				Value construct	
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
		people who started a main benefit before the introduction of the service a process evaluation of the Youth Service implementation, six months after it started, that assessed whether it operated as intended, what worked well, and what could have been improved.		a process evaluation of the Youth Service implementation, six months after it started, that assessed whether it operated as intended,	
NZ44	136	Youth Service implementation, six months after it started, that assessed whether it operated as intended, what worked well, and what could have been improved.	Operating as intended (coded as implemented as intended)	Not defined.	
NZ50	137	How effectively and efficiently has the NZ ETS been delivered and operated (for participants and govt agencies)?	Effectiveness and efficiency	P. 12 - 3 agencies involved in operation of program, thus need to work effectively and efficiently. p. 12 - as above, 'improving operational efficiencies in the delivery of the NZ ETS is a key objective…'	Evaluation question
NZ50	138	What are the impacts of the NZ ETS on emissions and behaviour?	Impact of the NZ ETS on emissions and behaviour	In sub questions (p. 17): outcomes of program; do these deliver on the objectives? Impact on emissions, impact of NZ ETS on participants and business behaviour, impact of price on emissions and participant behavior.	Descriptive focus informing judgement
NZ51	139	Assess how effectively the WMA was implemented from a stakeholder perspective. How well was it implemented? How useful and applicable was the role of the Ministry?	Effectiveness of the implementation from the perspective of stakeholders	In Table 1 (administrative efficiency, relationships - collaboration in the sector, good practice - building capability and capacity (including infrastructure) across the sector), information, awareness and compliance (MfE's performance). Stakeholders are	Evaluation question

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
				defined in Table 3 (p. 16) - environment is not a stakeholder though.	
NZ51	140	How valuable were the outcomes to stakeholders?	Value of outcomes to stakeholders	Value of administration to applicants.	Evaluation question
NZ51	141	How did the new regulatory environment for waste management and minimalisation impact on the range of stakeholders? Has behaviour change started to embed or do stakeholders slip back into old behaviours when not monitored?	Impact on stakeholders of the new regulatory environment for waste management and minimisation	Dimensions of impact: administrative efficiency, relationships/collaboration in sector; good practice - building capability and capacity across sector; info, awareness and compliance.	Descriptive focus informing judgement
NZ53	142	Assess whether the regulations are achieving their intended purpose in terms of the policy objectives of the NESTF.	Effectiveness of the regulations in relation to the policy objectives of the NESTF	Policy objective: to provide for consistent and certain regulatory planning provisions that apply on a national basis, to assist in network and equipment design and equipment sourcing for roll-outs, and a reduction in compliance costs and timeframes. (p. 9). Further defined in Table 1 (p. 11). There are five sub-objectives: 1) assist in network and equipment design and equipment sourcing for roll-outs, 2) create a reduction in compliance costs and timeframes for service providers; 3) reduce the timeframe and lower the costs for the availability of new services to consumers; 4) contribute to a reduced workload for councils in processing and determining consent applications; 5) set an appropriate balance between local participation in community planning and cost-effective national infrastructure investment.	Evaluation objective

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
NZ53	143	Extent to which NESTF is still fit for purpose.	Fit for purpose	p. 28	Evaluation objective
NZ53	144	Ascertain whether key risks associated with the NESTF have been realised.	Realisation of key risks	Impact of the program on RMA decision-making. RMA decision-making is broken down into two sub questions: a) what impact has the NPS-REG had on regional policy statements, regional plans and district plans?; b) what impact has the NPS-REG had on resource consent decision-making?	Descriptive focus informing judgement?
NZ54	145	What impact has the NPS-REG had on RMA decision-making?	Impact on decision making	Impact of the program on RMA decision-making. RMA decision-making is broken down into two sub questions: a) what impact has the NPS-REG had on regional policy statements, regional plans and district plans?; b) what impact has the NPS-REG had on resource consent decision-making?	Descriptive focus informing judgement?
NZ54	146	To what extent is the NPS-REG achieving its objective?	Achievement of program objective	Effectiveness in contributing to NZ's electricity generation. Objective defined on p. 10: 'to recognise the national significance of renewable electricity generation activities by providing for the development, operation, maintenance and upgrading of new and existing renewable electricity generation activities, such as the proportion of NZ's electricity generated from renewable energy sources increases to a level that meets or exceeds the NZ govt's national target for renewable electricity generation'.	Evaluation question
NZ54	147	Does the NPS-REG remain appropriate in the light of government and sectoral developments?	Appropriateness	Appropriateness of the NPS-REG objective.	Evaluation question

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
NZ70	148	Extent to which the program design remains fit for purpose or needs to adapt.	Fit for purpose	Extent to which SSF is fit for purpose (makes a worthwhile and valuable contribution to primary industries and rural communities) and occupies a unique position compared to other funding programs with similar aims (findings, p. 15).	Evaluation objective
NZ70	149	To what extent and in what ways has the SFF been worth the investment to date?	Worth	Value for money in this evaluation refers to whether government and SFF beneficiaries are using their combined resources well. Funding and other resources (such as time, knowledge and skills) are limited. There is an opportunity cost associated with their use. It is therefore desirable to allocate resources to activities that return high-value outcomes. This provides the basis for addressing KEQ 2: To what extent has the SFF been worth the investment to date? Includes cultural, environmental, economic and social perspectives (Table 5).	
NZ70	150	What are the opportunities to maximise the value derived from the SFF?	No criteria evident	Defined as 'Getting a better return'.	Descriptive focus informing program development
NZ72	151	Review progress made in each of the four projects and make any recommendations as to their future direction and funding priorities.	Progress	How the programme is tracking towards its goals as set out in the original business plan, as modified from time to time by annual plans (p. 19).	Descriptive focus informing judgement
NZ72	152	Review project outputs to date and expected programme outcomes within the time frames established in the business plan (and as updated in the annual plans).	Progress	Notes the different definitions used to describe the outcome and suggests the term 'marbled grass-fed beef' be used, to reduce misunderstandings and confusion in future. (p. 10).	Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
NZ72	153	Review internal and external factors affecting the programme including management and governance.	No criteria evident	The management, governance, budgeting and reporting of the programme (p. 20).	Descriptive focus informing judgement
NZ72	154	Review the likely benefits for New Zealand of the programme including commercial, economic and spillover benefits and how to maximise benefits.	Benefits	?	Descriptive focus informing judgement
NZ73	155	Assess programme progress to date as a whole, in each of the objectives, and in particular the likelihood of the programme delivering the expected outcomes.	Achievement of progress against objectives	No definition of what 'progress' would look like.	
NZ73	156	Identify any unintended consequences (positive and/or negative).	No criteria evident		Descriptive focus informing judgement
NZ73	157	Assess internal and external factors affecting the programme including management and governance.	?		Descriptive focus informing judgement
NZ73	158	Identify any key risks to achieving the contracted outcomes.	No criteria evident		Descriptive focus informing judgement
NZ73	159	Identify potential additional benefits (spillovers).	No criteria evident		Descriptive focus informing judgement

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
NZ73	160	Identify opportunities that the programme could address in the future.	No criteria evident		Descriptive focus informing program development
NZ74	161	Review progress made to date in each of the Programme work streams.	Progress toward achievement of objectives	The objectives of the Programme are to improve the efficiency of nutrient use in agriculture in New Zealand and to reduce the losses of N and P to the environment. Economic benefits of $348 million per year by 2025.	Descriptive focus informing judgement
NZ74	162	Assess progress made against the intended Programme outcomes.	Progress toward achievement of outcomes	Outcomes defined as: A 50% improvement in N fertiliser use efficiency in pastoral systems by 2018; 20% improvement in P fertiliser use efficiency by 2020; reduction by 30% in losses of N and P. and Economic benefits of $348 million per year by 2025.	Descriptive focus informing judgement about expected outcomes
NZ74	163	Review internal and external factors affecting the Programme including management and governance.	No criteria evident	Includes contractor management, governance, stakeholder engagement and reporting. Not defined but reported in findings. Includes judgements about each.	Descriptive focus informing judgement
NZ75	164	To what extent have Māori Agribusiness Prototype projects generated public value and delivered on MPI's Our Strategy 2030 and Government Priorities?	Public value Delivery on Strategy and Government Priorities	Producing public knowledge, increasing capacity of land owners or MPI and increasing productivity. Improving the productivity of Māori primary assets; doubling the value of primary industry exports by 2025).	Evaluand aim
NZ75	165	How and in what ways have Māori Agribusiness Prototype projects assisted	No criteria evident	Defined as support to overcome challenges to increasing productivity.	Evaluand aim

			Value construct		
Report ID	**Statement No.**	**Statement indicating criteria or value construct**	**Value term**	**Definition/s found in report**	**Question type**
		Māori land owners to overcome challenges to increasing productivity?			
NZ75	166	Which new ways of working for MPI worked well and not well, and why (considering different contexts, constraints and resources)?	Working well	Relationship based (described in findings).	Descriptive focus informing judgement

Appendix D: Expert review of criteria matrix

Please review the ethics' information in the draft criteria matrix handbook, in particular section 1.4 (on page 4). For each principle, please consider whether:

A	The definition included is accurate? If not, what changes would you include?
B	The information provided is pertinent to what evaluators need to know to undertake the task (identifying dimensions of value)? If not, what changes would you include?

You may include track change comments on the criteria matrix document.

Please include your comments in the table below and return to: **mathea.roorda@unimelb.edu.au** by 27 June 2018.

Thank you.

Principle	Comment
1. Consequences	
2. Duty	
3. Rights	
4. Equity	
5. Ethic of care	
Any other comments about the principles OR the matrix heuristic?	

Appendix E: Feedback form for criteria matrix handbook reviewers

Please review the criteria matrix using the three criteria in the left-hand column. You may also include track change comments on the criteria matrix document. Please return to: mathea.roorda@unimelb.edu.au by 06 July. Thank you.

Criteria		Comment/Suggested changes
Clarity of Purpose	A succinct title clearly identifies what the tool is about	
	A brief introduction orients the user (the evaluator) to the tool's purpose, including the following: The circumstances in which it should be used How it should be used Intended users	
Organisation of content	Content is presented in a logical order.	
	Content is organised in sections labeled with concise, descriptive headings.	
	Complex steps or components are broken down into multiple smaller parts.	
Clarity of writing	Content is focused on what users should do, rather than questions for them to ponder.	
	Everyday language is used, rather than jargon or highly technical terms.	
	Verbs are direct and action oriented.	
	Terms are precise.	
	Terms are used consistently.	
	Definitions are provided where terms are used but might not be obviously known.	
	Sentences are concise.	
Additional comments?		

Appendix F: Criteria matrix handbook

Developing Defensible Evaluation Criteria

A handbook for evaluators

mathea.roorda@unimelb.edu.au

Glossary

Consequentialist	The name given to a group of normative ethical theories that focus on the consequences or outcomes of particular actions or rules, to decide whether something is good or not.
Deontological	The name given to a group of normative ethical theories that views acts as right or wrong in all circumstances regardless of the consequences.
Dimensions of value	This describes 'good' or 'right' (etc) for each interest group through the lens of the five normative principles.
Normative ethics	This refers to a branch of philosophy that is concerned with understanding right or wrong. In western philosophy theories about normative ethics can be grouped into one of three categories: consequentialist, deontological or virtue based.
Perspectives	These are ways of understanding concepts such as 'good' or 'bad'; 'right' or 'wrong' that broadly align with consequentialist, deontological or virtue-based theories.
Principles	These are orientations within a perspective. In the criteria framework there are five principles. The deontological perspective includes three distinctive orientations in how it views 'good' or 'bad'. They are the principles of duty, rights and equity.
Sources	These refer to the people, places or things where information about value perspectives originates.
Virtue-based	This refers to a perspective that aligns with ethic of care theories. It prioritises inter-personal relationships and caring. A virtue-based perspective rejects universalism and argues that care is always contextual.

Introduction

This handbook provides a step-by-step guide to identifying and justifying criteria for public sector evaluations. It introduces a tool called the criteria matrix.

What are criteria?

Criteria are the aspects of a program that will be examined to determine its quality. They describe what a good / bad / right / wrong / successful / non-successful /valuable / detrimental (etc.) program will look like.

What is the criteria matrix?

The criteria matrix applies five principles, derived from normative ethical perspectives in western philosophy, to a set of selected groups with a legitimate interest in a program. Each principle is considered from the perspective of a selected interest group, resulting in up to five descriptive 'dimensions of value' for each interest group (Box 1).

Box 1: Criteria matrix template

	Normative perspectives				
	Consequentialist	**Deontological**			**Virtue-based**
	Consequences *Maximum benefits*	**Duty** *Meeting obligations*	**Rights** *Minimum protection*	**Equity** *Being fair or equitable*	**Ethic of care** *Relationships & context*
Interest group 1					
Interest group 2			**Dimensions of value**		
Interest group 3					

Why do we need such a tool?

Systematic development of criteria is fundamental to reaching a defensible evaluative conclusion, yet this is an aspect of evaluation that has been neglected by many evaluators. A review of evaluation reports in Australia and New Zealand (Roorda 2018) suggests criteria tend to be implicit in evaluation questions or objectives. Often, criteria are not well defined or justified.

Whether we recognise it or not, criteria are informed by normative perspectives (ways of viewing good / bad / successful/ non-successful etc.). It therefore makes sense that evaluators understand what the different perspectives are and systematically consider how each might contribute to development of criteria for a particular program evaluation.

When can the criteria matrix be used?

The criteria matrix can be used as part of evaluation scoping, e.g. in a facilitated workshop with stakeholders who have knowledge of a program and/or its context, or during an evaluation, e.g. by the evaluator as a desk-based exercise, using a range of sources.

What are the interest groups?

The 'interest groups', or affected parties, depend on the program context, but might include different groups of people as well as non-humans such as animals, forests, or the environment generally.

A critical factor is that the interest groups are subjects of consideration in their own right, and not just a means to others' ends.

What are the normative principles?

Consequences: assesses a program on the basis of its outcomes. One of the most well-known consequentialist theories is utilitarianism, which views actions as right or good when they produce the greatest possible good for the greatest number of people. A program is successful if for example, it increases wellbeing or lessens suffering. The focus is on securing the best outcome for the group as a whole. A potential limitation of this perspective is that the outcomes of minority groups may

be given less priority. This perspective may also be used to justify unjust policies by citing the greatest good for the greatest number.

Duty: assesses the goodness of something based on expectations, commitments and contractual obligations. A potential limitation of this perspective is that it may over-emphasise a managerial perspective and risk neglecting the interests of other important interest groups.

Equity: assesses the goodness of something based on concepts of equity and fairness. This may require looking beyond outcomes to examine the political and social conditions that have led to these outcomes and how the program takes these conditions into account. This perspective is concerned with allocating resources fairly. A potential limitation is that people may adopt a sense of entitlement to outcomes.

Rights: assesses the goodness of something based on the rights a group has to do certain things. A right may establish a minimum level of protection for individuals, groups or entities. These minimum expectations are independent of outcomes. Rights may be justified by legal rules and regulations or principles that guide human conduct. A potential limitation of this perspective is that it may encourage individualistic behaviour, and sometimes rights are in conflict (e.g. gender versus religious rights).

Ethic of care: assesses the goodness of something based on the strength of relationships, trust and responsiveness to need. It emphasises connectedness, e.g. through concepts such as social responsibility. It includes a focus on 'care' and 'caring' and this is always contextually applied. A potential limitation of the contextual nature of care is that it can be ambiguous.

Origins of the criteria matrix

The criteria matrix has been developed by a PhD student, Mathea Roorda. It draws on a decision-making tool developed by Mepham, Kaiser, Thorstensen, Tomkins and Millar (2006), which draws on two normative perspectives (consequentialist and deontological). The principles used in the criteria matrix are those used by Newman and Brown (1996) as these authors draw on a third normative perspective (virtue-based ethics) and differentiate between three principles within a deontological perspective (duty, rights and fairness).

Limitations

The principles presented in the criteria matrix are from 'western' philosophy. A limitation is that the matrix may miss other normative perspectives, for example from te ao Māori or Aboriginal or Torres Strait Islander worldviews.

Steps for using the criteria matrix

This section describes the steps for using the criteria matrix. For the purposes of field testing, please reflect on your use of the criteria matrix using the feedback form in Appendix A.

Specifying the interest groups and principles

The first step is to customise the matrix for your evaluation. This involves identifying the relevant interest groups and the five principles.

If groups are similar in the way they are affected by a program, they can be collapsed into a broad interest group (e.g. 'government'). A critical factor is that the interest groups are subjects of consideration in their own right, and not just a means to others' ends.

The principles may need to be defined or reworded using language that makes sense for the program being evaluated. If appropriate, you may want to collapse rights and fairness into one category. Use Box 2 as a guide.

Box 2: Summary of principles used in the criteria matrix

	Defined as:	Focus is on:
Consequences	Maximum benefits; action that achieves the best outcomes	Outcomes, effects, value for money where the focus is on return on investment
Duty	Meeting obligations	Contractual obligations, commitments
Rights	Minimum protection; doing the right action	Minimum protection, rights of individuals, groups, entities
Equity	Being fair, just, equitable	Equity, fair deal for those most at risk/ disadvantaged
Ethic of care	Being virtuous, e.g. trustworthy, loyal	Meeting the needs of others; attention to context and strengthening relationships

Identifying dimensions of value

The next step is to consider dimensions of value as they relate to the five normative principles. For example, if the program is to achieve maximum benefit, what will that look like from the perspective of Interest Group 1, 2, 3? If the program is to

meet certain obligations, what will that look like for Interest Group 1, 2, 3? Box 3 includes an example of a matrix completed for an evaluation of a seasonal labour initiative in New Zealand.

Box 3: Example of a criteria matrix for a seasonal labour initiative

About the initiative: Offshore workers come to New Zealand for seven months per year to be employed in the horticulture and viticulture industries. The Recognised Seasonal Employer (RSE) initiative allows for the return of experienced workers, who have an offer of employment, and meet certain criteria, in subsequent seasons. Selection of workers is determined by employers' requirements.

The number of offshore workers approved to come to New Zealand is subject to the availability of suitable New Zealand workers. Employers are responsible for the pastoral care of workers. An inter-agency understanding exists between the New Zealand Government and selected Pacific Island countries. This sets up expectations for participating parties.

	Maximising outcomes	Rights	Fairness	Duty	Ethic of care
Employers	Sustainable supply of skilled workers. Benefits outweigh costs (business efficiency gains). Employers have the confidence to invest/grow. Horticulture and viticulture industries competitive in international markets.	Employers select workers based on their requirements.	Employers share travel costs.	Employers abide by employment contract. Employers abide by RSE permit. Illegal employment practices eradicated.	Employers help with travel costs, suitable accommodation, link to services. Employers provide pastoral care to workers (help with suitable accommodation, link to services).
Pacific workers	Earning and experience are worth the effort. Workers gain new skills. Family benefits (school fees paid, houses renovated, news businesses start up).	Workers are guaranteed 240 hours work per season. Experienced workers are rewarded with better pay. Access to dispute resolution mechanisms.	Cost is not a barrier to participate in RSE. Pacific Island workers are not charged recruitment fees. The recruitment process is fair. Workers are treated fairly in their workplaces.	Workers have a positive work ethic and are highly productive. Workers abide by their employment contract. Workers comply with the visa requirements. Workers are ambassadors for their community.	Workers are well prepared for life in New Zealand. They know their employer's name and the name of the place they are going to. Returning workers and leaders help to support new workers.
New Zealand Government	Improved governance, regional integration in Pacific.		The Government provides equal opportunities for Pacific countries to participate alongside other countries.	Government effectively manages the scheme. Government addresses problems promptly and responsively.	

	Maximising outcomes	Rights	Fairness	Duty	Ethic of care
Pacific Governments			Offshore recruitment agents are regulated.	Participating countries abide by the inter-agency understanding. They provide a work-ready pool; promote good work ethic, upholding the country's reputation. Potential workers have health checks (HIV, TB).	
New Zealand workers	New Zealand workers can earn a satisfactory income from seasonal labour.		New Zealand workers are not displaced by foreign workers. They are treated the same as other workers.	Workers have a positive work ethic and are highly productive. Workers abide by their employment contract.	
Pacific communities	Communities benefit financially from the scheme.	Community has a right to decide who can participate.	There are opportunities for more remote communities to participate.	Workers return to their community, share learning and resources. Workers who bring shame on their community are not allowed to return.	A strong relationship exists between the employer and the community (employer deals directly with the village). Employers empower leaders to watch over workers while in New Zealand.

To identify dimensions of value, it may be useful to consider two situations; first the problem the program aims to address, and second, the situation that is expected to result once the program has been implemented. Draw on appropriate sources to identify dimensions of value for each cell, such as one or more of the following:

- *Affected stakeholders*
- *Content experts*
- *Cultural standards*
- *Cultural experts*
- *Ethical guidelines*
- *Evaluation commissioner*
- *Evidence, e.g. previous evaluations or research*
- *Financial directive*
- *Government strategy*
- *International agreement*
- *Legislation*
- *Literature on good practice*
- *Lobby groups*
- *Needs assessment*
- *Objectives*
- *Policy documents*
- *Professional standards*
- *Regulations*
- *You (if you're an expert)*
- *Other?*

As you work through the matrix, you may identify value perspectives that link to others. For example, if one interest group has a right, that probably means another has a corresponding duty.

Don't be too concerned about blank cells. Not every principle is applicable for every interest group. More important is that you systematically consider each interest group with respect to each principle.

Justifying the criteria

The final step is to justify the criteria against which you will be making your evaluative judgement/s. Using the example of the seasonal labour initiative, we need to provide good reasons for defining 'effectiveness' in the way we have. The

importance of this step is illustrated through the following imaginary dialogue between an evaluator and a key stakeholder:

Evaluator	The selection process was a great success! (Judgement)
Stakeholder	How did you get to that conclusion?
Evaluator	Well, for one thing, good relationships were developed with communities, and workers felt they got to know their employer before they started work. (Evidence/factual premise)
Stakeholder	I admit your data shows this is the case. But why is having a relationship with communities and workers relevant to the success of the selection process?
Evaluator	…? (Reasoning about values)

To answer this question satisfactorily, the evaluator needs to give a good reason (or reasons) why 'relationship' should be included as part of a criterion on effectiveness.

It is important to include this reasoning in your evaluation report. As well as supporting the defensibility of your evaluative conclusions, it helps to make the criteria explicit. Information about the development of your criteria should also be included in the methodology section of your report.

Constructing and defining criteria

The information in the completed matrix is now ready to be analysed and developed into criteria and sub-criteria. For example, an analysis of the value dimensions from Box 3 has resulted in a high-level criterion of 'effectiveness of selection and pre-departure' with three sub-dimensions: selection, information and relationship, each of which can be further defined based on information from the matrix. These are outlined in Box 4.

Mathea Roorda #579498

Box 4: Criterion of 'effectiveness' (selection and pre-departure of workers)

	Effectiveness of selection and pre-departure of workers defined as:
Selection	No barriers for eligible citizens in remote locations to participate in the selection process (including access to health checks, passports, interviews, not charged a recruitment fee).
Information	Pre-departure orientation includes information about expectations of workers in New Zealand. Leaders support new workers. Information is culturally appropriate.
Relationships	Workers meet employer before they arrive in New Zealand. Workers know the name and location of the employer they will be working for. Employer develops a relationship with the workers' communities.

Next steps

At this point, you may want to develop standards for each criterion, using a rubric that includes a spectrum of goodness for each criterion (i.e. how the program components will be labelled in levels of quality) and descriptors, or examples of what performance will look like for each criterion at each performance level (Martens 2018).

Appendix B includes the components of a rubric, for your information. However, as this aspect of criteria development is not being field tested, the handbook is confined to identifying and developing a set of defensible criteria.

References

Davidson, E. J. 2005. *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Martens, K. S. 2018. "How Program Evaluators Use and Learn to Use Rubrics to Make Evaluative Reasoning Explicit." *Evaluation and Program Planning 69*(14):25–32.

Mepham, B., M. Kaiser, E. Thorstensen, S. Tomkins, and K. Millar. 2006. *Ethical Matrix Manual*. The Hague: LEI.

Newman, D. L. and R. D. Brown. 1996. *Applied Ethics for Program Evaluation*.

Thousand Oaks, CA: Sage Publications, Inc.

Roorda, M.B. (2019). Developing defensible criteria for public sector evaluations. Draft PhD thesis.

Appendix A (handbook): Fieldworker feedback

Thank you for field testing the Criteria Matrix. This form is for recording your feedback based on your use of the tool to develop criteria for an evaluation. Please complete and return to Mathea (**mathea.roorda@unimelb.edu.au**) by end September 2018. This information will form the basis of a follow-up interview.

1. What is the name of the program for which you have developed criteria while field testing this tool?

2. What was especially helpful about the criteria matrix tool?

3. What did you find confusing or especially difficult to apply when using the tool?

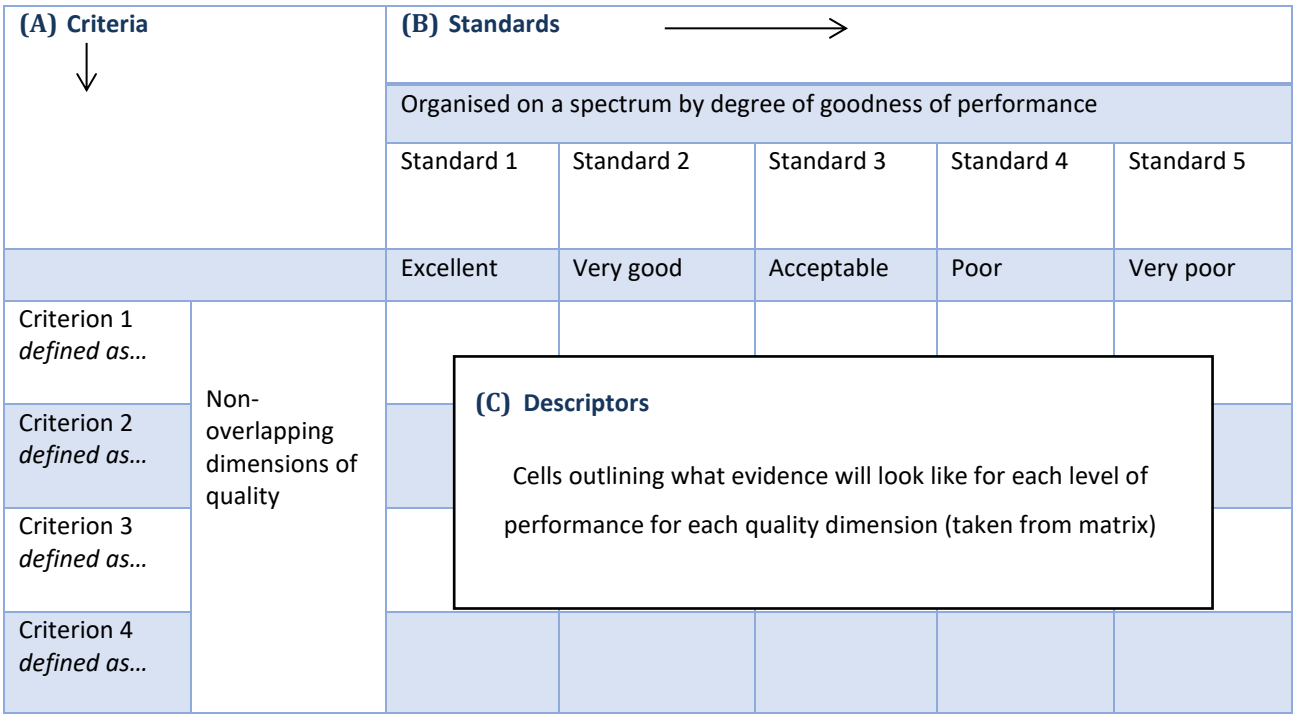
4. What would you add, modify, or remove?

5. If using the tool affected how you developed the criteria for your evaluation, please describe how it influenced you?

Thank you for your feedback!

Please return your feedback, along with a copy of the completed matrix and documents related to developing and justifying criteria for your program.

Appendix B (handbook): Components of a rubric



(Based on Martens 2018, p. 26)

What is the difference between the criteria matrix and the rubric?

Criteria matrix: The criteria matrix is a tool designed to help evaluators systematically *identify* and *justify* criteria of merit for a program evaluation. The aim is to develop a defensible set of evaluation criteria that incorporate all relevant dimensions of value.

Rubric: The rubric takes the relevant criteria (component A) and adds standards (component B) and descriptors for each criterion (component C). The descriptors in a rubric are examples of what performance looks like for each criterion at each performance level (e.g. excellent, very good, adequate, poor, very poor). The performance for each criterion can be stated quantitatively, qualitatively or in combination. For further information about developing an evaluation rubric, check out Davidson, (2005) or https://www.betterevaluation.org/en/resources/guides/rubric_revolution

Appendix C (handbook): Preparation for a criteria development workshop

Use this document as a guide to prepare for your workshop.

Recruitment of participants

- Identify key interest groups whose value perspectives need to be included.
- Recruit participants based on their experience and/or expertise.
- Consider sending a briefing document about the program to participants ahead of the workshop.

Preparation of matrix

- Define each principle using language that makes sense for your program and participants.

Opening and introductions

- Overview of the program and the workshop
- Note which interest group representatives are not in the room. You may need to encourage participants to step beyond their individual or representative interests to explore underlying value perspectives in the context of other interest group perspectives, that is, to place themselves in the shoes of others.

Identifying dimensions of value

- Give participants a copy of the matrix with example questions relevant to each principle. E.g. for consequences: 'if the program is benefiting [interest group] what would this look like?' or 'If X is the current situation, what do you expect to see if the program is working effectively'?
- The discussion should be documented by a member of the evaluation team. This may be done by typing statements and questions raised in the

respective cells of a blank ethical matrix and projected onto a screen for the participants to see, discuss or amend during the discussion.

- Ask participants to identify sources that support their statements, or back up / justify their statements, e.g. 'Why is that important?'