Author/s:
King, Julian Challis

Title:
Evaluation and value for money: development of an approach using explicit evaluative reasoning

Date:
2019

Persistent Link:
https://hdl.handle.net/11343/225766

Evaluation and Value for Money

Development of an approach using explicit evaluative reasoning

PhD Dissertation, July 2019

Julian King #602732

Title Page

Title of Thesis: **Evaluation and Value for Money: Development of an approach using explicit evaluative reasoning**

Author: Julian Challis King

Student ID number: 602732

ORCID ID: 0000-0002-8630-1423

Degree: Doctor of Philosophy

Date of Submission: 13 July, 2019

Department: Centre for Program Evaluation, Melbourne Graduate School of Education, The University of Melbourne

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

Abstract

There is increasing scrutiny on social investments to determine whether they deliver value for money (VFM), but current approaches to assessing VFM are incomplete. The disciplines of economics and evaluation share an interest in valuing resource use, but tend to operate as complementary or rival disciplines rather than being integrated within an overarching logic. Cost-benefit analysis (CBA) is often regarded as the gold standard for evaluating VFM, but has recognised limitations. For example, collective values, distributive justice, power dynamics, public dialogue, and qualitative evidence are peripheral to the method. Conversely, program evaluation offers more capacious approaches to determining value but rarely includes costs, let alone reconciling value added with value consumed. This disciplinary divide may diminish capacity for good resource allocation decisions.

The aim of this theory-building research was to develop a model to guide the evaluation of VFM in social policies and programs. A conceptual model was developed through critical analysis of literature, proposing requirements for good evaluation of VFM. Gap analysis was conducted to determine the extent to which CBA can meet the requirements of the conceptual model. Cumulative findings from the first two studies were dissected into a series of theoretical propositions. A process model was developed, identifying a series of steps that should be followed to operationalise the conceptual model. Case studies of real-world VFM evaluations in two international development programs were analysed to assess the conceptual quality of the theoretical propositions.

This research makes seven significant and novel contributions to the field of evaluation. First, VFM is an evaluative question, demanding a judgement based on logical argument and evidence. Second, VFM is a shared domain of two disciplines, because it is concerned with merit, worth and significance (the domain of evaluation) and resource allocation (the domain of economics). Third, CBA is not a rival to evaluation; it is evaluation. It evaluates an important dimension of VFM (aggregate wellbeing) and can strengthen the validity of an evaluation. Fourth, CBA is not the whole evaluation; it is usually insufficient on its own because of limitations in its scope and warrants. Fifth, a stronger approach involves explicit evaluative reasoning, with methods tailored to context including judicious use of economic methods where feasible and appropriate. Sixth, program evaluation standards should guide economic evaluation, and this has implications for the way CBA is used including the nature and extent of stakeholder involvement, the use of CBA in conjunction with other methods, and decisions about when not to use CBA. Seventh, the case studies are themselves a contribution, modelling the use of probative inference to corroborate the propositions of the conceptual model. Ultimately, this thesis provides proof of concept for a practical theory to guide evaluation of VFM in social policies and programs.

Declaration

This thesis comprises only the original work of Julian Challis King toward the degree of Doctor of Philosophy.

Due acknowledgement has been made in the text to all other material used.

The thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices.

Signed

Julian Challis King

13 July 2019

Acknowledgements

Completing this research is a big deal. Its time line spans a quarter of my career. When it began, my children were juniors at high school; now they are university students. Throughout those years, this research occupied my mind and significant chunks of time. Throughout those years, my wife Carolyn, and my two daughters, Alexandra and Victoria, remained unwaveringly supportive.

The catalyst for this research was a chance conversation with Professor Patricia Rogers who, in the space of a few minutes, awoke an ambition I didn't know I had, and helped me to realise that this was a path I should take. I remain ever thankful to Patricia for that conversation.

I am indebted to my supervisors, Professor Janet Clinton, Laureate Professor John Hattie, and Dr. Ghislain Arbour, who artfully helped me navigate the perplexing maze of postgraduate research, always knowing the right times to offer wisdom, debate, moral support, and space to ruminate. I also thank my Research Committee, Dr. Diane Mulcahy (Chair), and Professor Brian Yates for their support.

My professional peer group makes me want to be a better evaluator. There are too many to name here but in particular I want to acknowledge my Kinnect Group colleagues, Kate McKegg, Judy Oakden, and Nan Wehipeihana. I also acknowledge Dr. E. Jane Davidson as a significant influencer and mentor. Many thanks also to Professor Brian Yates and Dr. John Gargani for reviewing drafts of this thesis and providing valuable feedback.

I have been fortunate that this research has brought me into contact with people and programs that have, in turn, enriched the research. In 2015 I presented a prototype model for evaluation and value for money at the American Evaluation Association Conference in Chicago. In the audience was Luize Guimaraes, who at the time was establishing an innovative new female economic empowerment program in Mozambique. Luize saw the potential in the model and invited me to work with her team, and this in turn led to a series of further collaborations with Oxford Policy Management (OPM). The importance of these collaborations to this research is evident in the case studies presented in this thesis. I am thankful to my colleagues at OPM, the MUVA team in Mozambique, the Sub-National Governance team in Pakistan, and the UK Department for International Development, without whom the two case studies would not have been possible.

Contents

Chapter 1: Introduction 8

Chapter 2: Background 13

Evaluation 13
Evaluative reasoning 16
Economic methods of evaluation 22
Social policies and programs – an 'investment' view 31
Value for money 33
An interdisciplinary gap – and an opportunity 38
Summary 40

Chapter 3: Method 42

Methods 42
Staged approach 51
RQ1: Conceptual model 54
RQ2: Gap analysis 56
RQ3: Process model 57
RQ4: Case studies 58
Ethics 66

Chapter 4: Conceptual model 67

Introduction 67
Findings 67
Discussion 81

Chapter 5: Gap analysis 84

Introduction 84
Findings 85
Summary 123
Theoretical propositions for a 'Value for Investment' model 125

Chapter 6: Process model 132

Introduction 132
Results 135
Summary 142

Chapter 7: Case studies 143

Introduction 143
Case study 1: MUVA female economic empowerment program 144
Case study 2: The Pakistan Sub-National Governance Program 170
Replication 189

Overview of findings 195
Summary 199

Chapter 8: Discussion 201

Summary of findings in this thesis 201
Is it a good theory? 205
Research opportunities 208
Conclusions 209
Concluding remarks 212

References 214

Tables

Table 1: Sources of documentary evidence 63
Table 2: Rubric for coding propositions in case studies 64
Table 3: Economic methods as evaluative reasoning 91
Table 4: Working logics of economic methods 94
Table 5: Factors influencing selection of valuing methods 111
Table 6: Assessment of CBA against program evaluation standards 119
Table 7: MUVA VFM Criteria from original VFM framework 152
Table 8: MUVA VFM standards from original VFM framework 154
Table 9: MUVA VFM Criteria, Mk 2 158
Table 10: MUVA VFM standards, Mk 2 160
Table 11: Rubric for assessing effectiveness and scalability 161
Table 12: VfM criteria used in the SNG Program 177
Table 13: VfM Standards used in the SNG Program 178
Table 14: Replication analysis 193

Third party copyright material included in the thesis

1. King, J. (2017). Using Economic Methods Evaluatively. *American Journal of Evaluation, 38*(1), pp. 101-113. Reproduced with permission.
2. Figure 2: VFM conceptual framework. Reprinted from King & OPM (2018). Reprinted with permission.
3. Figure 5: MUVA program theory of change. Reprinted from OPM (2016). Reprinted with permission.
4. Figure 6: Sub-national governance program theory of change. Reprinted from OPM (2016). Reprinted with permission.

Fileref: PhD thesis Julian King July 2019.docx
Last saved: 13-Jul-19

Chapter 1: Introduction

This research began with a hunch. The author's career pathway started in public policy analysis, branched into economic evaluation of health care interventions, and then broadened into policy and program evaluation more generally. Program evaluators skilled in economic evaluation are relatively rare, and the author was regularly sought out to provide assessments of something called 'value for money' (VFM).

This raised an interesting set of conundrums. VFM was a familiar term in public policy, but an amorphous concept. When pressed to explain what they meant by VFM, people's responses would vary from 'are we getting enough outcomes to justify what we're spending?' to 'just reassure us we're not tipping money down the drain'. Even official governmental publications defined the term in diverse ways, ranging from sound practices to ensure reasonable prices are paid for quality inputs, to aspirational statements about hard-to-measure concepts such as "maximising the impact of each pound spent to improve poor people's lives" (DFID, 2011, p. 2). Could there be a unifying definition of VFM? Could VFM be defined in such a way that it expresses something meaningful, important, and evaluable?

Views differed, too, on how VFM should be evaluated. Some people held a conviction that the only way to really tell if something provides VFM is to conduct a randomised controlled trial (RCT; to determine whether the program causes outcomes) followed by a cost-benefit analysis (CBA; to determine whether the benefits of the program, valued monetarily, exceed the costs). Others, however, questioned the applicability of this rigid methodological stance to social policies and programs. They pointed out that imposing these methods might limit the circumstances in which it is feasible to assess VFM, and could limit the sorts of outcomes that would be included in the analysis. For example, future savings to the public purse may be easier to estimate than a 'cultural return on investment' where the primary objective of a policy is to support the sustainability of indigenous language and culture. Some went so far as to reject CBA entirely, arguing that the most important values in a social policy or program are intangible and either cannot or should not be valued monetarily.

Neither economic evaluation nor other evaluation methods seemed able to provide a satisfactory answer to a VFM question, but they appeared to bring complementary insights. Economic evaluation can provide valid and useful indicators of efficiency, but has recognised limitations in circumstances where outcomes, processes or equity trade-offs are difficult to value in the required metrics, or in circumstances requiring balancing of conflicting values. Conversely, program evaluation offers a more diverse set of approaches to determining value, but not how these might be used to compare value created with value consumed. This disciplinary divide may diminish capacity for good resource allocation decisions in social policies and programs.

In international development, for example, aid programs are routinely subjected to process and outcome evaluations which do not examine costs. Economic evaluation is often applied prospectively in assessing a business case for a new program but is rarely revisited during program delivery. In practice, VFM is often assessed on the basis of a narrow set of indicators, devoid of explicit evaluative reasoning. There is a risk that such assessment could focus on activities that are easy to measure but unimportant, or on the quantification of outputs and outcomes at the expense of more nuanced consideration of their quality and value. Could evaluation of VFM be enhanced by reaching across disciplinary boundaries to combine evaluative reasoning with economic methods of evaluation?

An explicit theoretical foundation did not exist to explain whether it is valid to combine the theory and practice of program evaluation and economic evaluation, nor to guide when and how this should be done. An opportunity was identified for theory-building research to fill this gap in the existing body of evaluation theory.

Accordingly, the aim of this research is to develop a model to guide the combined use of evaluative reasoning together with economic evaluation, to assess VFM in social policies and programs. The research commences with a review of existing theory and knowledge to develop a conceptual model, identifying the features of good VFM evaluation of social policies and programs. The purported gold standard, CBA, is then systematically assessed against the conceptual model to expose its strengths, limitations, and potential use within the model. A process model is developed, guiding the application of the conceptual model in practice. The model is applied in real-world evaluations of VFM, and these evaluations are analysed as case studies to learn whether the theory works in practice. The structure of the thesis mirrors these research steps, as follows.

Overview of thesis

Background (Chapter 2) introduces the key spheres of theory and practice that are within the scope of the thesis, namely evaluative reasoning and economic evaluation. The topic of VFM is explored. It is found that there are a multitude of definitions of VFM. A unifying definition is proposed. An interdisciplinary opportunity is identified where economic and other valuing methods should be better integrated. It is argued that a theoretical foundation should be developed to address this opportunity.

Methods (chapter 3) describes the overarching design and methods used in the research. A deductive approach has been taken, involving a cumulative process of theory building and empirical investigation. The approach is multi-method, including critical analysis of literature to develop a model, and case studies to investigate its conceptual quality. A series of four investigative

phases, and their corresponding methods, are explained. Each phase of the research is reported sequentially in the subsequent four chapters, as follows.

Conceptual model (chapter 4) sets out a logical argument, developed through critical review and synthesis of literature, about what VFM means and how it should be evaluated. The proposed conceptual model identifies requirements for good VFM evaluation. This model provides a foundation for the subsequent study.

Gap analysis (chapter 5) systematically compares the methodological prescription for CBA against the requirements of the conceptual model, to determine the extent to which CBA is theoretically able to meet the requirements of the model, and the manner and circumstances in which CBA might enhance the validity of the model. The chapter concludes by defining a series of theoretical propositions. These propositions, summarising the cumulative findings of the first two studies, provide the analytical framework for the case studies.

Process model (chapter 6) translates the theoretical requirements of the conceptual model into a set of steps for planning and undertaking an evaluation of VFM. The process model is a prototype, designed for testing and refinement in a specific context: international development programs where VFM assessments are mandated by the United Kingdom's Department for International Development (DFID).

Case studies (chapter 7) provide intensive analysis of two VFM evaluations conducted with fidelity to the model: a female economic empowerment program in Mozambique, and a governance reform program in Pakistan. Through documentary analysis, the two cases are described, providing concrete illustrations of the application of the model in practice. The facts of each case are used to systematically assess the conceptual quality of the theoretical propositions. Findings from the two case studies are triangulated to investigate the extent to which the findings are replicated. Observations and experiential learning from the case studies inform refinements to the model.

Discussion (chapter 8) provides a summary of results from the research and what they contribute to the field of evaluation. Opportunities for future research are identified. The chapter concludes with remarks about the nature of VFM and the requirement for sound evaluative reasoning, with methods such as CBA being used in the service of robust evidence and logical argument.

Summary of findings

The research makes seven significant and novel contributions to the field of evaluation. First, VFM is an important construct that is broader than efficiency or return on investment. It is an evaluative question about how

well resources are used and whether a resource use is justified. It demands a judgement based on logical argument and evidence.

Second, VFM is a shared domain of two disciplines – evaluation and economics. It is an evaluative question (concerned with merit, worth and significance) about an economic problem (resource allocation). Evaluation and economics can and should be combined to address questions about VFM.

Third, CBA is not a rival to evaluation – it is evaluation, in a literal sense, because it conforms to the general logic of evaluation. Further, CBA estimates something important to VFM – net benefit or aggregate welfare – and it does so in ways that can enhance the validity of evaluation. Evaluators should use CBA more.

Fourth, CBA is not the whole evaluation. CBA evaluates efficiency, and efficiency is only one possible criterion of VFM. Additional criteria such as relevance, sustainability, and distributive justice are also relevant to VFM in social investments. Multiple forms of evidence and knowledge creation should contribute to evaluative judgements about complex social issues. Given the centrality of issues like these in social policies and programs, CBA will usually be insufficient on its own. CBA should be more widely used in evaluation, but not as a stand-alone method. CBA should have a supporting role, contributing part of the evidence toward an evaluation of VFM.

Fifth, evaluative reasoning is essential to making judgements about VFM. This research highlights the worth of evaluation as a discipline, and evaluative reasoning as the backbone of evaluation. As much as evaluators should use economic analysis, economists should make their evaluative reasoning explicit to ensure the validity of CBA design, methods and findings. Evaluative judgements should not be subbed out to a formula, but CBA can supply valuable insights to support sound judgements.

Sixth, program evaluation standards should guide economic evaluation. Any evaluation that stands to affect the welfare of people should be open to scrutiny to determine whether it meets scientific and ethical obligations. Although there is no definitive checklist for such a purpose, a number of existing and widely used program evaluation standards offer a basis for judging the quality of an evaluation. Such standards are the culmination of debate and formalise some degree of consensus about evaluation as a field of practice. Commonly agreed features of high quality evaluations are that they should be useful, practical, ethical, accurate, and accountable. Applying these principles to CBA requires the evaluator to remain open to the possibility of not conducting an economic evaluation.

Seventh, the case studies conducted in this research are a contribution to the evaluation field. The case study analysis models the use of probative inference (inference to the best classification) which underpins the general logic of evaluation. Theoretical propositions were identified that differentiated

the strengths and limitations of CBA relative to the features of a proposed model of evaluation involving explicit evaluative reasoning. These propositions, together with evidence from the case studies, underpinned a process of reasoning to evaluate the conceptual validity of the proposed model. The theoretical propositions could be used again in future research and meta-evaluation of the model. Moreover, the general approach of developing such propositions about the nature of evaluation, and testing them empirically through case studies, can be used in other research on evaluation.

Investigation of the model through case studies provides proof of concept. When it comes to investments in social change, VFM is often as concerned with criteria such as social justice, equity and fairness as it is with economic efficiency. Economic methods of evaluation can enhance evaluation of VFM but are insufficient on their own. A stronger approach involves explicit evaluative reasoning, with methods tailored to context including judicious use of economic methods where feasible and appropriate. Such a model can incorporate the strengths of economic evaluation without being limited to economic criteria and metrics alone.

Chapter 2: Background

This chapter introduces the key spheres of theory and practice that are within the scope of this thesis. Gaps in knowledge are identified that lead to the research questions.

Evaluation, the host discipline for the thesis, is introduced. In particular, the domain of values and valuing, and the manner in which explicitly evaluative conclusions are reached from empirical evidence, are central to this thesis. This form of reasoning, labelled evaluative reasoning, has transdisciplinary application and is at the heart of what it means to evaluate. Economic evaluation is introduced, a powerful set of methods for comparing alternative courses of action with regard to their costs and consequences. Arguably, these methods are under-utilised by program evaluators.

The topic of value for money (VFM) is explored. Analysis of literature reveals a multitude of definitions of VFM as well as some striking commonalities among those definitions. An all-encompassing definition is proposed, suggesting that VFM sits at an intersection between evaluation and economics. It is argued that there may be advantages to combining economic evaluation with other methods and forms of evaluative reasoning, and that this possibility warrants theoretical and empirical investigation.

Evaluation

Evaluation has been defined as "the process of determining the merit, worth or significance of something" (Scriven, 1991, p. 139). Merit, worth and significance are common terms in evaluation, used to describe what matters to people – also known as *value*. In the broadest sense, all human beings are evaluators – indeed, the capacity to evaluate is an essential factor underpinning technological and social progress (Davidson, 2005). In everyday life, people evaluate things on the basis of many considerations, sometimes explicit but often implicit. Professional evaluation seeks to identify and make explicit relevant considerations and thus to systematically determine value (Davidson, 2005).

Evaluation involves making judgements:

> *The distinguishing feature of evaluation as a practice is that it directly engages questions of the value of programs and policies. Evaluation is a judgment-oriented practice – it does not aim simply to describe some state of affairs but to offer a considered and reasoned judgment about the value of that state of affairs.* (Schwandt, 2015, p. 47).

This view of evaluation – in which judging value is central – is distinct from an alternative view of evaluation as applied research, oriented to the use of social science methods to establish patterns of causality (Schwandt, 2015).

This thesis holds the position that a judgement-oriented view of evaluation is what defines and distinguishes evaluation from other fields of inquiry (Davidson, 2005). The alternate view is relevant, in that social science research designs and methods are often employed in evaluation to gather empirical evidence and make causal inferences. But it omits the fundamentally evaluative purpose of gathering the evidence. Greene (2007) argued that "methodological decisions are always made in service to substance" (p. 114). This thesis views the 'substance' of evaluation as the reasoning used to establish evaluative conclusions and holds that the process of "developing, strengthening, and clarifying reasoning that leads to legitimate evaluative conclusions" is central to evaluation theory and practice (Fournier, 1995, p. 15).

The discipline of evaluation can be viewed as a transdiscipline, just as statistics and logic are transdisciplines, because it applies "across broad ranges of the human investigative and creative effort while maintaining the autonomy of a discipline" (Scriven, 1991, p. 1). Evaluation is an intellectual process "that technology and science share with all other disciplines, with the crafts, and with rational thought in general" (Scriven, 1991, p. 4). Evaluation is necessary in order for these disciplines to function – leading to an argument that evaluation could even be viewed as "the alpha discipline", overseeing how well evaluation is performed in all other disciplines (Scriven, 2015, p. 19).

Program and policy evaluations – the overarching domain for this thesis – are undertaken by and for governments, non-profit organisations, philanthropists and think tanks, to assess the value of social programs and policies. Evaluators make judgements about various aspects of programs and policies, such as their: quality of implementation; goal attainment; effectiveness; outcomes; impact; and costs (Schwandt, 2015). This research focuses on evaluation of social programs and policies, with case studies further narrowing the focus to international development programs – though the research may have broader relevance.

This thesis builds on the notions that no one method is perfect and that multiple methods can be used to answer an evaluative question (Greene, 2005; 2013; Mertens & Hesse-Biber, 2013). Such approaches afford the evaluator scope to balance the relative virtues of different methods, for example, independent observation and participatory approaches. It allows evaluation to be positioned as a contextually responsive practice (Patton, 2011; Schwandt, 2015). With this flexibility, however, comes a responsibility to defend the validity of the choices of evaluator orientation, evaluation design and methods, in addition to the validity of the evidence gathered and judgements made (Griffith & Montrosse-Moorhead, 2014).

Validity, in contemporary evaluation theory and practice, extends beyond technical considerations such as accuracy and reliability to questions of

'validity to whom?'. House (1980), for example, argued that validity includes beauty (placing the evidence within an evaluation story that resonates with stakeholders), truth (evidence that makes sense and is acceptable to stakeholders), and justice (e.g., ensuring the voices of marginalised stakeholders are heard) (Montrosse-Moorhead, Griffith, & Pokorny, 2014). This notion of validity also extends to multicultural validity (Kirkhart, 2010), recognising the interplay between cultural worldviews, values and ways of knowing (Deane & Harré, 2016; Wehipeihana & McKegg, 2018).

This situationally responsive, mixed methods model of evaluation reflects an increasingly sophisticated view of complexity, recognising that "many aspects of economic and social development are complex, unpredictable, and ultimately uncontrollable" (Archibald, Sharrock, Buckley, and Young, 2018, p. 74). The theory and practice of evaluation under this orientation requires evaluators to respond to complex and evolving situations, in contextually sensitive ways. This cannot be entirely based on algorithms or rules; rather, rigorous evaluation requires evaluative thinking – a mix of "critical thinking, creative thinking, inferential thinking and practical thinking" (Patton, 2018, p. 21). Evaluative thinking includes evaluative reasoning (detailed later) but is conceptually broader, reflecting:

> *. . . the ability to creatively arrive at evaluation-specific solutions by determining what combinations of methods, tools, techniques, and theories are appropriate in light of contextual particulars. Making reasoned, evidence-based choices set critical thinking and evaluative thinking apart from judgements based only on deeply held and unchallenged beliefs (e.g., stereotypes). Making a reasoned choice about value and being able to defend it is what distinguishes evaluative thinking from critical thinking* (Vo, Schreiber, & Martin, 2018, p. 40).

Evaluators determine the value of policies and programs for a practical purpose – informing decision making – and with a view to providing findings that are useful (Julnes, 2012c; Schwandt, 2015). Undertaking an evaluation does not, however, guarantee its usefulness or its use. Evaluators have long sought to address the gap between knowledge and action (Alkin & King, 2016; Patton, 2008). Evaluation use extends beyond instrumental use of findings to inform decisions and includes, for example, process-based use through evaluation capacity building and learning (Patton, 2011).

An example of a situationally responsive model aimed at facilitating evaluation use is utilisation-focused evaluation: "evaluation done for and with specific primary intended users for specific, intended uses" (Patton, 2011, p. 13). This model is predicated on the idea that good evaluation is useful and gets used, that intended users are more likely to use evaluations if they have been actively involved in them, and that evaluators should therefore facilitate specific and concrete end-use through the design and conduct of the evaluation (Patton, 2008). Utilisation-focused evaluation requires the

evaluator to form a working relationship with intended users, and to negotiate an evaluation design and methods that meet users' needs while also meeting professional evaluation standards such as accuracy, feasibility, propriety, and being sensitive to the diversity of values and interests of affected parties (Yarbrough, Shulha, Hopson, & Caruthers, 2011).

In summary, this thesis builds on premises that program and policy evaluation "exists to inform real-world decisions" (Julnes, 2012b, p. 3), that it needs to be useful and actually used (Patton, 2008), that judgements of the value of policies and programs "can and should be made on the basis of evidence and argument" (Schwandt, 2015, p. 46), and these judgements need to be warranted – that is, valid and acceptable to relevant stakeholders such as program architects, funders, deliverers, and affected communities (Fournier, 1995). In order to meet these aims, evaluation needs to be responsive to complex and evolving contexts, requiring evaluative thinking (Vo & Archibald, 2018) and accommodating multiple paradigms and methods (Mertens & Hesse-Biber, 2013).

Central to evaluative thinking, evaluation validity, and evaluation use, is an understanding of valuing – how evaluators systematically incorporate values held by individuals and groups into their work, and how they use values to reach evaluative conclusions (Gargani, 2018). There are multiple approaches to valuing. The following section provides a general overview before focusing on two approaches that feature in this research: qualitative valuing and synthesis, and economic evaluation. This thesis explores the relative merits and potential compatibility of these two sets of approaches in particular.

Evaluative reasoning

This chapter began with an observation that evaluation involves making judgements about merit, worth and significance. The "fundamental problem" in this endeavour is "how one can get from scientifically supported premises to evaluative conclusions" (Scriven, 1991, p. 51). *Evaluative reasoning* is the solution to this fundamental problem, and is the process by which an evaluator uses criteria of merit, worth and significance to draw explicitly evaluative inferences from evidence (Davidson, 2005; House & Howe, 1999; Yarbrough et al., 2011).

Criteria of merit, worth and significance are "aspects, qualities or dimensions that make it possible to determine that a program or policy is good, poor, successful, unsuccessful, better or worse than some alternative, effective, ineffective, worth its costs, morally defensible, and so on" (Schwandt, 2015, p. 48). "Most common concepts", Scriven (2007) argued, "are 'cluster concepts' that are learnt, explained, and defined in terms of a cluster of properties" (p. 5). For evaluation purposes, a 'good' policy or program can be

defined criterially, and evidence can be gathered to determine the extent to which the desired cluster of properties is apparent.

Criteria of merit, worth and significance are determined on the basis of *values* – "principles, attributes, or qualities held to be instrinsically good, desirable, important, and of general worth" (Stufflebeam, 2001b, p. 1) or "normative beliefs about how things should be, strong preferences for particular outcomes, or principles that individuals and groups hold to be desirable, good or worthy" (Schwandt, 2015, p. 48). Values, in other words are an expression of what matters to people.

Scriven (2012) argued that values, as the basis of criteria, can be validated "by observation, inference or definition" – making it possible "to infer beyond reasonable doubt to evaluative conclusions" (p. 17). This argument rests on an objectivist view of merit, worth and significance as being "real, although logically complex, properties of everyday things" which can be determined to "an acceptable degree of objectivity and comprehensiveness" (Scriven, 1993, p. 9).

The use of criteria to reach evaluative conclusions has been termed *probative inference* – that is, inference that is not deductive or inductive, but rather inference that has the "quality or function of proving or demonstrating something" (Scriven, 2012, p. 22) or "an inference to a conclusion that should be believed until disproved" (Scriven, 1994, p. 373). The *General Logic of Evaluation* (Scriven, 1980; 1991; 1994; 1995; 2012) offers an overarching logic that describes this process. Just as disciplines like law, medicine, and science use specific patterns of reasoning or basic logic to guide and inform judgements, the general logic of evaluation provides evaluators with "the rules for constructing and testing claims, and. . . the basic conditions under which rationally motivated argumentation can take place" (Fournier, 1995, p. 16).

Scriven was not the first theorist ever to describe this logic, but was first to articulate it in the evaluation literature. Nunns, Peace and Whitten (2015, p. 138) noted:

> *Up until the mid-20th century, the rules of formal logic made it logically impossible to reason from a factual premise to an evaluative claim (Scriven, 2013). Developments in informal logic such as those articulated by Hare (1967), Rescher (1969), and Taylor (1961) meant reasoning about values became logically possible.*

Evaluation theorists diverge in their application of the general logic of evaluation. In particular, "there is little consensus in the field of evaluation about how to aggregate findings across multiple criteria or across stakeholders' differing perspectives on important criteria" (Schwandt, 2015, p. 59). For example, the extent to which a systematic criterial approach

might guide, complement, or supplant more intuitive judgement is debated (House, 1980; Stake et al., 1997).

Schwandt (2015) distinguished four general approaches to evaluative reasoning: technocratic, tacit, all-things-considered, and deliberative. The technocratic approaches are empirically-based and systematic, and span rule-governed, algorithmic and rubric-based approaches, which include both quantitative and qualitative approaches to weighting and synthesis. Tacit approaches are intuitive, holistic and based around a narrative construction, in contrast to all the other approaches to synthesis, which are "arguments structured around premises that logically lead to conclusions" (p. 61). All-things-considered approaches involve comparing and weighing reasons for and against a particular conclusion in order to reach a judgement. Deliberative approaches are a way of reaching an all-things-considered judgement through a process of collective public reasoning (House & Howe, 1999; Schwandt, 2015). In some evaluation circumstances, little or no synthesis is attempted; as an alternative to synthesis, a policy or program's performance is described for a range of criteria with no attempt to reach an overall conclusion (Schwandt, 2015).

This research focuses primarily on technocratic approaches to synthesis. The validity of other approaches to synthesis, whether as alternatives to technocratic approaches or as compatible processes, is not discounted, however (Stake & Schwandt, 2006). For example, a technocratic approach to synthesis could be supplemented with tacit, all-things-considered, or deliberative approaches as strategies for mutual checking, challenging and validation of conclusions – or, a technocratic approach could be invoked as a way of structuring a deliberative approach.

Technocratic approaches, for the purposes of this research, include quantitative and qualitative approaches to valuing and synthesis. These approaches involve establishing criteria of merit, worth and significance (the aspects of performance to be assessed), constructing standards (the level of performance that must be achieved), gathering and analysing evidence of performance against each criterion, and synthesising results across all criteria to make a judgement (Fournier, 1995). The synthesis process requires a determination of the relative importance of criteria, as well as taking into consideration the level of performance against each criterion – and it is in the approach to valuing and synthesis that the following approaches differ, with the former being based on a numerical scoring and weighting system, using cardinal numbers, and the latter being based on a qualitative system in which weighting and scoring is carried out ordinally. The two approaches to valuing and synthesis are summarised as follows.

Quantitative valuing and synthesis

Quantitative valuing and synthesis involves comparing options by reference to an explicit set of criteria. These criteria are generally assigned importance weights using an interval or ratio scale, and a scoring system is used to derive a weighted score for each criterion. The option with the highest overall score (sum of weighted scores) is the preferred option. This form of evaluative reasoning is referenced both in program evaluation literature, as *Numerical Weight and Sum (NWS)* (Scriven, 1991), and in economic literature, where it is called *Multi-Criteria Decision Analysis* (Dodgson, Spackman, Pearman & Phillips, 2009).

The most comprehensive version of multi-criteria decision analysis is based on multi-attribute utility theory (Von Neumann & Morgenstern, 1947), to which we return later, and decision analysis (Keeney & Raiffa, 1976) but the general approach also includes simple matrix-based approaches (Dodgson et al., 2009), which can be applied prospectively to support decisions, or retrospectively to evaluate the performance of alternatives.

Decision analysis is "a philosophy, articulated by a set of logical axioms, and a methodology and collection of systematic procedures, based upon those axioms, for responsibly analysing the complexities inherent in decision problems" (Keeney, 1982, p. 806). The axioms of Von Neumann and Morgenstern (1947) provide the foundations for decision analysis, proposing a rational model of decision making in which alternatives can be ranked based on the relative probabilities of their possible consequences and an individual's preference for those consequences. This model provides for the systematic appraisal of alternatives and for the inclusion of "judgments and values in an analysis of decision alternatives" (Keeney, 1982, p. 807).

A key feature of multi-criteria decision analysis is that it explicitly requires judgements from evaluators or decision makers – in establishing criteria, assigning importance weights, and judging performance in relation to each criterion (Dodgson et al., 2009). The approach is often applied by economists where complexities, often present in decision-making environments, present conceptual and practical difficulties to the conduct of a cost-benefit analysis (Dodgson et al., 2009). Examples of these complexities, which are commonplace in program evaluation, include: multiple objectives; difficulty identifying good alternatives; intangible factors (such as goodwill, morale, and aesthetics); long time horizons and the need to consider future implications of alternatives; many impacted groups with different attitudes and values; risk and uncertainty (e.g., stemming from lack of data, time and costs of acquiring data, unpredictable future events with significant consequences, changing priorities over time); risks to life and limb; the need for input from a number of relevant disciplines, with no overall experts; multiple decision makers; value trade-offs; risk attitudes; and the sequential nature of decisions (Keeney, 1982).

The use of numerical weights and scores brings a transparent and systematic approach to the valuing and synthesis process. It can work adequately, for ranking options, where there are relatively few criteria and there is a clear basis for setting weights (Davidson, 2005). However, Scriven (1991, p. 380) noted that conducting an evaluation through NWS has a number of limitations which mean that although it is "sometimes approximately correct, and nearly always clarifying" it can lead to fallacious conclusions.

The problems in NWS arise from: treatment of all values as continuous variables, whereas some may in fact involve minimum expectations or 'bars' (requiring a different mechanism to rule out noncomplying options); a spurious effect of having a high number of dimensions, in which multiple trivial considerations can swamp a few important criteria in the aggregate score; interactions between dimensions (such as double-counting); and the assumption of linear utility across the range of variables (whereas real-life human decision-making does not reflect this property) (Scriven, 1991; 1994). NWS also requires a basis for setting and justifying the selected weights, a condition that Scriven (1991, p. 293) argued can only be met through empirical curve-fitting. The upshot of these limitations is that NWS does not always work as well as intended.

> *The reality is that although NWS seems simple and intuitive, it can often leave the evaluation team looking at a conclusion that does not seem quite right. The temptation at that point is often to fiddle with the numbers to see whether the right answer can be coaxed out of the data. An alternative is to work with a synthesis strategy that incorporates the key elements of how the human brain naturally weights considerations, making them explicit so that they can be applied to larger numbers of dimensions* (Davidson, 2005, p. 177).

For these reasons, Scriven (1991) favoured the use of qualitative weighting and synthesis in the majority of evaluation circumstances, where the requirements for numerical approaches cannot be met – in particular, in contexts where criteria are determined by users and the relative weights assigned to different criteria are determined prescriptively.

Qualitative valuing and synthesis

Scriven (1991; 1994) argued that in order to avoid the fallacies of numerical weighting and synthesis, criteria should be weighted using an ordinal scale. Under this approach, the range of possible weights is simplified down to a few levels – for example, essential, very valuable, valuable, marginally valuable, and not valuable. Under this approach, it becomes simpler in practical terms to justify weights – for example, weights can be assigned based on criterial definitions or needs assessments (Scriven, 1994). "Beyond this very modest level", Scriven (1991) argued, "validity in allocating utility points is hard to justify" (p. 294). This approach also overcomes the

problems of bars (by defining minimum requirements) and swamping (because the weights are incommensurable; if a program receives high scores on three criteria and a low score on a fourth criterion, the three high scores cannot overpower the low one) (Scriven, 1994).

An evaluative rubric is a tool commonly used to summarise and define ordinal weights. A rubric "captures all the elements of the logic of evaluation as presented by Scriven and explicitly outlines the systematic reasoning called for by Fournier" (Martens, 2018b). Davidson (2005; 2014) pioneered the use of rubrics in program evaluation, building on their use in personnel evaluation and education (Martens, 2018a).

Rubrics can be used in two distinct ways in program evaluation: for the purpose of transforming, classifying or categorising data; and as an evaluation-specific methodology to support evaluators in "the process of combining evidence with values to determine merit, worth, or significance" (Martens, 2018a, p25). It is the second of these functions that is most relevant here.

An evaluative rubric comprises three components: criteria of merit, worth or significance; performance standards; and descriptors or "examples of what performance looks like for each criterion at each performance level" (Martens, 2018a, p. 26). These components are generally arranged in a matrix – for example, with criteria as row headings, standards as column headings, and descriptors within the body of the table.

The content of rubrics can be developed on the basis of documented evidence (for example, evidence of needs and past performance of similar interventions), documented expectations (such as policy documents and service contracts) and/or as a participatory activity with stakeholders (Martens, 2018b). The process of identifying criteria and standards can be intentionally used as an aid to fostering stakeholder engagement in evaluation and as a means to enhance situational responsiveness and evaluation use (Davidson, 2005; Dickinson & Adams, 2017).

For example, King, McKegg, Oakden, and Wehipeihana (2013) described their use of rubrics in evaluation practice, finding that rubrics could be used in ways that helped support participatory processes to surface different viewpoints and values from stakeholders. Rubrics provided a visual and intuitive representation of a shared (inter-subjective) set of values that provided the basis for making evaluative judgements. By involving stakeholders in rubric development collectively and incrementally, common ground was identified, and differences could be accommodated or acknowledged. This process was found to increase transparency about how evaluative judgements were made. It provided a focal point for negotiating a contextually appropriate mix of methods to provide evidence that would be credible to stakeholders. The rubrics were used as a critical point of reference

when facilitating stakeholder participation in weighing the evidence and reaching evaluative judgements. These processes increased the likelihood that findings would be accepted – facilitating evaluation credibility, utility, and use.

While a rubric-assisted approach to valuing and synthesis overcomes specific limitations of NWS, it is not without its own limitations and risks. Rubrics remain vulnerable to shared bias in the development of criteria and standards, and in the judgements made from the evidence – that is, either the rubrics, the ratings, or both, can be wrong (Scriven, 1994; Stake & Schwandt, 2006). Determining the characteristics of good criteria is an area of ongoing research (Roorda, 2019). Irrespective of the approaches and tools used to guide evaluative reasoning, evaluation is "not first and foremost about methods, but is about making sense of evidence and creating a coherent, logical, and, ultimately, if successful, persuasive argument about what the evidence shows" (Patton, 2018, p. 18).

Our attention now turns to a different and distinct set of approaches to evaluation and valuing. Existing alongside, and with little overt connection to the general logic that underpins program evaluation, are a set of economic evaluation methods. These methods tend to operate in a parallel universe to that of program evaluation, despite having a similar intent of informing sound decisions in the public interest, and despite their use, on occasion, in similar contexts to evaluate social policies and programs.

Economic methods of evaluation

Economics has been described as "the study of how people choose to use resources" (American Economic Association, 2013), or the "study of how societies use scarce resources to produce valuable commodities and distribute them among different people" (Samuelson, 1948). The term *economics* was first used by the Ancient Greek philosopher Xenophon, who combined *oikos* (household) with *nomos* (rules or norms), in reference to the art of household management (Raworth, 2017). Indeed, in the broadest sense, all human beings are economists. Resource scarcity, and our consequent need to ration resources, means that choices have to be made in all areas of human activity (Bills, 2013; Drummond et al., 2005).

Economics comprises a number of sub-disciplines. For example, distinctions are made between microeconomics (the study of "how households and firms make decisions and how those decisionmakers interact in the marketplace") (Mankiw, 1999, p. 12) and macroeconomics ("the study of the economy as a whole") (Mankiw, 1999, p. 538).

Welfare economics uses microeconomic techniques to study the distribution of resources within an economy. While much of economics is positive (seeking to reveal and describe how people actually use resources), welfare

economics has the capacity to also be used normatively (Drummond et al., 2005), proposing how decisions *should* be made and taking preferences into account when seeking to maximise "social welfare" (the overall wellbeing of society) through the allocation of resources. Approaches to welfare economics are diverse and their history complex. The focus of this research is restricted to the theory and practice of economic methods of evaluation – and, in particular, cost-benefit analysis (CBA). These methods, as we shall see in the gap analysis chapter, are forms of evaluative reasoning.

Economic methods of evaluation are a suite of methods for comparing the costs and consequences of alternative resource allocations (Drummond et al., 2005). Economic evaluation is fundamentally concerned with choices, necessitated by resource scarcity. As with evaluation, "these choices are made on the basis of many criteria, sometimes explicit but often implicit" (Drummond et al., 2005, p. 9). Unlike evaluation, "economic analysis seeks to identify and make explicit *one* [emphasis added] set of criteria that may be useful in deciding among different uses for scarce resources" (Drummond et al., 2005, p. 9): namely, efficiency.

Choosing to allocate resources to one program or policy instead of another – or to particular activities within a program or policy – entails opportunity cost. That is, the opportunity to allocate resources for alternative uses is foregone (Drummond et al., 2005). Accordingly, economic evaluation is typically formulated as a choice between competing alternatives. For example, a project may be evaluated against its existing next-best alternative, or a baseline of doing nothing. Usually, when comparing alternatives, only the difference in costs and consequences is considered. This is called an incremental analysis (Drummond et al., 2005).

Economic methods of evaluation are used to systematically identify, measure, value, and compare the costs and consequences of those alternatives (Drummond et al., 2005). All of the different economic methods yield indicators of efficiency, but vary in regard to their scope and the units of measurement used. Traditionally, they inform summative judgements – though they can be adapted for formative purposes (Yates, 1996). For the purposes of this summary, economic methods are grouped into three approaches: Cost-benefit analysis (CBA), cost-effectiveness analysis (CEA) and cost-utility analysis (CUA).

Cost-Benefit Analysis

For the purposes of this research, CBA refers to any approach to economic evaluation that values costs and consequences in the same units – typically, though not necessarily, monetary units (Yates, 1996).

CBA involves a systematic comparison between the incremental costs and the incremental consequences of an investment (such as a policy or program). In

simple terms, where benefits exceed costs, the investment can be deemed worthwhile on the basis that it provides a net benefit (Drummond et al., 2005).

Usually, costs and benefits are distributed across different points in time, and may include past, present and/or future costs and benefits. The value of a cost or benefit varies with time (e.g., an immediate benefit is more desirable and hence more valuable than the same benefit years in the future); therefore the analysis applies a *discount rate* to adjust all costs and benefits to their present day value (Levin & McEwan, 2001). This can be summarised algebraically as follows.

If there are i possible social investments (where $i = 1 \ldots I$), then the net present social benefit of project i ($NPSB_i$) is:

$$NPSB_i = \sum_{t=1}^{n} \frac{b_i(t) - c_i(t)}{(1+r)^{t-1}}$$

Where:

$b_i(t)$ = benefits, in monetary terms, in year t

$c_i(t)$ = costs, in monetary terms, in year t

$1/(1+r)$ = a discount factor at annual interest rate r and

n = the lifetime of project (Drummond et al., 2005).

The main output of a CBA is an indicator of the investment's net value. In the formula above, the indicator is net present social benefit (or net present value) which represents a summative assessment of whether a program is worth doing. As the formula specifies, the factors taken into account in the assessment include monetary valuations of the opportunity costs of value consumed and value created (costs and benefits, expressed monetarily), the points in time at which they occur over the life of the project, and the discount rate applied.

The discount rate requires some explanation and is explored in greater detail in subsequent chapters. In brief, the discount rate reflects the time value of money – the observation that we would rather be given a dollar now than later (Destremau & Wilson, 2017). A dollar we have to wait for is less desirable, and therefore has a lower value, than a dollar we receive immediately. This lower value reflects the opportunity cost associated with the delay – that is, the use to which we could have put that dollar in the intervening period, had we received it immediately. If we could invest it at eight percent per annum, turning it into $1.08 after one year, then the discounted value of a dollar we have to wait one year to receive is 1/(1.08) =

93 cents. In this sense, "discounting is interest the other way round" (Gargani, 2017, p. 118).

More formally, the discount rate represents the opportunity cost of the investment – the rate of return that could be earned from the next-best use of resources, such as another investment with similar risk characteristics to the one being evaluated (Levy & Sarnat, 1994). The discount rate is the means by which one investment is compared with a hypothetical universe of relevant alternatives.

The implication of the discount rate, when interpreting net present value, is that when the net present value is equal to zero, the policy or program under consideration does not have zero value; rather it is of equal value to its next-best alternative (the discount rate at which this occurs is known as the *internal rate of return* of the project). The higher the discount rate, the higher the standard a project must meet to be considered worthwhile (Drummond et al., 2005).

CBA is widely used to inform capital investment and financial decision-making (Levy & Sarnat, 1994), where records or forecasts of capital invested and revenue generated, and analysis of different scenarios, can support rational decision-making about the projects that will give the best return on investment. Its application is far wider than financial investments, however. CBA is widely used in diverse areas of policy making from transportation (Damart & Roy, 2009) to healthcare (Drummond et al., 2005) and education (Levin & McEwan, 2001). Since the 1980s all regulatory proposals in the US have been required to be subjected to CBA (Adler & Posner, 2006). By intent, the scope of CBA is just as capacious as related methods such as Social Return on Investment (SROI) (Nicholls, Lawlor, Neitzert, & Goodspeed, 2012). Money is a way of representing value and, in principle, anything can be valued monetarily (Nicholls et al., 2012; Svistak & Pritchard, 2014).

Assuming the scope of a CBA is comprehensive, and its valuations are accurate, comparisons can in principle be made across diverse alternatives that do not share the same objectives (Levin & McEwan, 2001) – for example, a health care intervention could be compared with an education program to determine which returns the greatest benefit relative to its costs. Net present value is used when determining the worth of mutually exclusive investment options; when allocating resources across a portfolio of investments, another indicator is used: the benefit:cost ratio which is the sum of discounted benefits divided by the sum of discounted costs (Gargani, 2017).

Although a single indicator like net present value is the principal output of a CBA, wider considerations, including qualitative considerations, may also be examined. For example, an investment may have social, cultural, environmental or ethical considerations that are not fully reflected in the

analysis and these considerations may be explored qualitatively (Executive Order No. 13563, 2011, p. 3821; Sinden et al., 2009). No guidance is given, however, on how these additional considerations should be integrated with the core findings of a CBA.

The widespread use of CBA, like that of program evaluation, has a fairly recent history. The origins of CBA trace back to the 19th century French economist Jules Dupuit, Chief Engineer for the City of Paris, whose seminal works were guided by the principle that "the only utility is that which people are willing to pay for" (Talvitie, 2018, p. 14). Dupuit's ideas were further developed by British neoclassical economists Alfred Marshall and A.C. Pigou in the early 20th century (Chapple, 2017). CBA was developed and used for military purposes during the second world war. During the 1960s and 1970s it came into wider use as a tool across many areas of public policy. In practice, it has been more widely used for assessing the potential value of planned or proposed policies, than for summative evaluation of the value of established policies or programs (Mintrom, 2017). The body of literature on CBA has become increasingly extensive and detailed from the 1950s onwards. The method continues to develop; Sunstein (2018, p. xii) described CBA as being in its "early adolescence" and predicted it will improve in coming decades. Current theoretical foundations of CBA are summarised as follows.

Theoretical foundations of CBA

Economic evaluation has multiple theoretical layers. While there is a standard formulation of CBA as a method, which bundles these layers together, some theorists have unbundled and analysed key principles individually (Sen, 2000). The first and most fundamental theoretical layer underpinning CBA is the proposition that advantages ('benefits') and disadvantages ('costs') of any intervention can be identified. The second layer adds the notion that, in order to have meaning, relevant costs and benefits should be measured and weighted in some way – and, since a cost is a foregone benefit, that they can be weighted in commensurable units and aggregated. As a mathematical concept this is uncontroversial – though some points of disagreement arise; for example, over the relative merits of additive, multiplicative, or other forms of aggregation (Sen, 2000).

The third theoretical layer is concerned with how costs and benefits should be weighted – and it is here that points of view diverge over what should be measured, and why. The foundation of CBA is a positivist approach to knowledge creation, which believes facts are separate from values, facts can be determined empirically, and that the goal of policy research is to reveal universal laws that can be applied to all policy problems (Shapiro & Schroeder, 2008). In the case of CBA, the universal dictum is that overall wellbeing should be the sole criterion for choosing between policies (Adler & Posner, 2000). This rests on a consequentialist utilitarian perspective – the

consequentialist part of the argument being that "the right course of action is the one that leads to the best consequences" (Frank, 2000, p. 77) and the utilitarian part being that "consequences can be ranked according to highest total utility" (Frank, 2000, p. 79). Accordingly, costs and benefits are weighted according to their utility.

Adler & Posner (2006, p. 13) described CBA as "a device for converting utility losses and gains from a project or regulation into dollar values, and aggregating". Everybody affected by a project experiences a change in utility, and this change can be measured by a *compensating variation*, the monetary value that should change hands to make a person as well off as they would be in the status quo, based on their preferences. Compensating variations can be revealed, for example, by observing prices determined by the behaviour of markets or through surveys and other methods to directly elicit people's 'willingness to pay' (Drummond et al., 2005). CBA provides the analytical structure for aggregating these valuations.

In CBA, the aggregated gains or losses in utility are evaluated against the Kaldor-Hicks criterion. An allocation of resources is said to be Pareto-efficient if there is no alternative allocation in which one person can be made better off without making somebody else worse off (Drummond et al., 2005). The Pareto criterion is too restrictive, however, to be practical for evaluating real-world policy proposals, which usually have distributive implications. Kaldor (1939) modified the Pareto criterion by arguing that for an action to be in the public interest, those who gain from it must be able to compensate those who lose from it, and still find the action worthwhile. The compensation, however, does not have to actually occur (Drummond et al., 2005). Hicks (1939) added that the losers must not be able to bribe the gainers from forgoing the action.

The Kaldor-Hicks criterion reflected paradigms of "New Welfare Economics" which emerged at the end of the 1930s. The New Welfare Economics attempted to establish standards by which policies could be judged as economically desirable and could be optimised, based on maximising welfare. It sought to develop propositions that were 'scientifically' ethics-free (Backhouse, 2016). The Kaldor-Hicks criterion offered economists such a standard, for evaluating the allocative efficiency of resource allocations. In CBA, this notion of allocative efficiency is accepted as uncontroversial (Brent, 2006); a "moral criterion [that] identifies the features of outcomes that make them morally better or worse than alternatives" (Adler & Posner, 2006, p. 62). Weighing costs and benefits in this way means that "any change that increases the 'social wealth' according to the Kaldor-Hicks criterion. . . is routinely interpreted as an 'increase in efficiency', particularly in the law-and-economics literature, cost-benefit analysis, policy analysis, and other parts of applied welfare economics" (Ellerman, 2014, p. 126-127).

Features of CBA such as valuing gains and losses in utility in commensurable units, evaluating net benefits against the Kaldor-Hicks criterion, and the possibilities this holds for comparing diverse evaluands on a like-with-like basis, has led to CBA often being regarded as the gold standard for evaluating the worth of a program or policy (Julnes, 2012b). CBA is the most comprehensive of the economic methods of evaluation (Drummond et al., 2005) and for these reasons this thesis focuses primarily on CBA.

Two further economic evaluation methods warrant mention, however. CEA and CUA differ from CBA in that they quantify costs and consequences in differing units. The cost measure and the consequence measure are brought together into a ratio, which can be used to help a decision-maker choose between alternatives when making resource allocations across a portfolio of investments (Drummond et al., 2005).

Cost-Effectiveness Analysis

CEA is an approach to economic evaluation that values costs monetarily, and consequences in natural or physical units (Drummond et al., 2005; Levin & McEwan, 2001). As costs and consequences are measured in different units, they cannot be directly compared as in a CBA. Instead, the output of a CEA is a ratio of costs to consequences, known as the *cost-effectiveness ratio* (e.g., the average cost per year of life saved). As economic methods of evaluation are concerned with the opportunity cost of an intervention compared to its next-best alternative, an *incremental cost-effectiveness ratio* (ICER) is calculated that compares the difference in costs and consequences of an intervention with those of a comparator (Drummond et al., 2005).

As with CBA, costs and consequences are discounted to adjust for differential timing, though there are some controversies over whether, or when, consequences should be discounted and, if so, whether this should be done at the same discount rate as costs or at a different rate (Drummond et al., 2005).

If there are i possible investments (where $i = 1 \ldots I$), then the ICER of project i relative to next-best alternative project *0* is:

$$ICER_{i,0} = \sum_{t=1}^{n} \frac{c_i(t) - c_0(t)}{e_i(t) - e_0(t)}$$

Where:

$e_i(t)$ = effects, in natural or physical units, in year t

$c_i(t)$ = costs, in monetary terms, in year t

both $e_i(t)$ and $c_i(t)$ are discounted using $1/(1+r)$ = a discount factor at annual interest rate r and

n = the lifetime of project (Drummond et al., 2005).

The ICER requires careful interpretation. In some circumstances, a CEA might find that a program is more effective, and less costly, than its next-best alternative. In such cases it is said to be 'dominant' and unequivocally worth investing in. This is illustrated in the *cost-effectiveness plane* shown in Figure 1. Conversely if it is more costly and less effective, the finding is also clear. Even in these circumstances, however, the choice of outcome indicator and scope of costs evaluated may determine whether the findings are positive or negative.

The finding is less clear, however, where the ICER indicates that a program is more costly and more effective than its next-best alternative, or less costly and less effective. In such cases there is no clear decision rule built into CEA to determine whether the program is 'worth doing'; this requires a standard-setting exercise to determine appropriate benchmarks or thresholds to indicate what is an 'acceptable', 'good', or 'unacceptable' ICER (Drummond et al., 2005).

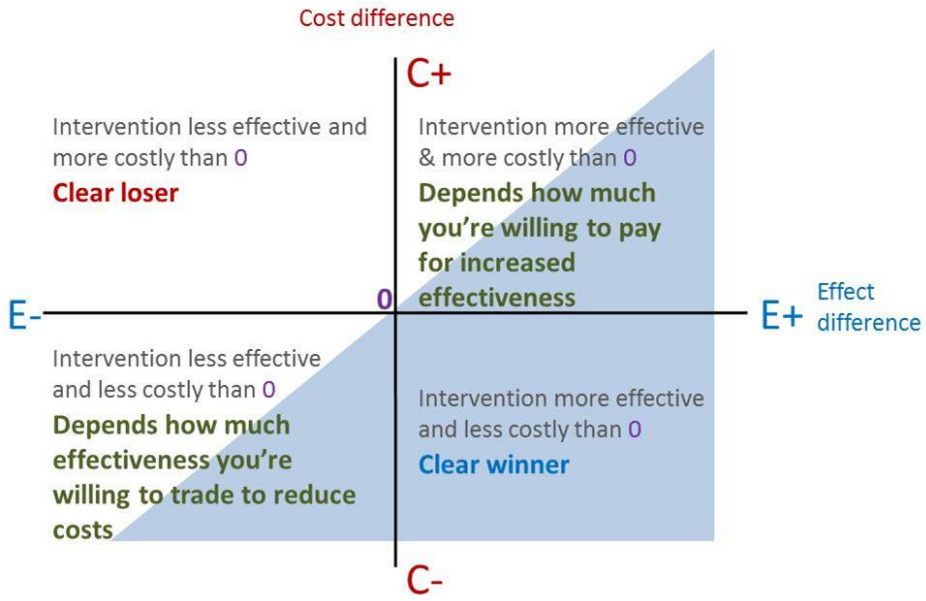

Figure 1: The cost-effectiveness plane. Adapted from Drummond et al. (2005).

CEA is not grounded in welfare theory, but in a decision-making paradigm which "assumes that a decision-maker seeks to maximise achievement of a defined objective by using a given budget" (Drummond et al., 2005, p. 24). It is most often used in situations where a limited range of options (two or more programs) are being considered to achieve a specific outcome, and where the competing alternatives can be adequately evaluated in terms of a single, quantifiable outcome measure (such as life-years gained). CUA follows the same general form as CEA, with costs and consequences measured in

different units to produce a ratio, but contains more information in the consequence measure, as described below.

Cost-Utility Analysis

While the 'effectiveness' measure in CEA is "single, specific, and unvalued" (Drummond et al., 2005, p. 137), the measurement of consequences in CUA is broadened to incorporate utility. In CUA, outcomes "may be single or multiple, are generic as opposed to program-specific, and incorporate the notion of value" (Drummond et al., 2005, p. 137).

In simple terms, a utility measure could comprise a composite index of outcomes, weighted for their relative importance (Levin & McEwan, 2001). In health economics, the relative value of outcomes is measured in accordance with multi-attribute utility theory (Von Neumann & Morgenstern, 1947; Keeney & Raiffa, 1976), with utility being represented by empirically derived measures such as quality-adjusted life years (QALY) and disability-adjusted life years (DALY), weighted measures which adjust the measurement of extended life/health spans to represent the utility of those additional years (Drummond et al., 2005).

The Von Neumann-Morgenstern utility theorem (Von Neumann & Morgenstern, 1947) differs from the notion of utility used in welfare economics in that it is a normative model proposing how individuals should make decisions prospectively, when future outcomes are uncertain. Von Neumann-Morgenstern utility does not represent experienced utility, happiness or wellbeing. Where there are two or more options, the outcomes of which are uncertain but can be defined probabilistically, the theorem argues that people should maximise the expected value of a utility function that defines the potential outcomes at a specified time in the future. Maximising expected utility is more than simply multiplying the probability of an outcome occurring by its value if it occurs – for example, the Von Neumann-Morgenstern utility function also accommodates an individuals' attitude toward risk (Drummond et al., 2005).

Because of these features, CUA has broader applicability than CEA and can be used in situations where a more diverse range of alternatives are being considered, including alternatives that are not pursuing similar outcomes. Costs and utility are brought together into an *incremental cost-utility ratio* (ICUR) which, like the ICER, can represent relative value but requires a decision rule or threshold to determine whether a stand-alone proposal is worth investing in.

What CBA, CEA and CUA have in common is that they all evaluate costs and consequences, and all yield indicators of efficiency. Additionally, all of the methods are often used to estimate future value, even when based in part on evidence of past effectiveness. Strengths of economic methods of evaluation

include transparently modelling and forecasting future scenarios under various different combinations of assumptions (scenario analysis) and assessing the sensitivity of model outputs to varying input assumptions (sensitivity analysis). "Sensitivity and scenario analysis facilitate transparency and robust thinking about relationships between benefits and costs, taking uncertainty into account, which can lead to insights that could otherwise be difficult to gain" (King, 2017, p. 104).

An area of overlap in the focus of economic evaluation and program evaluation is the application of these methodologies to assessing the value of social policies and programs. The following section briefly introduces this context, and in particular the notion of social spending as an investment, before turning to the core focus of the thesis: value for money.

Social policies and programs – an 'investment' view

This research concerns a contest of ideas; about what it means to use resources well, what it means to evaluate resource use, and how this should be done. The setting for this contest is any policy, program, or intervention where resources are used to pursue social goals – that is, to change society or the lives of people within it, in ways that are deemed to be positive. Under neoliberal models of social spending it has become fashionable to conceive of social policies and programs as investments (Destremau & Wilson, 2017).

The term "social investment" as it is used here applies to traditional publicly-funded 'welfare state' models of social policy, but not exclusively so. The principles canvassed in this research also pertain to the growing role of philanthropists, social enterprise, and corporations as investors in social change (Gargani, 2017; Svistak & Pritchard, 2014).

Different societies display different sets of social values and politico-economic traditions when it comes to the role of the welfare state (Esping-Andersen, 1990; Morel, Palier & Palme, 2012) and these continue to evolve over time (Destremau & Wilson, 2017). This research acknowledges these differences without dwelling on them; it is primarily concerned with *how* those values, whatever they may be at a particular place and time, should be used to evaluate and inform resource allocation decisions.

Nonetheless, it is important to briefly acknowledge that the idea of social spending as an investment has itself arisen out of these evolving welfare state models. For instance, Morel, Palier & Palme (2012) described the evolution of welfare from a Keynesian approach, with a passive social policy focused on the 'here-and-now', to a neoliberal perspective in which social policies have come to be seen more as a 'cost' and social policy is increasingly evaluated through a lens of economic efficiency (Destremau & Wilson, 2017). The notion of 'investing' in social change can be framed as another step along this evolutionary path, with Giddens (1999) seeing social

investment as part of a 'third way social democracy' – including a shift in focus from costs to returns, and the growth in non-government actors as social investors, in addition to governments. In this way, the notion of social investment is connected to political ideologies about the roles of markets, work, individual responsibility, and collective social rights and duties (Destremau & Wilson, 2017).

An 'investment' is the use of resources to increase capital:

> *Under traditional accounting practice an investment is a form of expenditure which results in an asset that lasts longer than the current accounting period. The cost of the asset over the course of its life is recorded as a non-cash expenditure for depreciation year by year. The asset produces either revenues from sales or user charges, while decisions about whether to make such investments are based on an assessment of the size of these revenues compared with the cost of the asset – the rate of return. In a purely commercial context the streams of cost and revenue are also implicitly the social evaluation of the investment. There is assumed to be no market failure driving a wedge between private and social cost and benefit* (Scott, 2017, p. 421).

In a financial investment, the investors and beneficiaries are either the same people, or at least their interests and objectives tend to be relatively well-aligned. In contrast, in a social investment, those who invest and those who benefit are different people and may have very different perspectives about what constitutes a return on the investment (Carlton & Perloff, 1994; Destremau & Wilson 2017; Gargani, 2017). In contrast to a financial investment, where the principal goal is to profit financially, 'social investment', defined herein as any allocation of resources in pursuit of social goals, encompasses a far greater diversity of possible returns:

> *Potentially, returns can include many different things – with the parameters varying according to what is believed to represent a desirable outcome. Typically, returns are thought to include benefits of an economic nature, such as improved economic performance, as reflected, for instance, in higher aggregate output, greater productivity, lower inflation, more employment and better fiscal outcomes. But returns can also be defined more broadly to embrace desirable social, cultural or environmental outcomes. These might include gains in objective well-being, as measured, for example, by improvements in health status, longevity, educational attainment and housing quality. Additionally, they might include gains in subjective well-being and/or the achievement of broader and perhaps more intangible goals. Among the latter might be improvements in social cohesion, social mobility, distributional fairness, societal trust and sustainability, or changes to deeply entrenched but harmful cultural practices and attitudes (e.g., racism, sexism etc.).* (Boston 2017, p. 93).

As notions of return on investment have become increasingly linked to social change, the term 'value for money' (VFM) has entered widespread use (Dumaine, 2012; Julnes, 2012c; Svistak & Pritchard, 2014; Yates, 2012). In the face of constrained resources, contested policy priorities, and political pressures to be accountable for effective targeting and management of resources, evaluators are often asked to determine whether policies and programs provide VFM. This is challenging, however, not least because VFM has not been clearly defined.

Value for money

There have been notable calls for many years for program evaluators to consider cost when evaluating programs and policies (Herman et al., 2009; Levin, 1987; Scriven, 1993; Yates, 1996) and increasing calls for investments in social change to deliver and demonstrate VFM (Dumaine, 2012; Gargani, 2017; Julnes, 2012c; Svistak & Pritchard, 2014; Yates, 2012). In international development for example, because of limited aid budgets and political pressures to be accountable for the use of taxpayers' funds, it is accepted that aid should be well targeted and managed effectively (Adou, 2016), leading to an increased interest in VFM (Fleming, 2013). There is no universal definition of VFM, however, nor are there any internationally standardised guidelines on how to approach VFM in programming, hence how to apply VFM methods continues to be a field of debate (Renard & Lister, 2015).

VFM is variously defined in the literature (Adou, 2016) and appears to exist more as a bureaucratic or political concept than an academic one. Few mentions were found of VFM in academic literature. Schwandt (2015, p. 52) argued that the term is a common expression of "the extent to which monetary costs, time, and effort are well used in achieving specific outcomes", and linked VFM with the economic concept of efficiency. Similarly, Levin & McEwan (2001) associated the colloquial expression "bang for the bucks" (p. 2) with efficiency and economic methods of evaluation. VFM also has a loose association with the evaluative concept of worth (Davidson, 2005; Liddle, Wright & Koop, 2015). In other publications, the term "value for money" was not formally defined but was used consistently with these everyday meanings (Dumaine, 2012; Julnes, 2012c; Patton, 2008).

Efficiency is defined in multiple ways, but generally encapsulates the notion of maximising desired outputs, outcomes or utility for a given level of inputs. For example, the formal economic definition of efficiency references the Paretian concept of maximising aggregate social welfare from available resources (Drummond et al., 2005). This concept may be applied at whole-of-society level, or to the activities of a particular market, firm or other unit of production, where it means that no additional output or outcome could be

achieved without increasing inputs, and production occurs at the lowest possible average cost (Carlton & Perloff, 1994).

Worth, in evaluative terms, is associated with the extrinsic "value of something to an individual, an organization, an institution, or a collective" (Scriven, 1991); "more often than not in evaluation, we are looking at whether something is "worth" buying, continuing to fund, enrolling in, or implementing on a broader scale" (Davidson, 2005, p. 2). Efficiency is one criterion of worth – but a social program could be worth funding for other reasons such as equity or fairness (Pinkerton et al., 2002). Indeed, Coryn and Stufflebeam (2014, p. 14) argued: "We assert that equity, in the broadest sense, is an important criterion for all evaluations that involve delivering programs to groups of people".

It is beyond the pages of academic texts, and in particular within political and bureaucratic publications, where diverse definitions of VFM in everyday use are revealed. In public policy contexts, governmental publications have defined VFM in various ways that include maximising value in the procurement of inputs; delivery of outputs; achievement of outcomes or objectives; maximising outputs or outcomes for a given level of input; environmental sustainability; ethical resource use; and/or impacts on equity (DFID, 2011; Dumaine, 2012; Fleming, 2013; ICAI, 2011; s, 2012).

For example, the Organisation for Economic Co-operation and Development defined VFM as "the optimal combination of whole-life cost and quality (or fitness for purpose) to meet the user's requirement" (Jackson, 2012, p. 1). The World Bank defined VFM as "the effective, efficient, and economic use of resources" (World Bank, 2016, p. 1). Both of these definitions appear focused principally on accountability in the use of resources to purchase inputs and outputs.

Similarly, the Australian Government Department of Finance (2018) used the term "value for money" only in relation to procurement, noting that "achieving value for money is the core rule of the CPRs [Commonwealth Procurement Rules]" (web page). The Department did not define VFM directly, but directed officials to "consider the relevant financial and non-financial costs and benefits of each submission" – including, but not limited to, their quality, fitness for purpose, the supplier's experience and performance history, flexibility of the proposal (including innovation and adaptability over the lifecycle of the procurement), environmental sustainability of the proposed goods and services (e.g., energy efficiency), and whole-of-life costs.

The Australian Department of Foreign Affairs and Trade (DFAT) has not explicitly defined VFM but set out eight principles, organised under four themes, "to guide decision making and maximise the impact of its investments" (web page, n.d.). The four themes and their associated

principles are labelled 'economy' (cost-consciousness and encouraging competition), 'efficiency' (evidence-based decision making and proportionality), 'effectiveness' (performance and risk management, results focus, experimentation and innovation) and 'ethics' (accountability and transparency). These principles hint at an important relationship between resource use and outcomes, but remain primarily accountability and procurement-focused.

New Zealand's aid programme (Ministry of Foreign Affairs and Trade, 2011) referenced outcomes more explicitly, defining VFM as "achieving the best possible development outcomes over the life of an activity relative to the total cost of managing and resourcing that activity and ensuring that resources are used effectively, economically, and without waste" (p. 1).

All United Kingdom (UK) government departments are required to achieve VFM in their use of public funds. The UK Treasury defined VFM as "the optimum combination of whole-of-life costs and quality (or fitness for purpose) of the good or service to meet the user's requirement" (HM Treasury, 2006). While focused on *ex-ante* VFM assessment to guide policy decision-making and procurement, the Treasury's guidance took a broad perspective on VFM that included both quantitative and qualitative assessment including consideration, for example, of whether an investment is viable, achievable, desirable, operationally flexible, innovative, and equitable.

Of particular relevance to this research, the UK Department for International Development (DFID) defined VFM as "maximising the impact of each pound spent to improve poor people's lives" (DFID, 2011, p. 2) and disaggregated this concept into five dimensions, labelled 'economy', 'efficiency', 'effectiveness', 'cost-effectiveness' and 'equity'. These dimensions were associated with particular levels of a generic program 'results chain' (Figure 2). For example, economy focused on the conversion of resources into inputs; efficiency on the conversion of inputs into outputs. The effectiveness, cost-effectiveness, and equity criteria are outcome-focused.

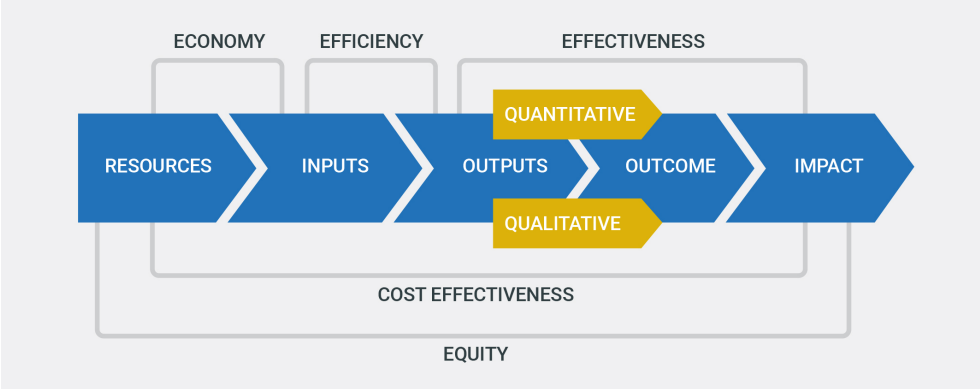

Figure 2: VFM conceptual framework. Reprinted from King & OPM (2018). Reprinted with permission.

DFID (2011, p. 4) defined these criteria as follows.

> *Economy: Are we or our agents buying inputs of the appropriate quality at the right price? (Inputs are things such as staff, consultants, raw materials and capital that are used to produce outputs)*
>
> *Efficiency: How well do we or our agents convert inputs into outputs? (Outputs are results delivered by us or our agents to an external party. We or our agents exercise strong control over the quality and quantity of outputs)*
>
> *Effectiveness: How well are the outputs from an intervention achieving the desired outcome on poverty reduction? (Note that in contrast to outputs, we or our agents do not exercise direct control over outcomes)*
>
> *Cost-effectiveness: How much impact on poverty reduction does an intervention achieve relative to the inputs that we or our agents invest in it?*

These definitions all relate to the notion of efficiency – that is, they are concerned with maximising productivity along a notional results chain by procuring as much input/output/outcome/impact as possible with a given level of resources.

DFID did not stop at efficiency, however. DFID's approach to VFM also stressed the importance of equity – acknowledging that VFM of development aid depends not just on maximising outcomes for a given level of resources, but also on how those outcomes are distributed:

> *When we make judgements on the effectiveness of an intervention, we need to consider issues of equity. This includes making sure our development results are targeted at the poorest and include sufficient targeting of women and girls.* (DFID, 2011, p. 3)

The inclusion of equity in DFID's definition of VFM means that, at least in some circumstances, VFM is not solely concerned with efficiency. VFM is a multi-dimensional concept, requiring judgement (DFID, 2011) and balancing trade-offs (ICAI, 2018). Similarly, the other definitions of VFM referenced above all define VFM in terms of multiple objectives and criteria.

A unifying definition of VFM

Despite the lack of a consensus definition of VFM, the working definitions in the public domain reveal some pervasive themes. First, all definitions of VFM are concerned with some aspect of resource use – for example, what quantity and/or quality of resources are used, how the resources are used, whether the resource use is relevant (meets an identified need), whether it is affordable (achievable within available resources), whether it is frugal

(minimises wastage), whether sound procurement policies and processes were followed, whether the resource use is ethical, and so on.

One can take stock of resources used without valuing what results the investment has achieved. However, this will result in a partial understanding of VFM if differences in consequences are not also examined. For example, two interventions might involve similar costs while differing markedly in their productivity or outcomes (Levin & McEwan, 2001).

Second, therefore, working definitions of VFM are often (though not always) concerned with the consequences of resource use – for example, outputs and/or outcomes, how they are causally related to the resource use, their aggregate value, their incidence (how the consequences are distributed and whether they reach intended groups), and equity (how fairly they are distributed).

One can examine consequences without taking resource use into account. This again results in a partial understanding of VFM. For example, two programs might achieve similar outcomes but differ in regard to their resource use. Therefore, the most comprehensive definitions of VFM examine both costs and consequences. However, one could separately take stock of resource use and its consequences, and still not reach a comprehensive understanding of VFM unless the relationship between the two is examined.

Third, therefore, the most comprehensive definitions of VFM are explicitly concerned with the relationship between resource use and its consequences. This sometimes includes investigation of the dose-response relationship between resource use and consequences – for example, whether it is possible to spend a little more in return for a disproportionately greater increase in consequences (Drummond et al., 2005).

Fourth, VFM is concerned with the fundamental economic problem of resource allocation. Resource scarcity necessitates that choices be made including what (and how much) to do, and what not to do. VFM should therefore be concerned with opportunity cost – that is, whether something is worth doing, bearing in mind resource use, consequences, and the next-best alternative use of resources (Drummond et al., 2005). Reconciling these factors involves going beyond a reductive approach of analysing costs and consequences separately, and requires some form of aggregation or synthesis to consider them holistically (Julnes, 2012b).

VFM is an evaluative question, because it requires a determination of how well resources are used, and whether the resource use is justified. Therefore, in the absence of a unifying definition, VFM is defined for the purposes of this research as "the merit, worth and significance of resource use" – bringing the investment and its consequences together. This definition encompasses all of the working definitions cited herein. Of particular importance, this definition

connects VFM to the evaluative concepts of merit, worth and significance, and to the economic concept of resource scarcity.

"Resources", moreover, are not limited to the financial resources provided by a donor, philanthropist or social investor. Resources, for the purposes of this thesis, include anything with an opportunity cost, invested in producing a policy or program by any person or group. Such factors of production include the classical economic triad of land, labour and capital, and may also include other, less tangible factors. For example, a community leader might invest and risk her reputation and relationships when granting a new program access to her community.

This framing of VFM, as an evaluative question about an economic problem, asserts that both economics and evaluation share an interest in valuing resource use, and that both can be oriented toward the betterment of society (Henry & Mark, 2003; Mark, Henry & Julnes, 2000; Sunstein, 2018). Despite this shared interest, however, the two disciplines diverge in their application. Although the goals of evaluation and economics overlap when it comes to evaluating VFM, in practice the two disciplines tend to approach evaluation in distinct and separate ways. Consequently, economic evaluation tends to be applied either in isolation from, or in parallel to other forms of evaluation, usually without explicit synthesis of economic analysis and other perspectives. This 'interdisciplinary gap' is explored in the next section.

An interdisciplinary gap – and an opportunity

Economics is an influential discipline, and economic methods of evaluation have a high status in policy making. Over the last half-century, and particularly in the context of neoliberal ideas about the role of the market and the state, economics has been "the mother tongue of public policy; the language of public life and the mindset that shapes society" (Raworth, 2017, p. 5). In the United States, the use of CBA in the regulatory process has been enshrined in Executive Orders since the Reagan Administration (Adler & Posner, 2006), with policy makers following the maxim that "no action may be taken unless the benefits justify the costs" (Sunstein, 2018, p. 3).

Evaluation is in many ways a marginalised discipline; the theory and practice of evaluation remains poorly understood outside of the evaluation profession, and contested within it (Gargani, 2016). While Economics 101 is widely taught as a component of many university degrees globally (Raworth, 2017), the fundamentals of evaluation are not. Whereas economics as a distinct discipline has its own university faculty, evaluation typically does not; it tends to remain a niche interest of a few academics within other departments such as education, psychology, or public health. Evaluation has been described as "the largest profession no one has ever heard of" (Gargani, 2016).

To the economics discipline, CBA, CEA and CUA are the core tools of evaluation, and the outputs of analysis provide the answer to the evaluative question of whether the evaluand is good enough to justify the resources used. Indeed, in public policy, CBA is often seen as the gold standard for assessing value for money (Julnes, 2012b). Economists, however, understand that economic analysis addresses one criterion: efficiency (Drummond et al., 2005).

To the evaluation discipline, economic methods of evaluation are just one set of evaluation tools, among many others (Julnes, 2012b). But they are a set of tools requiring specialist skills, and few evaluators are trained in economic analysis (Levin, 1987; Persaud, 2007). Moreover, the vast majority of program evaluations do not include costs (Herman et al., 2009; Levin, 1987; Yates, 2012). Such studies may, for example, be able to evaluate whether outcomes are being achieved, and the value of those outcomes from various perspectives, but cannot relate this value to the cost of achieving those outcomes – and therefore cannot conclusively determine whether the program is a worthwhile use of resources.

Scriven (1993) argued that "cost is crucial" (p. 32) but also that cost analysis should be integrated within broader evaluative frameworks:

> *We must move evaluation away from the task of determining a program's progress toward its goals and instead emphasize the task of determining its comparative cost-effectiveness in meeting needs, integrating into this conclusion consideration of unexpected outcomes, generalizability, and various value standards, including those from needs assessments, ethics, the law, and other relevant disciplines.* (p. 37).

Conversely, where economic evaluation is conducted in isolation from other forms of program evaluation, there is a risk that these evaluations may provide an incomplete picture of a program's worth, with the risk that this might result in "a distorted understanding of the public interest and a diminished capacity for evaluation in general to serve that interest" (Julnes, 2012a, p. 1). For example, a key criticism of CBA is concerned with its construct validity – that is, whether the output of a CBA (such as net present value) actually represents overall wellbeing (Adler & Posner, 2006; Eberle & Hayden, 1991) and whether maximising net value should have primacy as a social goal (Adler & Posner, 2006).

In contrast to the notion of Kaldor-Hicks efficiency as an external "moral criterion" (Adler & Posner, 2006), "resources may also be allocated for reasons of equity, which itself has multiple possible conceptions" (Chapple, 2017). Questions of equity require evaluators to adopt a normative position on how resources and opportunities should be distributed, necessitating explicit value judgements. This would require either modifications to CBA (Adler & Posner, 2006) or a different set of evaluation methods.

The interdisciplinary gap identified here has been recognised within the evaluation community and it has been suggested that economic and other valuing methods should be better integrated (Davis & Frank, 1992; Julnes, 2012c).

> *The aspirations and challenges of economic valuation offer several lessons for the field of evaluation. First, economic valuation needs to be more fully appreciated and used by evaluators... Second, the thoughtful criticisms of economic approaches should make us sceptical of accepting any valuing methodology as the gold standard for this task. Building on this, the third lesson is the need to retain, and even defend, the traditional valuing methodologies in evaluation, such as checklists, surveys, focus groups, case studies, and evaluator judgment. These other approaches complement the insights of economic valuation, although there is little consensus on their relative contributions, nor on the contexts in which they are most useful* (Julnes, 2012c, p. 111).

These considerations raise a question of what might be gained by using economic and other methods in combination to evaluate the merit, worth and significance of resource use. The limitations from economic methods are not necessarily addressed by other evaluation methods – but the possibility warrants investigation. Currently, "the formal treatment of public policies as investments, entailing the combination of cost-benefit analysis and programme evaluation, is largely still in its infancy" (Mintrom, 2017, p. 76). This is an area ripe for research.

Summary

This research seeks to make an original, necessary and important contribution to the fields of evaluation and economics: A model to guide the use of economic methods within evaluation, to address questions about VFM in social programs and policies.

VFM is variously defined in the literature. For the purposes of this research VFM is defined as the merit, worth and significance of resource use. This definition encompasses the multiple working definitions of VFM in use, and conceptually links VFM to fundamental concepts underpinning program evaluation and economic evaluation. There may be potential to strengthen evaluation of VFM by bringing together the fields of program evaluation and economic evaluation.

A gap has been identified in the existing body of knowledge: An explicit theoretical foundation has not been established to determine whether it is valid and desirable to combine evaluative reasoning with economic evaluation, guide when this might be valid or desirable, nor to suggest how this might be done. An opportunity exists for theory-building research to fill this gap. Such research should start by reviewing existing theory and

knowledge (Shepherd & Suddaby, 2017) to develop a conceptual model, identifying the features of good VFM evaluation. This new model should be systematically compared with its existing rival, CBA (Yin, 2009). To be useful, the theory must also be operationalised (Patterson, 1986; Wacker, 1998) and empirically investigated (Wacker, 1998; Yin, 2009). Accordingly, the following research questions are addressed in this research:

RQ1: What are the requirements of a model for evaluating value for money in social programs?

RQ2: To what extent and in what circumstances can cost-benefit analysis meet the requirements of the proposed model?

RQ3: How should the model be operationalised?

RQ4: To what extent are the model's theoretical propositions applicable in evaluation in real-world contexts?

In this thesis, a model of VFM assessment is developed, implemented and evaluated through a deductive process of theory-building and investigation. The research commences with a review of existing theory and knowledge to develop a conceptual model, identifying the requirements for good evaluation of VFM. The purported gold standard, CBA, is then systematically assessed against the requirements of the conceptual model to understand its strengths, limitations, and the manner and circumstances in which CBA might contribute to meeting the requirements of the model. A process model is developed, guiding the application of the conceptual model in practice. The model is used in two real-world evaluations of VFM, which are analysed as case studies to determine whether the theory works in practice, and to identify areas for further research and development. The next chapter describes the overarching design and methods used in the research.

Chapter 3: Method

The aim of this research is to develop and empirically assess a conceptual model. The model is intended to guide the combined use of evaluative and economic thinking to evaluate VFM in social policies and programs. The research takes a multi-method approach underpinned by a theory-building methodology, using critical analysis of literature together with practice-based experience to develop the model, and case studies to investigate the conceptual quality of the model in two real-world settings.

In order to address the research in a logical manner, a series of investigative phases, with corresponding research questions, are followed. First, a conceptual model is developed, proposing a set of requirements for good evaluation of VFM. Second, a gap analysis is undertaken to systematically assess the potential efficacy of CBA, the purported 'gold standard' against the requirements of the conceptual model. Results from these first two studies are parsed into a series of theoretical propositions to be tested empirically. Third, the principles of the conceptual model are operationalised by identifying a series of steps that should be followed in practice. Finally, case studies are systematically examined to investigate the conceptual validity and replicability of the theoretical propositions.

The following **research questions** are addressed:

RQ1: What are the requirements of a model for evaluating value for money in social programs?

RQ2: To what extent and in what circumstances can cost-benefit analysis meet the requirements of the proposed model?

RQ3: How should the model be operationalised?

RQ4: To what extent are the model's theoretical propositions applicable in evaluation in real-world contexts?

The objective of this chapter is to introduce the general approach and methods used. First, theory building methodology is explained. Second, a phased approach is presented, and the specific methods used in each phase are described. Finally, relevant ethical issues are addressed.

Methods

This study involves a cumulative process of theory-building and empirical investigation. The approach is multi-method. Theory development is undertaken through literature synthesis, critical analysis and triangulation of program evaluation and economic evaluation theory. Empirical investigation of the theory's conceptual validity is carried out, with case study

methodology providing a framework for analysis. The steps in this approach are described later. This section focuses on the methodological principles that underpin the research from beginning to end.

Theoretical, conceptual and practical models

'Theory' has multiple meanings. For example, the term can refer to causal, conceptual, philosophical, hermeneutical, or normative theories, among others (Abend, 2008). A theory can be conceptualised as a "a set of analytical principles or statements designed to structure our observation, understanding and explanation of the world" (Nilsen, 2015, p. 2), or sometimes, predictions or generalisations about a field or phenomenon (Scriven, 1991). The purpose of a theory is "to organise (parsimoniously) and communicate (clearly). . . by offering a coherent explanation of a phenomenon, making assumptions and building on those assumptions to logically derive predictions, offering conjectures that allow for refutation or falsification, and testing" (Shepherd & Suddaby, p. 75).

Models are closely related to theories, and the boundary between the two is not always clear (Nilsen, 2015; Scriven, 1991). Models are not as general as theories, and they typically entail a deliberate simplification of reality, which need not be completely accurate to be of value (Nilsen, 2015). In this research, two models were developed: a conceptual model and a process model.

The first model delves into evaluation theory, to consider the meaning of VFM, what it means to evaluate VFM and the requirements for good evaluation of VFM. The conceptual model is not a theory in the 'hard sense' (Flyvbjerg, 2011), for example, it does not comprise explanation or prediction in the manner of Abend's (2008) *Theory*$_1$ or *Theory*$_2$; rather it seeks to provide understanding about the concepts of evaluative reasoning, economic evaluation, and VFM, and how they are related in regard to meaning, in the manner of Abend's (2008) *Theory*$_5$. In this sense it is similar to the depiction by Miles, Huberman and Saldaña (2014) of a conceptual framework which "explains, either graphically or in narrative form, the main things to be studied – the key factors, variables, or constructs – and the presumed interrelationships among them" (p. 20). The conceptual model developed in this research is also aligned with the notion of a 'soft theory' as described by Flyvbjerg (2011) in that it involves developing and testing propositions.

The second model has a more operational focus and sets out a stepped process to guide the translation of the theoretical model into evaluation practice. The process model is an amalgam of published evaluation theory and accumulated practice-based knowledge.

The objective of the research is to make a novel, significant contribution to evaluation theory and practice by developing and empirically investigating a

normative model for the evaluation of VFM in social policies and programs. In having a normative component, the model is aligned with Abend's (2008) *Theory$_6$.* The model describes what evaluators should do, as distinct from what they actually do (Scriven, 1991).

Theory-building methodology

Theory building often aims to contribute cumulatively to knowledge, with each new theoretical contribution being presented as an interim struggle adding to an ongoing process of theorising (Shepherd & Suddaby, 2017). Such contributions should be novel – they should reveal something previously not known or encourage reconsideration of something generally thought to be settled, or something counterintuitive. Contributions should also be useful, offering scientific utility (such as advances in conceptual rigour) and/or practical utility (application to real-world problems facing practitioners) (Shepherd & Suddaby, 2017). Accordingly, the aim in this research is to contribute a new model, of practical use to evaluators. This thesis is ambitious, in that it questions the received wisdom that CBA provides the best answer to a VFM question, and develops and tests an alternative model. At the same time the ambition is modest, in that the results of this research are a contribution to the interim struggle, subject to ongoing logical debate and empirical investigation.

Theoretical contributions may be "triggered by tensions that exist between what we know and what we observe" (Shepherd & Suddaby, 2017, p. 80). Such a tension was indeed the catalyst for this research, as the author observed a gap between explicit evaluative reasoning and economic evaluation. The theory and practice of both disciplines appear to have common objectives and focal points, yet tend to operate in a siloed manner, missing an opportunity to integrate insights and findings. This study aims to make an original, useful contribution by bridging a gap between explicit evaluative reasoning and economic evaluation, as a way of understanding their respective orientations to VFM and guiding their use within an integrated methodological model.

The two general objectives of research are rational theory-building and empirical fact-finding (Wacker, 1998). This research has a theory-building objective and proceeds in a deductive, 'top-down' manner, developing a conceptual model through analysis of literature, which is then operationalised and tested empirically through case studies. This contrasts with inductive approaches, which use empirical research as the starting point for build a theory from the 'bottom up' (Wacker, 1998). This research does not rely on deductive reasoning alone, however. Some theorists acknowledge that theory development can involve an interplay between deductive and inductive reasoning, drawing on the formal and experiential knowledge of the theorist (Shepherd & Suddaby, 2017).

The process of theory development involves deliberate exploration of tensions between two epistemological realms, the rational world of theoretical literature and the empirical world of observable phenomena. While rationalism values the abstraction of knowledge into generalisable theoretical principles and relationships, empiricism emphasises the value of direct observation without the constraint of theory. While rationalism seeks to understand new empirical knowledge through deduction from prior knowledge or theory, empiricism seeks to accumulate knowledge through induction, building observation-upon-observation (Shepherd & Suddaby, 2017).

These features of rationalism and empiricism hint at a symbiosis between the two, in which empirical observations are rationalised into theories, and theories are tested through empirical observation. Although this can be framed as a sequential process, theorists can and do move continuously between the two worlds in a process that Shepherd and Suddaby (2017) described as "pragmatic empirical theorising". This model of theorising emphasises "abductive reasoning as a practical compromise of induction and deduction" in order to more realistically encompass "the authentic process by which theorizing occurs" (Shepherd & Suddaby, 2017, p. 79).

Through this process of abduction, theories are developed, tested and refined. Surprising or anomalous results can be used to trigger inquiry, and theoretical propositions can occur *a priori* (using existing knowledge or theory), or *a posteriori* (after testing and experience of a theory in practice). Such theorising can include deductive and inductive phases, and may corroborate and/or refine theories as well as raising new questions or speculations that might inform future research (Shepherd & Suddaby, 2017). While the overarching theory-building design in this research is deductive, use is also made of abductive reasoning to bridge theory and practice. This is consistent with the observation that evaluation theorists have often drawn their ideas from practical experience (Coryn & Stufflebeam, 2014).

Although the conceptual and process models developed in this research do not represent 'hard theory' in the causal or predictive sense, the methodology described in the general literature on theory building nevertheless offers relevant insights into the theory-building process and the features of a good theory (and by extension, a good conceptual model). These are summarised as follows.

What makes a good theory?

Theories are commonly considered to have four components: conceptual definitions; a domain defining the settings or circumstances where the theory applies; a set of relationships between variables; and specific predictions or factual claims (Wacker, 1998). The conceptual model developed in this thesis proposes a novel definition of VFM, linking definitions of evaluative reasoning

and economic evaluation. The domain to which the theory applies is evaluation of VFM in social programs and policies (though the domain is narrowed to international development programs in the empirical component of the research). The conceptual model does not make predictions, but does make conceptual claims about potential differences between CBA and an alternative approach. The conceptual claims are developed through critical analysis of literature and are investigated empirically through case studies.

A good evaluation theory should "be useful in efficiently generating verifiable predictions or propositions concerning evaluation acts and consequences and . . . provide reliable, valid, actionable direction for ethically conducting effective program evaluations" (Coryn & Stufflebeam, 2014, p. 52). In this research, propositions are concerned with the features of CBA compared to an alternative approach to evaluating VFM. Empirical testing aims to determine whether the conceptual model is valid and fit for purpose.

Literature reveals a number of general features of 'good' theory. Wacker (1998, p. 365) noted that despite a lack of agreement on the relative importance of these "virtues" of good theory, "there seems to be a fairly widespread agreement as to what they are". According to theorists from evaluation and other disciplines, these features include: uniqueness, conservation, generalisability, fecundity, parsimony, internal consistency, empirical riskiness, and abstraction (Coryn & Stufflebeam, 2014; Miller, 2010; Popper, 1957; Quine & Ullian, 1980, Wacker, 1998). The applicability of these features to the conceptual model is summarised as follows.

Uniqueness means that the theory should be differentiated from other theories (Wacker, 1998). For example, this research develops a unique definition of VFM. Developing a new theory, however, can involve a balancing act between novelty and conformity – it must be sufficiently "different from received wisdom to warrant a second look", yet "similar enough to what is known to be comprehensible" (Shepherd & Suddaby, 2017, p. 77). The aim in this theory-building research is to link existing evaluation and economic theory through a unique definition of VFM, to develop a new theory about how evaluative reasoning and economic evaluation should be conducted in an integrated manner.

Conservation dictates that a current theory should not be replaced unless the new theory is superior in relation to the features listed here (Wacker, 1998). A theory with fewer acceptable alternative explanations is better than a theory with many alternatives (Shepherd & Suddaby, 2017). In this context, 'current theory' is represented by the popular view that CBA is the gold standard for VFM assessment (Julnes, 2012b), and this study seeks to determine whether such a label might be deserved, and, if not, whether there is an alternative that can combine the respective strengths of evaluative thinking and economic evaluation. In this sense, the theory-building endeavour explores a struggle between two alternatives: the

received wisdom of CBA, and the possibility of an alternative. This struggle is not necessarily expected to produce an overall 'winner' – rather it is anticipated that it might arrive at a clearer set of conclusions about the circumstances in which CBA, or a novel alternative (if viable) should apply.

Generalisability (sometimes labelled utility) means that a theory that can be widely applied is better than a theory with narrow application (Shepherd & Suddaby, 2017; Wacker, 1998). Although this criterion applies most directly to causal or predictive theories, rather than conceptual theories, it is relevant in the context of Miller's (2010) argument that a good evaluation theory is clear about the circumstances and evaluation questions to which it applies. In this research, the aim is to develop a general conceptual model that can be applied to evaluation of VFM in any social policy or program. For practical reasons, the scope of study is narrowed to programs in international development when it comes to testing the theory in case studies.

Fecundity refers to the capacity of the theory to generate new models or hypotheses that can be tested, leading to new knowledge (Patterson, 1986). Theories that explore new conceptual areas of investigation are considered superior to those that remain within the boundaries of established areas of research (Wacker, 1998). This criterion relates primarily to causal or predictive theories. However, it is relevant to note that this thesis enters unexplored territory by seeking to integrate evaluative and economic thinking to address questions of VFM.

Parsimony states that a simple theory, containing minimal complexity and assumptions, is better than an overly complicated one (Coryn & Stufflebeam, 2014; Shepherd & Suddaby, 2017; Wacker, 1998). Accordingly, the theoretical and practical models developed in this research aim to cover only those details that are necessary and sufficient to describe a good evaluation of VFM.

Internal consistency requires that the theory logically explains relationships between variables; for example, that the theory identifies all relationships and gives adequate explanation (Patterson, 1986; Wacker, 1998). This is important in statistical and mathematical models but is less relevant in the current research. By extension, however, the principle of internal consistency highlights that the conceptual model should not contain logical contradictions.

Empirical riskiness means that a good theory should not merely support something obvious or very likely, and there should be a credible chance that the theory might be refuted. Where empirical evidence and logical evidence conflict, the logical evidence is generally considered more trustworthy on the basis that there is less chance of error. However, empirical evidence may be used to refine the theory (Wacker, 1998). This research may be considered empirically risky on the basis that it questions the superiority of CBA, a well-

established and commonly accepted method for evaluating VFM and instead proposes a hybrid model that has not been tried before.

Abstraction refers to the theory's independence from time and place (Wacker, 1998). For example, Miller (2010) argued that the impacts of the theory should be reproducible over time, on different occasions, and by different evaluators. Theories can be described on a continuum from high-abstraction theories with a generalised scope, to middle-abstraction explaining limited sets of phenomena, and low-abstraction, empirical generalisations that have limited scope and application (Nilsen, 2015; Wacker, 1998). For example, evaluation theories include high-abstraction theories about the nature of evaluation, such as the General Logic of Evaluation (Scriven, 1980; 1991; 1994; 1995; 2012) middle-abstraction theories suggesting attributes evaluation should have in specific circumstances, such as deliberative democratic evaluation (House & Howe, 1999) or utilisation-focused evaluation (Patton, 2008), and low-abstraction theories such as theories about how to evaluate student learning in Australian primary schools.

The aim in this research is to develop a middle-abstraction theory, suggesting requirements and a process for addressing VFM in social policies and programs. Within the confines of this research, however, empirical investigation of the theory is limited to a lower level of abstraction. The empirical evidence from one case study represents a critical instance (US Government Accountability Office, 1990), investigating the theory at a low level of abstraction because it only applies to the time and place of one program (Wacker, 1998). Two case studies together provide the opportunity for replication (Yin, 2009), supporting abstraction at a slightly higher level. In this way, lower-abstraction theories can be used to build higher-abstraction theories (Wacker, 1998).

In addition to the core virtues of theories described above, it has been suggested that theories should demonstrate: **importance** (being significant and relevant in the real world); **practicality** (being useful to practitioners); **operationality** (capable of being converted into a procedure in order to test its propositions); and should have a discernable **impact** (intended or unintended) (Miller, 2010; Patterson, 1986; Shepherd & Suddaby, 2017).

Approaches and strategies to theory-building

Wacker (1998) synthesised theory-building literature to suggest a parsimonious model applicable to both conceptual and statistical theory-building procedures. Four stages were identified, corresponding to the four components of a theory: definitions, domain, relationships, and predictions or propositions. Although presented sequentially, the domains interact with each other and may be applied iteratively. Throughout this process, Wacker

(1998) stressed the importance of literature review in supplying accepted definitions and domains, as well as previously identified relationships.

The first step, defining constructs or variables, "helps to separate the phenomenon of interest from the mass noise of everyday experience and prior research" (Shepherd & Suddaby, 2017, p. 66). Theoretical definitions are conceptual and not observable (Wacker, 1998). For example, in this research it is noted that VFM has been defined variably in the literature, and so a construct is developed to ground the concept of VFM in a definition that reflects its core concepts and everyday use.

The second step, limiting the domain, has already been described and relates to the evaluation of VFM in social policies and programs. It is in the third and fourth steps where the detail of this thesis resides. The third step, model building, involves logically assembling the reasoning and relationships between definitions that comprise the conceptual model, through a cumulative process of critical analysis of literature. The fourth step involves defining a set of theoretical propositions and empirically investigating their conceptual quality through case studies (Wacker, 1998).

Throughout this process of model building, theoretical propositions and empirical investigation, a range of theory-building strategies are brought into play. One such strategy is the concept of *blending* – combining different theories to provide "a basis for transforming constructs and relationships. . . to generate new insights" (Shepherd & Suddaby, 2017, p. 66). In this research, the theory-building endeavour seeks to understand the interplay between evaluative reasoning and economic methods of evaluation as potential complementary approaches for evaluating VFM.

Another relevant theory-building strategy is *bricolage*: flexibly and responsively assembling "different knowledge elements that are readily available to the researcher" (Shepherd & Suddaby, 2017, p. 74) into fluid knowledge constructs. Boxenbaum and Rouleau (2011) observed that theorists engage in bricolage by combining concepts already to hand, that are sufficiently diverse that their combination might provide original and useful insights. Selecting and combining these concepts can draw on experience and common sense as well as literature. For example, in this research, the possibility is explored that economic methods of evaluation might be combined with other methods, under an umbrella of evaluative reasoning.

A third strategy is that of *thought experiments:* "posing problem statements, making conjectures on solutions to the problem, trialling conjectures, and selecting and retaining those that show promise enable the theorist to move through disciplined imagination to build a theory" (Shepherd & Suddaby, 2017, p. 67). For example, during early theory-building a hypothetical

example of a social enterprise café (based on a real café evaluated by the author) is used to illustrate and broadly test the proposed alternative to CBA.

A fourth strategy is *problematising,* through immersion in the literature, to "reveal paradoxes, problems, challenges, and puzzles" (Shepherd & Suddaby, 2017, p. 61) related to the definition and evaluation of VFM. Problematising involves "challeng(ing) the value of a theory and explor(ing) its weaknesses and problems in relation to the phenomena it is supposed to explicate" (Alvesson & Karreman, 2007, p. 1265-1266). For example, why is CBA regarded by some as the gold standard, despite it only accommodating limited values and types of evidence?

Shepherd and Suddaby (2017) argued that problematisation requires not only an understanding of the literature, but that the literature be approached with an open mind, allowing the literatures to reveal problems or gaps.

> *The work of the theorist is to move iteratively between gaps observed in the phenomenal world and those observed in the extant literature. It is often the tension created by a gap between the literature and world that ultimately triggers the need for new theory. Having triggered the theorising process by discovering or generating a conflict, the theorist conceives of a research idea, perhaps first as a simple construct or guess, that is then constructed into a theory* (Shepherd & Suddaby, 2017, p. 65).

A fifth strategy, labelled *engaged scholarship*, involves combining formal and experiential knowledge through "a collaborative form of inquiry in which academics and practitioners leverage their different perspectives and competencies to co-produce knowledge about a complex problem or phenomenon that exists under conditions found in the world" (Van de Ven & Johnson, 2006, p. 803). Similarly, Coryn & Stufflebeam (2014) characterised evaluation theory development as a creative and complex process drawing on formal and informal knowledge.

Shepherd and Suddaby (2017) noted that engaged scholarship is problem-driven research, involving researcher(s) and practitioner(s) in addressing real-world problems, in a collaborative learning environment, in which the participants employ multiple frames of references and remain open to new experiences influencing theories. Such an approach can help to address gaps between theory and practice, and solutions can contribute to both academic and practitioner knowledge. In this research, the author brings two perspectives – theorist and practitioner – to the process of theory building. As a theorist, the author looks to the published literature on evaluative reasoning and economic evaluation. As a practitioner, the author draws on accumulated practice-based knowledge in both disciplines, some of which has also been published (King et al., 2013; King, 2015; McKegg, Oakden,

Wehipeihana & King, 2018; Oakden & King, 2018; Wehipeihana, Oakden, King & McKegg, 2018).

Different theory-building strategies are emphasised at different stages of the research process, reflecting the specific aims of each stage. The following section summarises the stages followed.

Staged approach

Wacker (1998) distinguished two different types of theory-building research: analytical (based on deductive rules), and empirical (based on empiricism and induction). The theory-building undertaken in this research is analytical; it uses primarily deductive methods based on critical review of literature, to build a logical argument. Analytical theory-building research is further disaggregated into three different procedures: conceptual, mathematical, and statistical. This research is aligned with the notion of *analytical conceptual research*: its purpose is to add new insights "by logically developing relationships between carefully defined concepts into an internally consistent theory" (Wacker, 1998, p. 373). When it comes to empirically testing analytical theory, the approach to fact-finding is in large part determined by the type of theory building undertaken. In the case of conceptual theory building, empirical evidence usually comes from case studies (Wacker, 1998).

This theory-building research, accordingly, follows a deductive approach. After defining concepts and domain, the research proceeds by developing a model proposing requirements for good evaluation of VFM in social policies and programs, systematically assessing the ability of CBA to meet the requirements, and identifying a set of theoretical propositions, describing anticipated qualities of the proposed alternative that differentiate it from CBA. The conceptual model is then operationalised in a process model, and evidence is gathered by investigating the application of the model in two case studies, to see whether the anticipated qualities occur in practice. In this way, the research comprises a phase of theory development, followed by a phase of fact-finding.

While the overarching research design follows a pattern of deductive reasoning, other reasoning strategies are brought into play in the different stages of the research. The development of conceptual and process models is supported by abduction and engaged scholarship as described by Shepherd and Suddaby (2017). The systematic assessment of theoretical propositions in the case studies is consistent with the evaluative model of probative inference as described by Scriven (2012).

A cumulative approach is followed, comprising three sequential stages: development of a conceptual model providing rationale and requirements for the use of explicit evaluative reasoning and mixed methods to evaluate VFM; development of a process model setting out a series of steps and guiding

principles to be followed when conducting such an evaluation; and the conduct of two case studies, applying the model in real-world evaluations of international development programs to test the theoretical propositions and identify refinements to the conceptual and process models. Figure 3 presents these stages as a cycle, emphasising the expectation that theory and practice will continue to evolve; the models contributed through this research may be subjected to further research.

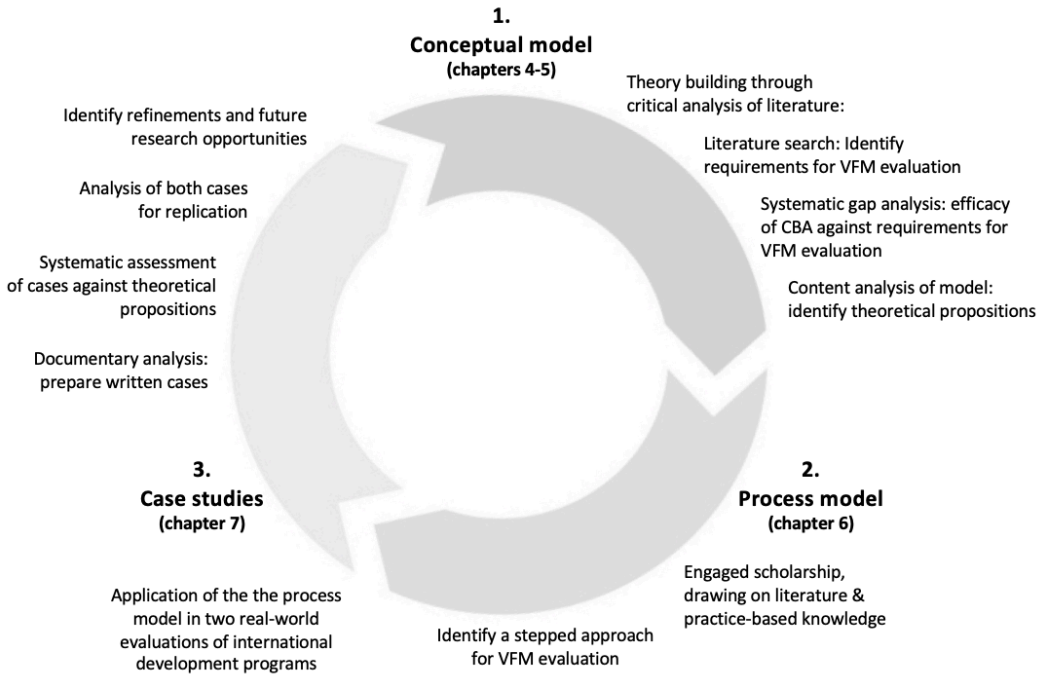

Figure 3: Overview of staged approach to the research

Stage 1: Conceptual model

RQ1: What are the requirements of a model for evaluating value for money in social programs?

The research begins by developing a conceptual model, building a logical argument through critical review and synthesis of published literature. The model proposes requirements for VFM evaluation – criteria of merit that can be used to judge a good evaluation of VFM. The model, published in the *American Journal of Evaluation* (King, 2017), provides a foundation for the second study, as follows.

RQ2: To what extent and in what circumstances can cost-benefit analysis meet the requirements of the proposed model?

A gap analysis is undertaken, systematically comparing the methodological prescription for CBA against the requirements of the conceptual model, to determine the efficacy of CBA. If CBA is a gold standard approach for evaluating VFM, it should meet all of the requirements of the model. If limitations of CBA are exposed, then the gap analysis will identify conditions and limitations of its use.

Stage 1 concludes by defining a series of theoretical propositions. These propositions summarise the findings from the first two studies and represent a set of qualities that differentiate the proposed conceptual model from CBA. The conceptual model will underpin the development of a process model for evaluating VFM. When the model is investigated empirically through case studies, the applicability of the theoretical propositions will be systematically assessed.

Stage 2: Process model

RQ3: How should the model be operationalised?

Here the conceptual model is operationalised for testing. Implementing the model requires additional specification, to translate the theoretical requirements into a practical process. A process model is developed, drawing on implementation science literature and underpinned by the general logic of evaluation. The model sets out a sequence of steps for planning and undertaking an evaluation of VFM (while acknowledging that the process may not be purely sequential and may involve some iteration between steps). The model is a parsimonious prototype, intended for testing and refinement through case studies.

Stage 3: Case studies

RQ4: To what extent are the model's theoretical propositions applicable in evaluation in real-world contexts?

In conceptual theory-building, a theory developed through deductive analysis should be empirically investigated (Wacker, 1998). In analytic conceptual theory-building, case studies are commonly used for this purpose (Wacker, 1998). Case studies provide intensive analysis of complex social phenomena and support abstraction through analytic generalisation (Yin, 2009). When case studies are used deductively for theory-testing, as in this study, theoretical propositions developed in advance of the case studies guide data collection and analysis (Baskarada, 2013). The case study research gathers evidence to see whether the qualities of the conceptual model, described by the propositions, occur in practice (Wacker, 1998).

The process model has been applied to design and conduct VFM evaluations in two real-world evaluation settings, international development programs where VFM assessment is mandated. Through documentary analysis, two

case studies are described and analysed. The case studies provide concrete illustrations of the application of the model in practice. Individually, each case study is used as a critical instance for examining the conceptual quality of the theoretical propositions. Together, the findings from the two case studies are triangulated to investigate the extent to which findings are replicated. Observations and experiential learning from the case studies inform refinements to the model.

The specific methods used in each stage of the theory-building research are detailed in the following sections, addressing each of the four research questions in turn.

RQ1: Conceptual model

The research objectives of this study are to identify a set of requirements for defensible evaluation of VFM in social policies and programs. The study involves a search and critical analysis of the literature, using strategies of theory-building (as described above) adapted to the purpose of building a conceptual model. The methods used are summarised as follows.

Literature search

Before commencing the literature search it is necessary to identify domains for investigation (Wacker, 1998). The primary domains to be investigated are evaluative reasoning, economic evaluation, and value for money. Relevant sub-domains will emerge through exploratory reading of the literature.

To select articles and books for review, keyword searches were used in the body of program evaluation and economic evaluation literature, with a view to uncovering seminal journal articles and books addressing VFM and the strengths, limitations, and compatability of evaluative reasoning and economic evaluation in the context of social policies and programs. The search was subsequently widened to include guidance documents and web pages on VFM published by government departments and multilateral aid agencies, to supplement the coverage of VFM in academic texts which primarily focused on efficiency and economic methods of evaluation. Additional references were also located using citation searches and snowballing.

The initial list of articles and books was refined by reading abstracts (and, where necessary, further exploration of the full texts) to determine their appropriateness to the review. References were excluded that did not have explicit evaluative reasoning or cost-benefit analysis or value for money at their core. Where multiple sources were found covering the same topic, they were narrowed down to the most authoritative sources on that topic to minimise duplication. The resulting 11 books and 27 papers were categorised into topics based on the three key themes.

Literature review

In line with the strategy of engaged scholarship (Shepherd & Suddaby, 2017), the analysis is shaped by the researcher's assumptions and experiences. Analysis was conducted through multiple readings and interpretations of the text, in an exploratory manner. Themes were developed iteratively, with a hierarchical structure stemming from the three domains of evaluative reasoning, economic evaluation, and value for money. Emergent sub-themes and links between them were added progressively during exploration of the literature. When new sub-themes emerged, the frame was revised and relevant documents were re-read according to the new structure. The process continued until no new themes emerged, suggesting that saturation had been reached (Thomas, 2006).

The draft model was prepared through an iterative process. Model development, employing theorising strategies of labelling constructs, blending, bricolage, thought experiments, and problematising (Shepherd & Suddaby, 2017), as described above, took place throughout the duration of immersion in the literature, and continued during the writing process. Themes were rationalised during the writing process, to observe the principle of parsimony, with the final manuscript containing only those themes considered necessary and sufficient to describe and justify the key elements of the theoretical model.

Coryn & Stufflebeam (2014) argued that program evaluation theories must themselves be evaluated, and "must pass muster with critics and the theorist's broader audience" (p. 57). Similarly, the draft model was subjected to peer debriefing and review. The developing model was presented to academic colleagues at two conferences,[1] providing opportunities for dialogue and debate with evaluation theorists and practitioners. A draft paper on the model was subsequently prepared, and was reviewed by three academic experts. These processes were not intended to validate the model, but rather to invite experts to critique and challenge it, in order to expose aspects that might require revision or clarification.

Finally the paper was submitted to a peer reviewed journal. The manuscript was published in the *American Journal of Evaluation,* initially online in May 2016, and in print in the March 2017 edition (King, 2017).

The findings from this study provided a framework for the analysis conducted in the next study, to which we now turn.

[1] Australasian Evaluation Society Conference (September, 2015), Melbourne; American Evaluation Association Conference (November, 2015), Chicago.

RQ2: Gap analysis

The objectives of this study are to systematically assess the relative strengths and weaknesses of CBA against the requirements of the conceptual model. In this research, the term 'gap analysis' refers to the systematic analysis of CBA against the model's requirements. If CBA is a gold standard approach for evaluating VFM, it should comprehensively meet the model's requirements. If not, then the gap analysis will identify conditions and limitations of its use. If there are particular circumstances in which CBA does or doesn't meet the model's requirements, then those circumstances should be taken into account to guide the selection of appropriate evaluation methods.

This analysis returns to the literature, with further targeted searching to review economic evaluation and explicit evaluative reasoning in additional detail. In this study, however, the emphasis shifts to a process of systematic assessment to examine the theoretical and methodological prescription for CBA against the requirements identified in the preceding study.

The perspective taken is that of *efficacy* (that is, it addresses the question of theoretical feasibility and validity: '*can* CBA meet the requirements?') as distinct from real-world *effectiveness* ('*does* CBA meet the requirements?' – which depends on a wider suite of contextual factors). Accordingly, the inherent capacity of CBA to fulfil the requirements is rated, rather than the extent to which CBA actually fulfills the requirements in real-world evaluations.

Three types of rating are assigned: prescribed, permitted, and precluded. A 'prescribed' rating means that the methodological standard requires the method to fulfil this criterion. 'Precluded' means that adherence to the methodological standard would logically make adherence to this criterion impossible. 'Permitted' means that meeting this criterion is neither prescribed nor precluded, so could theoretically be included as a customisation of the prescribed economic method.

Through systematic analysis, conclusions are reached about a set of principles to guide consideration of the circumstances in which CBA could strengthen an evaluation of VFM; circumstances in which CBA may be insufficient (on its own) to evaluate VFM; and circumstances in which CBA may be fit for purpose to support an evaluation of VFM if used in conjunction with other methods.

Theoretical propositions

Theoretical propositions serve a formal purpose in the theory-building approach: the propositions specify descriptive qualities that highlight differences between the proposed model for evaluation of VFM, and CBA.

Yin (2009) noted that collectively, theoretical propositions are not expected to be fully ironed out by should aim to have "a sufficient blueprint for the study" and represent a "hypothetical story" (p. 36). While Yin's (2009) characterisation of theoretical propositions relate to causal or predictive theories (aimed at explaining how or why something occurs), the theoretical propositions here are aimed at understanding conceptual meanings (i.e., how and why the proposed model for evaluation of VFM differs from CBA). Each proposition directs attention to aspects that are explicitly examined within the scope of the case studies.

Theoretical propositions are identified through content analysis and triangulation of findings from the first two studies. First, the findings and key contributions of the first study are parsed into a series of propositions that summarise the logic of the argument presented in the published manuscript. Second, additional insights from the second study are added.

The propositions represent a concise summary of the conceptual model, and are used as a critical point of reference to inform the development of a process model, as follows.

RQ3: Process model

The results of the first two studies, and in particular the theoretical propositions summarised from those studies, reveal broad principles (Patton, 2017b) to guide the design of an evaluation of VFM of a social policy or program, but do not specify how an evaluator should go about putting those principles into practice. If such an approach is to be implemented and tested, a more detailed operational model is also needed.

A prototype is developed for an operational model. The field of implementation science devotes attention to the study of methods to promote the systematic uptake of knowledge into practice (Greenhalgh, 2018; Nilsen, 2015). Although principally used in evidence-based health care, translating research findings into clinical practice, the principles and insights from implementation science are transferrable to other settings. A *process model* is a model that specifies steps, stages or phases to guide the translation of research into practice. One type of process model is a *planned action* (or how-to-implement) model, which provides practical guidance and/or strategies for planning and implementing knowledge. Planned action models usually depict a number of stages or steps that should be followed in the process of implementation, as well as important aspects to consider in implementation practice. Although a series of steps may be represented as a rational, linear process, "authors behind most models emphasise that the process is not necessarily sequential" and highlight the importance of using models flexibly to respond to context (Nilsen, 2015, p. 4).

The development of process models often draws on the experiences of those who develop them, as well as literature reviews of theories, models, and frameworks identifying characteristics of, and considerations for successful implementation (Nilsen, 2015). This process of combining experiential and formal knowledge aligns with the theorising strategy of engaged scholarship (Shepherd & Suddaby, 2017), in which theoretical and practical knowledge is combined. This process of engaged scholarship requires the author to wear two metaphorical hats – the academic researcher/theorist, and the reflective practitioner. Accordingly, the conduct of the current research involves review of literature, while drawing on expert knowledge, and the author's experience in evaluative reasoning and economic evaluation. Each locus of knowledge influences the other.

First, a concept sketch of a process model was developed. The steps in the process comprised Fournier's (1995) four steps describing the general logic of evaluation, together with additional steps to further elaborate on the process of evaluation design, implementation and reporting (Davidson, 2005; 2014), informed by practical experience (King et al., 2013; McKegg et al., 2018; Wehipeihana et al., 2018) and the general literature on evaluative reasoning as already summarised.

In order to develop a process model that was sufficiently specific to apply and test in real-world settings, it was necessary to narrow the focus of the research to a specific sector and purpose. The operational model was developed for the purpose of evaluating VFM in international development programs funded by the UK Department for International Development (DFID). Accordingly, the model reflected the evaluation purposes driving VFM assessment in this sector (accountability and learning), and criteria of VFM stipulated by this organisation (economy, efficiency, effectiveness, cost-effectiveness and equity). Nonetheless, the model remained sufficiently general and flexible that it would be a relatively simple matter to incorporate additional or alternative criteria for use in other settings.

The model was subsequently applied in two real-world evaluations. These evaluations were carried out by the author (as lead consultant) in collaboration with other evaluators, and with the participation of management and technical experts from the respective programs. These evaluations provided the documentary data for case studies which were used to systematically test the applicability of the theoretical propositions, and to refine the operational model as described in the following section.

RQ4: Case studies

The prototype model for evaluation of VFM is investigated through its application in two case studies – evaluations of VFM in international development programs. The purpose of the case studies is to compare how

the proposed model of VFM assessment – the use of explicit evaluative reasoning to systematically determine the merit, worth and significance of resource use – differs from the use of CBA in practice. This is achieved by testing the applicability of core theoretical propositions of the model to these real-world settings, abstraction from the cases back to the theory, and by identifying experiential learning to inform refinements to the model.

The case study method is appropriate in settings where the researcher needs to develop or test a theory in the real world, when the phenomenon under investigation is complex (Løkke & Sørensen, 2014). Case studies are commonly used in such situations, spanning disciplines as diverse as psychology, sociology, political science, anthropology, social work, business, public administration, public health, education, accounting, community planning and, notably, in economics and evaluation (Baskarada, 2013; Yin, 2009).

The two case studies are drawn from different development contexts. In both cases, the donor – the UK Department for International Development (DFID) – required VFM assessments on at least an annual basis. The two case studies are summarised as follows.

Case study 1: MUVA: A program testing approaches to improving female economic empowerment with the aim of influencing empowerment of women in urban Mozambique (King & Guimaraes, 2016).

Case study 2: Pakistan Sub-National Governance (SNG) Program: Provincial and district government reforms in public financial management, needs-based budgeting, governance, and geographic information systems, aimed at improving democracy in Pakistan by improving people's access to basic services (King & Allan, 2018).

The cases were selected on the basis that they are revelatory cases (Yin, 2009) – that is, they involve a novel situation. These two cases were the first and second instances, respectively, in which the model of VFM assessment, described in this thesis, was applied in international development contexts.

In theory-building research, case studies can be used either for confirmatory (deductive) or explanatory (inductive) purposes (Baskarada, 2013; Wacker, 1998). In this instance, the case studies are applied deductively, for theory testing. The case studies serve three purposes. First, they provide concrete descriptions of the model's application in practice. The US Government Accountability Office (1990) used the term *illustrative* case studies to describe this function. Second, they serve as critical tests of the conceptual model by providing empirical data that are systematically assessed against theoretical propositions. This function has been described as *instrumental* (Stake, 1978) or *critical instance* case studies (US Government Accountability Office, 1990). Third, the case studies together enable replication to be explored. This makes the study design a *multiple case design* (Yin, 2009). A

multiple case design also provides the possibility to identify context-specific variations or contradictions to the theory, which may suggest areas where the theory could be modified to raise its abstraction level (Wacker, 1998).

Units of analysis

Case study research involves "intensive analysis of an individual unit" (Baskarada, 2013, p. 1) such as a person, community, or organisation. The unit of analysis defines what the 'case' is (Yin, 2009). In these case studies, the unit of analysis is an evaluation of VFM. In contrast to an experimental design, which separates a phenomenon from its context (controlling context in order to isolate the phenomenon), case studies seek to understand a real-life phenomenon in depth, within its specific contextual conditions (Yin, 2009). In these case studies, the phenomenon is the evaluation. The context is the program, including its setting, objectives, processes, outcomes, stakeholders, as well as the purpose, objectives and audience of the VFM assessment.

Each individual case study consists of a whole study, in which evidence is sought regarding the facts and conclusions for the case (Yin, 2009). Each case study has been written up to a common structure. First, the case study is presented, including a summary of the context, the program, the approach and methods used in the VFM assessment. Second, the applicability of the theoretical propositions to the real-world context is systematically assessed. Third, a more general discussion is provided, contextualising findings and considering implications for refinement of the theoretical and operational models.

Generalising from case studies to theory

Yin (2009) characterises the role of case studies in deductive theory-building as supporting a process of 'analytic generalisation' – a generalisation to a conceptually higher level than the case (Løkke & Sørensen, 2014). Under analytic generalisation, a theoretical model, developed in advance, is used as a point of comparison with the empirical observations and interpretations of the case study. In this research, a conceptual model, developed in advance, is investigated by assessing whether the anticipated differences between the model and CBA exist in real-world contexts. In this instance the analytic generalisation logic posits that if a difference is anticipated in theory, and subsequently observed in practice, then the specific observation corroborates the general theory.

Analytic generalisation requires judgements about whether the findings from the case study can be a guide to what will occur in another situation. These judgements are made based on comparison of empirical evidence to the theory, in order to corroborate, modify, or reject the theoretical concepts, and/or develop new concepts (Løkke & Sørensen, 2014). Stake (1978) argued that generalisations from case studies should be viewed as

expectations rather than predictions, as they fall short of empirical and logical tests required of formal scientific generalisations.

Analytic generalisation can be contrasted with statistical generalisation, which involves generalising results to a population or universe based on empirical data collected from a sample (Yin, 2009). Cases are not sampling units and cannot be used for statistical generalisation. Rather, through "the force of example and transferability", it is often possible to "generalize on the basis of a single case" (Flyvbjerg, 2011, p. 305).

In contrast to an experimental design, which studies selected variables while controlling for context, a case study "copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result, relies on multiple sources of evidence, with data needing to converge in a triangulating fashion, and as another result, benefits from the prior development of theoretical propositions to guide data collection and analysis" (Yin, 2009, p. 18).

Analytic generalisation can be used whether a study involves one case or multiple cases. Where there is more than one case, "individual cases are to be selected as a laboratory investigator selects the topic of a new experiment" – and "multiple cases, in this sense, resemble multiple experiments" (Yin, 2009, p. 38). Irrespective of the number of cases, the goal is to abstract from the cases back to the theory (a generalising analysis) rather than to enumerate frequencies (a particularising analysis) (Yin, 2009).

In these case studies, the theoretical propositions developed in the preceding studies are systematically assessed to determine the extent to which they are corroborated. If the case studies corroborate the propositions, they support analytic generalisation. If empirical evidence does not align with the propositions, then rejection or modification of the theory should be considered. This is supported by logic underpinning theory testing, where it is assumed that verification (of 'truth') is impossible apart from trivial facts (Lakatos, 1970) and that corroboration is not the opposite of falsification – rather it is part of a continual process of theory development and testing (Lakatos, 1970; Løkke & Sørensen, 2014).

Replication

This research contains two case studies – making it a multiple-case design (Yin, 2009). While each individual case is instrumental (Stake, 1978), the use of two cases together follows a replication design (Yin, 2009). The logic of replication is analogous to replication in multiple experiments (Yin, 2009); each case study comprises a whole study with its own context and conclusions, while the conclusions of each case may be replicated across cases. The results may be considered more potent if multiple cases support the same theory (that is, where replication occurs), and if the cases do not support an equally plausible rival theory (Yin, 2009).

The number of cases should be determined not on a statistical concept of sampling, but on the basis of theoretical sampling, where cases are chosen on the basis of theoretical criteria to judge the analytical benefits added by each new case. The greatest incremental benefit comes from increasing the number of cases from one to two, as it makes replication possible. Beyond that point, diminishing marginal returns may be expected, though this depends on the nature of the theory being investigated. Increasing the number of cases in the context of this research might allow some of the characteristics, logical implications, exceptions, or flaws in the model to emerge. Nonetheless, it remains that this research does not attempt to generalise statistically from empirical data; it empirically tests the logical implications of the theoretical propositions in the cases.

In this instance, case selection was straightforward; the two selected cases are the only two examples in which the model was applied to evaluate VFM. Additional evaluations had commenced at the time of this research but were ongoing and had not been completed. Nonetheless, given the nature of the theory, these two cases provide adequate replication for the first empirical test of the theory, with the theory remaining ever open to further research (Lakatos, 1970; Shepherd & Suddaby, 2017; Yin, 2009). The two case studies investigate the applicability of the theoretical propositions, to see whether the theory predicts similar results in both cases. This type of replication is called *literal replication*, in contrast to *theoretical replication* which predicts contrasting results for reasons that can be anticipated (Yin, 2009). The selection of two cases that are believed to be literal replications is "the simplest multiple-case design" (Yin, 2009, p. 59). With as few as two cases, direct replication is possible: "analytic conclusions independently arising from two cases, as with two experiments, will be more powerful than those coming from a single case (or single experiment) alone" (Yin, 2009, p. 61).

Documentary analysis

Case study summaries and analysis were conducted through documentary research – a form of qualitative research in which documents are reviewed and coded thematically (Ahmed, 2010; Miles, Huberman & Saldaña, 2014). The documents are reviewed not simply to record facts, but to source information about a phenomenon being studied (Ahmed, 2010). In documentary research, relevant documents can include any written material, such as public records, personal documents, or physical evidence, with the exception of records prepared specifically in response to a request from the investigator (Guba & Lincoln, 1981).

While the VFM assessments both used a mix of quantitative and qualitative methods, the data for the case studies is qualitative, focusing on the nature and context of the VFM assessments that were designed and executed in each case. The data for these case studies is sourced from reports and

background documentation on the programs (the context) and their evaluations (the phenomena), including business cases, program inception reports, monitoring reports, VFM frameworks, VFM assessment reports, peer reviewed journal articles, and email correspondence. Data sources for the two case studies are listed in Table 1.

Table 1: Sources of documentary evidence

MUVA female economic empowerment program, Mozambique	**Sub-National Governance Program, Pakistan**
DFID. (2017). *Annual Review of Ligada.* Maputo, Mozambique: UK Department for International Development. Hansford, F., & King, J. (November, 2016). *Ligada Formative Evaluation.* Oxford, United Kingdom: Oxford Policy Management Limited. King, J. (August, 2016). *Ligada Value for Money Framework.* Auckland, New Zealand: Julian King & Associates Limited. King, J. (February, 2017). *MUVA Value for Money Report 1: May-December 2016.* Oxford, United Kingdom: Oxford Policy Management Limited. King, J. (June, 2017). *MUVA Value for Money Report 2: January-April 2017.* Oxford, United Kingdom: Oxford Policy Management Limited. King, J., & Guimaraes, L. (2016). Evaluating Value for Money in International Development: The Ligada Female Economic Empowerment Program. *eVALUation Matters, the quarterly knowledge publication of the Africa Development Bank*. Third Quarter, 2016, pp. 58-69. King, J., & Wate, D. (May, 2018). *MUVA 2018 VFM Assessment.* Oxford, United Kingdom: Oxford Policy Management Limited. OPM (April, 2016). *Ligada Inception Report*. Oxford, United Kingdom: Oxford Policy Management Limited. Email correspondence and meeting notes (January, 2016 – May, 2018).	DFID. (2012). *Sub-national governance programme: Business case.* Islamabad, Pakistan: UK Department for International Development. DFID. (March, 2016). *SNG Annual Review.* Islamabad, Pakistan: UK Department for International Development. DFID. (March, 2017). *SNG Annual Review.* Islamabad, Pakistan: UK Department for International Development. King, J., Allan, S. (September, 2016). *Measuring results and value for money on the Pakistan Sub-national Governance programme.* Oxford, United Kingdom: Oxford Policy Management Limited. King, J., Allan, S. (2018). Applying Evaluative Thinking to Value for Money: The Pakistan Sub-National Governance Programme. *Evaluation Matters—He Take Tō Te Aromatawai* 4: 2018. OPM. (February, 2017). *SNG value for Money Report.* Islamabad, Pakistan: Oxford Policy Management Limited. OPM. (March, 2018). SNG value *for Money Report.* Islamabad, Pakistan: Oxford Policy Management Limited. Email correspondence and meeting notes (April, 2016 – March, 2018).

Evidence was drawn from these documentary sources to build profiles of the case studies and to extract the necessary evidence to assess the applicability of the theoretical propositions. Using a predetermined thematic structure, the

documents were analysed and coded for content relevant to the case study purposes. Written case studies were prepared summarising the program, the VFM evaluation design, and the VFM assessments carried out. Following this introductory information, each case study was systematically assessed against the theoretical propositions (these propositions are summarised at the conclusion of the gap analysis chapter). For each proposition, commentary was provided on the extent to which the proposition was observed (or not) within the case, the limitations of the case in demonstrating the proposition, and a summative rating of whether, and to what extent, the case corroborates or does not corroborate the sub-proposition (Table 2). The findings from the two case studies were synthesised to determine the extent to which findings were replicated across the two assessments.

Table 2: Rubric for coding propositions in case studies

Rating	**Definition**
Case corroborates sub-proposition	Within the context of the individual case, the evidence supports a conclusion that the proposition is true in this specific instance.
Case does not corroborate sub-proposition	Within the context of the individual case, the evidence does not support a conclusion that the proposition is true in this specific instance.
Case refutes sub-proposition	Within the context of the individual case, the evidence supports a conclusion that the proposition is false.
Proposition not tested in the case study	The individual case does not enable this proposition to be assessed.

Assuring quality of the research design

Four tests have commonly been used to establish the quality of empirical social research, including case studies: construct validity; internal validity; external validity; and reliability (Yin, 2009).

Construct validity, in the context of this case study research, relates to the specification of a set of constructs (the theoretical propositions) with sufficient clarity to make them evaluable through empirical evidence, and the conduct of VFM evaluations with fidelity to the process model. To meet these requirements, the theoretical propositions are disaggregated into a number of sub-propositions, providing sufficient detail to define the conceptual model and its rationale. Assessment of the propositions in the case studies is systematic and transparent. Multiple sources of evidence have been used, in a manner encouraging converging lines of inquiry (Yin, 2009). A chain of evidence has been maintained so that conclusions can be traced back to the evidence and theoretical propositions. Case study reports were independently reviewed by personnel involved in each evaluation to validate fidelity to the model.

Internal and external validity are most literally applicable to appraising the quality of statistical causal inferences, though the underlying meaning of these concepts has been applied to the current research as follows.

Internal validity, in this instance, is concerned with whether the case study evidence is consistent with the theoretical propositions. Pattern matching is used to systematically compare the empirical pattern with the theoretical propositions, in order to judge the extent to which the evidence in each case corroborates or does not corroborate the propositions. In doing so, triangulation and convergence/divergence of evidence are investigated – including evidence for the 'rival theory' that CBA is the gold standard for evaluating VFM.

External validity, in this research, is concerned with defining the domain to which the case study findings can be generalised (Yin, 2009). The principal strategy used is a two-case design based on replication logic, providing analytic generalisation from the results of the two cases to the broader theory. Corroboration is presented as part of an ongoing process of theory = development and testing (Lakatos, 1970); an interim contribution to an ongoing struggle of theorising (Shepherd & Suddaby, 2017) that remains open to future research.

Reliability is concerned with demonstrating that the study procedures can be repeated (for the same cases) with the same results (Yin, 2009). To meet this criterion, the case study methods and procedures are documented. Theoretical sub-propositions are used systematically and transparently to assess evidence. Evidence is primarily drawn from program documentation and is therefore auditable. All relevant information is presented within the case study. The independent review of case study reports provides inter-rater reliability of the conclusions reached.

Together, the strategies described above, used to assure quality of the research design, serve to address a criticism that has sometimes been levelled at case studies that they risk bias toward verification of the researcher's preconceived notions. Flyvberg (2011) argued that well-designed and executed case studies are no more biased toward verification than other methods, noting that: "On the contrary, experience indicates that the case study contains a greater bias toward falsification of preconceived notions than toward verification" (p. 311).

Theory refinement

In addition to formally testing the theoretical propositions, the case studies provide a source of real-world experiential data to support critical reflection and refinement of the theoretical and/or operational models. In alignment with the methodology of pragmatic empirical theorising and the theorising strategy of engaged scholarship (Shepherd & Suddaby, 2017), insights from

the conduct of the two case studies are captured to inform post-hoc theory refinement.

The author facilitated a reflective workshop. The workshop was conducted at the head office of Oxford Policy Management in September 2017. Present at the workshop were evaluators from the MUVA and SNG programs. Also present were evaluators involved in applying the process model in further programs, where evaluations had commenced and were in various stages of design/implementation progress. In the two-hour workshop, participants considered four questions: What has been learned from the use of the model? What has worked well? What hasn't worked well or has been challenging? How can the model be refined to support future practice? Notes were taken during the workshop, and key themes were identified.

A draft document was prepared, describing the stepped model and incorporating insights from the two VFM evaluations and related practice-based learning. The model was peer reviewed by colleagues from the two evaluation teams. Reviewers were asked to provide feedback on the validity and clarity of the document, and to identify any specific concerns or areas where the model should be further developed or more clearly explained. Feedback from the reviewers was reflected in the document. The resultant document, published online and providing guidance in the use of the process model to evaluate international development programs, is not a formal part of this thesis but by way of an epilogue, a link to the document is provided in an Appendix (King & OPM, 2018).

Ethics

Ethical approval was obtained from the Melbourne Graduate School of Education Human Ethics Advisory Group.

The case studies make use of secondary data from the MUVA and SNG programs. Written permission was obtained from Oxford Policy Management Ltd (OPM), the organisation delivering the programs, to use the two programs and their VFM assessments as case studies. OPM, in turn, obtained permission from DFID, the UK government agency that funded both programs. All feedback from participants has been used and reported anonymously.

Beyond the participation of colleagues from OPM, the research did not involve human subjects and did not involve any related ethical issues such as informed consent, participant safety or protection and storage of personal data.

Chapter 4: Conceptual model

Introduction

This chapter addresses the first research question:

RQ1: What are the requirements of a model for evaluating value for money in social programs?

This research interrogates the literature to address questions about what VFM means, how VFM can be evaluated, and to propose how VFM should be evaluated. Economic methods of evaluation and explicit evaluative reasoning exist in parallel disciplinary universes. This study compared the two universes, exploring their potential fit through the prism of a common interest: evaluation of VFM. The product of this study was a peer reviewed journal article. The author-accepted manuscript is reproduced here in full.

This study involved building a logical argument through critical review of published literature, in combination with various theory-building strategies as described earlier (Shepherd & Suddaby, 2017). From the literature, criteria were identified that would lay the foundation for a conceptual model by specifying and justifying the minimum requirements of an evaluation of VFM. The resulting conceptual model proposes core elements of an approach to combining explicit evaluative reasoning with economic evaluation.

Findings

Building on the background literature around VFM, economic evaluation and evaluative reasoning, this study made a novel, significant contribution to the discourse on VFM by proposing a definition and a set of requirements for VFM evaluation. It explored the validity of CBA as an approach to evaluative reasoning and found that, while CBA has distinct advantages that support its use, it is too restrictive to necessarily be appropriate for evaluating VFM in social policies and programs. Other forms of evaluative reasoning were explored and it was concluded that qualitative weighting and synthesis could be used to combine insights from CBA with those from other methods of evidence acquisition. The following article is a contribution to theory on what VFM means, and how it should be evaluated.

Article

American Journal of Evaluation
1-13

DOI: 10.1177/1098214016641211
aje.sagepub.com
SAGE

Using Economic Methods Evaluatively

Julian King[1,2]

Abstract

As evaluators, we are often asked to determine whether policies and programs provide value for the resources invested. Addressing that question can be a quandary, and, in some cases, evaluators question whether cost–benefit analysis is fit for this purpose. With increased interest globally in social enterprise, impact investing, and social impact bonds, the search is on to find valid, credible, useful ways to determine the impact and value of social investments. This article argues that when addressing an evaluative question about an economic problem (the merit, worth, or significance of resource use), economic analysis can enhance evaluation but is usually insufficient to fully answer the evaluative question and that a stronger approach would involve explicit evaluative reasoning, supported by judicious use of economic and other methods. An overarching theoretical framework is proposed, and implications for evaluation practice are discussed.

Keywords

values, economics, cost benefit, evaluation theory, mixed methods

> An economist is [someone] who, when [s]he finds something works in practice, wonders if it works in theory.
>
> (Walter Heller, 1979, cited in Rivkin, 2015)

Introduction

There are increasing calls for evaluation to address questions of value for money (VFM) in social programs (Dumaine, 2012; Julnes, 2012; Svistak & Pritchard, 2014; Yates, 2012). In the last few decades, cost–benefit analysis (CBA) has come into widespread use to weigh the net value of public policies, taking into account both resources used and their consequences, valued in monetary terms. In the United States, for example, its use in the regulatory process has been enshrined in executive

[1] Centre for Program Evaluation, University of Melbourne, Melbourne, Victoria, Australia
[2] Julian King & Associates Limited, a member of the Kinnect Group, Auckland, New Zealand

Corresponding Author:
Julian King, Julian King & Associates Limited, a member of the Kinnect Group, PO Box 41-339, St. Lukes, Auckland 1346, New Zealand.
Email: jk@julianking.co.nz

orders since the Reagan Administration (Adler & Posner, 2006). More recently, with growth in social investment initiatives in the nonprofit, philanthropic, and private sectors, the notion of return on investment has become increasingly tied to social change. In this context, the application of CBA constructs in valuing social investments has gained attention. Examples of such methods include social return on investment and social CBA (Arvidson, Lyon, McKay, & Moro, 2010; Svistak & Pritchard, 2014). No universal definition of VFM exists, however, and there are different views about how it should be evaluated (Levin & McEwan, 2001; Pinkerton, Johnson-Masotti, Derse, & Layde, 2002; Tuan, 2008; Yates, 2009).

Although evaluation and economics share an interest in determining how well resources are used, the two disciplines tend to approach evaluation and valuing in distinct ways. As a consequence, economic evaluation tends to be applied either in isolation from or in parallel to other methods of evaluation. This gap has been noted within the evaluation community, and it has been suggested that economic and other valuing methods should be better integrated (Davis & Frank, 1992; Julnes, 2012). An overarching framework is needed to guide the use of appropriate methods to determine the value of social investments.

This article defines "value for investment" (VFI) as an evaluative question that connects the evaluative concepts of merit, worth, and significance to the fundamental economic problem of resource allocation. It then assesses whether economic methods of evaluation represent a universal "gold standard" for addressing this question. Finding that they do not, a case is argued that economic methods would be more appropriately situated within an overarching framework of evaluative reasoning. Implications for evaluation practice are discussed.

A Scenario

Recovery House, a fictitious nongovernment organization that provides community-based alcohol and drug (AOD) rehabilitation services, has just opened a social enterprise café on the ground floor of its largest service facility. The principal objective of the café is to contribute to the recovery of young clients by building their employment skills and experience, enhancing their future job prospects and well-being. At the same time, as a social enterprise, it is intended that the café will earn a healthy profit (i.e., its income will exceed its costs) and that Recovery House will reinvest the profit to enhance its rehabilitation services. It is also intended that the café will help to forge connections with the local community, providing a social hub where staff, clients, families, and the general public can meet each other. It is hoped that this will boost morale of staff and clients, enhance the Recovery House brand, and help to destigmatize AOD addiction.

There are some risks, however. Some clients might feel uncomfortable knowing that members of the public might see them visiting the service, and this might discourage them from attending treatment. Also, some clients have personality traits that make them prone to emotional outbursts. A violent incident in the café is considered low risk but, if it did occur, could have a significant negative impact on public perceptions of Recovery House. The Board of Recovery House has asked for an evaluation of the café, including whether it provides VFM.

An Evaluative Question About an Economic Problem

The imperative to consider VFM arises because, when resources are invested in a particular policy, program, or intervention, the opportunity to use those resources in another way is foregone. Economists call this loss of alternatives *opportunity cost* (Drummond, Sculpher, Torrance, O'Brien, & Stoddard, 2005; Levin & McEwan, 2001). Consequently, choices need to be made in resource allocation—with a "good" choice being one that compares favorably to the next-best alternative use of resources. What constitutes a good choice, however, is a matter of context and perspective.

Academic, governmental, philanthropic, and international development publications use a variety of working definitions of VFM that, depending on context, may be concerned with minimizing costs, achieving outcomes (e.g., effectively meeting identified needs), equitably distributing resources or outcomes, and/or maximizing outcomes for a given cost—with priority often given implicitly or explicitly to the last of these (Drummond et al., 2005; Dumaine, 2012; Fleming, 2013; Levin & McEwan, 2001).

In the absence of an all-encompassing definition for these various concepts, the generic term VFI, defined as *the merit, worth, or significance of resource use*, is proposed. Building on Scriven (2013), working definitions of the merit, worth, and significance of resource use are proposed as follows. *Merit* refers to the quality of the resource use, for example, using funds for their intended purpose, using funds ethically, minimizing wastage, and achieving intended outcomes. *Worth* refers to the value of the resource use (to a person, group, or society, at a particular time and place) relative to the next-best alternative use of resources. *Significance* refers to the importance of the resource use, beyond its merit and worth. For example, a social program may have low merit and worth (perhaps it is only moderately effective and quite costly) but may be viewed as an entitlement or may be seen as significant in being the only program meeting a particular need for a vulnerable group in society.

The ''investment'' in a social program often includes nonmonetary resources such as time, expertise, and relationships, as well as monetary resources. Similarly, the value derived from an investment may be monetary and/or nonmonetary. Accordingly, the term VFI is proposed in place of VFM in order to draw attention to the underlying value to society of all resources invested. Money, whether in its everyday sense as a medium of exchange, a unit of account or a store of value in an economy, or in its more rarefied sense as a proxy for things that matter, is simply a way of representing underlying value (Nicholls, Lawlor, Neitzert, & Goodspeed, 2012; Svistak & Pritchard, 2014).

The proposed definition of VFI connects the evaluative concepts of merit, worth, and significance to the fundamental economic problem of resource allocation. It frames VFI as an evaluative question about *how well* resources are being used, and whether they are being used *well enough* to justify their use in a particular way. This evaluative question is often addressed using economic methods. The following section describes economic methods of evaluation and considers their strengths and limitations in terms of their capacity to address evaluative questions about VFI.

Economic Methods of Evaluation

Economic evaluation offers a powerful set of methods for considering the costs and consequences of resource allocation decisions. All of these methods involve systematically identifying, measuring, valuing, and comparing the costs and consequences of alternative courses of action (Drummond et al., 2005). All the methods yield indicators of efficiency but differ in their scope and the units of measurement used. Three key examples of economic methods are cost-effectiveness analysis (CEA), cost-utility analysis (CUA), and CBA.

CEA measures costs in monetary terms and consequences in natural or physical units—usually, a single quantitative outcome measure with a strong counterfactual claim, such as years of life saved by a health intervention (Levin & McEwan, 2001). The output of a CEA is a *cost-effectiveness ratio* (e.g., the average cost per year of life gained). Often, an *incremental cost-effectiveness ratio* is calculated that compares the *additional* costs and consequences of an intervention to its next-best alternative (Drummond et al., 2005). For the Recovery House café, for example, CEA could examine the cost per young person who gains a defined set of employment skills in the café, compared to a traditional work skills program.

CUA is closely related to CEA in terms of its underlying structure but broadens the valuation of consequences to incorporate the notion of utility to people, which may include multiple attributes

(Von Neumann & Morgenstern, 1947). For example, empirically derived measures such as quality-adjusted life years and disability-adjusted life years scale the "raw" measurement of extended life spans to take into account the utility of those additional years (Drummond et al., 2005). For the Recovery House café, CUA could incorporate a composite index of outcomes weighted for their relative importance (Levin & McEwan, 2001)—for example, such an index might accommodate the utility of improvements in young people's employability, ongoing recovery from addiction, and community perceptions of the quality of the café experience. The *cost-utility ratio* would be the average cost per unit of improvement in the index value.

CBA values all costs and consequences in the same units, which are usually monetary. For example, when evaluating the Recovery House café, a CBA would include not only the financial costs and benefits of the enterprise but also the monetized value of outcomes such as the café's contribution to client employment and well-being, staff and client morale, community cohesion, and destigmatizing AOD addiction. In practice, some of these outcomes might be difficult to value in monetary terms; however, in principle, all values should be included and there are well-established economic valuation methods to support this approach (Levin & McEwan, 2001). The output of a CBA can take various forms such as net value (benefits minus costs), benefit cost ratio (benefits divided by costs), or return on investment (net value divided by costs; Drummond et al., 2005; Levin & McEwan, 2001).

Strengths of Economic Evaluation

Economic methods of evaluation offer a number of benefits that can enhance the evaluation of social investments. Systematically evaluating costs and consequences yields insights that cannot be gained by looking at either factor in isolation. For example, two alternative interventions may be equally effective in terms of a measured effect size, while differing markedly with regard to their costs. A decision to include or exclude costs from an evaluation may therefore result in different conclusions being reached about the relative worth of the two interventions (Drummond et al., 2005; Levin & McEwan, 2001).

Economic evaluation often involves modeling (creating a simplified representation of a system to facilitate analysis and understanding) and may also include forecasting (estimation of future value) as well as measurement of past performance. Both modeling and forecasting introduce assumptions and uncertainty. A strength of economic evaluation is the ability to apply *scenario analysis* (exploring the net value of a program under different sets of assumptions or circumstances) and *sensitivity analysis* (exploring the extent to which changes in a particular input variable affect the outputs of the model). Sensitivity and scenario analysis facilitate transparency and robust thinking about relationships between benefits and costs, taking uncertainty into account, which can lead to insights that could otherwise be difficult to gain (King, 2015). For example, in addition to cost-effectiveness ratios, probabilistic and statistical analysis together with graphs showing more complete and possibly more complex relationships between costs and consequences have been used in program evaluation for several decades.

Among the economic methods, CBA is often regarded as the gold standard when it comes to evaluating resource use (Julnes, 2012). It is the ability to value costs and benefits in commensurable units, and therefore to easily reconcile them in the final synthesis, that sets CBA apart from other methods and is one of the key reasons for its appeal as a powerful approach to estimating the worth of a policy or program. From an evaluative perspective, this article argues that CBA (and by extension, less comprehensive methods such as CEA and CUA) is likely to enhance evaluation of VFI but is often insufficient to fully evaluate VFI. CBA has a number of limitations that preclude the method from providing a full and comprehensive evaluation of investments in social change. These

include inherent limitations of the method itself as well as limitations in its practical application. Key examples of both sets of limitations are outlined as follows.

Inherent Limitations of CBA

Like any method, CBA has a number of embedded assumptions and values that are not always made explicit. Unless these limitations of CBA are understood, there is a risk of reaching incomplete or invalid conclusions about VFI. First, CBA provides an estimate of economic efficiency (Sinden, Kysar, & Driesen, 2009), which is a criterion of worth. It cannot provide a full evaluation of VFI because it does not explicitly consider the merit or significance of resource use. Economic efficiency is likely to be an important criterion of VFI in many cases but may not be the only criterion—it is "morally relevant, but not necessarily morally decisive" (Adler & Posner, 2006, p. 154). Therefore, CBA on its own may not offer sufficient breadth to provide a full evaluation of VFI.

Second, CBA reflects a particular set of premises about economic efficiency, based on normative values from welfare economics (Drummond et al., 2005). For example, the final synthesis in a CBA rests on warrant claims that all things that matter should be valued in commensurable units, that the actual or hypothetical behavior of markets adequately represents the value of anything, that competing alternatives can be compared and ranked according to their value, and that any net gain in overall value is worthwhile (even if it creates winners and losers). While this may be appropriate some of the time, it is too restrictive a basis to represent a universal approach to the evaluation of VFI.

Third, the aggregation of values using a common metric, though advantageous from the perspective of determining net utility, can diminish the visibility of qualitative differences between values held by different people or groups (Adler & Posner, 2006) and imply a value judgment that consensus is desirable and possible (Julnes, 2012). If an evaluation is being conducted in a context marked by power imbalances or diverging worldviews, values or interests, aggregating those values may serve to diminish clarity rather than enhance it (House & Howe, 1999).

Fourth, the aggregation of diverse things of value can disguise qualitative differences between them (Adler & Posner, 2006). For example, building a new highway can help save lives and move goods to market more quickly. If two alternative highway projects have the same net value, but one will save more lives and the other will have a greater impact on reducing driving times, this qualitative difference may be considered important notwithstanding the commensurable valuations used in CBA.

Fifth, CBA reflects a consequentialist perspective, which assumes program processes are only relevant to the extent that they achieve outcomes, and so are fully and adequately represented by the value of program outcomes. As Julnes (2012) points out, processes may in fact be relevant to the value of a program or policy independently from outcomes, and indeed the ultimate outcome may even *be* a process (e.g., ongoing development or "human flourishing").

Practical Limitations of CBA

In addition to the conceptual issues presented above, a number of practical challenges that can lead to bias in CBA estimates are possible. These limitations may, in principle, be addressed through professional standards and methodological innovation—but, nevertheless, may have a bearing on decisions about when and how to use CBA in evaluation.

One source of bias stems from the possible exclusion of some nonmonetary values. Although CBA can theoretically accommodate all things of value, it has been found that "CBA analysts, in practice, ignore welfare dimensions [that] are just too hard to estimate given current techniques" (Adler & Posner, 2006, p. 78). This is critical because the exclusion of particular benefits or costs can affect whether the net valuation obtained through CBA is positive or negative, and therefore

whether the investment is considered worthwhile or not. In social programs, some of the most valuable benefits may be the hardest to monetize. For example, in the social enterprise café, benefits that are hard to value in monetary terms include impacts on client well-being, community cohesion, and reducing stigma associated with addiction. Exclusion of any of these benefits would result in underestimation of the overall value of the café.

A second practical challenge is related to construct validity problems in monetary valuation (Drummond et al., 2005; Julnes, 2012; Pinkerton et al., 2002). For example, the use of market values in a CBA model rests on the assumption that a valuation set by a particular market for a particular purpose has sufficient external validity to represent the value of a similar input or outcome in the analysis. Does the market wage for café staff adequately represent the value of employment for the young people working in the café? What if the market wage is higher in another city—does that mean employment outcomes are more valuable there?

A third set of challenges concern the ability to derive a determinate estimate of value using monetary valuation techniques. One example of a determinacy problem is the *endowment effect*, in which people's "Willingness to Accept" (payment in compensation to forego a benefit they already have) can differ widely from their "Willingness to Pay" (for the same benefit if they don't already have it). The decision to accept or reject a project based on CBA can depend on which valuation is used (Adler & Posner, 2006; Sinden et al., 2009).

Due to these and other practical challenges, results of a CBA may be subject to significant uncertainty. More insidiously, the input variables in a CBA can be manipulated to obtain the desired results (Sinden et al., 2009). Such studies are often not peer reviewed and, though they may appear to be a genuine attempt to determine the worth of an intervention, can in fact be constructed to support a predetermined course of action. Similar criticism can, of course, be leveled at any evaluation method, but the technical complexity of CBA, nevertheless, introduces the risk that the findings can be misused by people with inadequate knowledge of CBA methods and assumptions.

CBA Is Not a Gold Standard

The conceptual and practical limitations identified above, though not exhaustive, are sufficient to demonstrate that CBA is not a gold standard for evaluating VFI. Like any method, CBA accentuates some considerations while reducing the visibility of others. In particular, CBA is only relevant where economic efficiency is a criterion of interest and is insufficient on its own if there are other criteria to consider beyond economic efficiency.

When evaluating the merit, worth, or significance of resource use in social investments, there may be values that need to be considered besides those inherent in a CBA.

Adler and Posner (2006) similarly find that "CBA is not a superprocedure" (p. 157) and argue that an overarching approach is needed that can incorporate other considerations alongside efficiency. They concede: "We suppose that that is a theoretical possibility—but we have absolutely no idea what the superprocedure would consist in" (Adler & Posner, 2006, p. 158). Although they outline a typology of alternative procedures including "intuitive balancing" and "balancing with explicit trade-off rates," the two law professors conclude "none of the competitors ... seem remotely plausible as this sort of superprocedure" (p. 158). The remainder of this article proposes that explicit evaluative reasoning is the superprocedure that eluded Adler and Posner (2006) and holds the key to using economic methods evaluatively.

Evaluative Reasoning and Economic Methods

Many evaluators accept the proposition that systematically addressing evaluative questions should involve explicit evaluative reasoning; indeed, this is specified in evaluation standards (e.g., Yarbrough, Shulha, Hopson, & Caruthers, 2011). CBA represents a form of evaluative reasoning,

but one that often requires augmentation to evaluate the merit, worth, and significance of resource use. These issues are explored below.

The General Logic of Evaluation

Evaluation, according to Scriven (1980, 1995, 2012), is definitionally underpinned by a specific process of reasoning involving the use of criteria and standards to reach conclusions about merit, worth, or significance from empirical evidence. Under this model of evaluative reasoning, the logic of valuing (Scriven, 2012) provides justification for criterial inference, a subtype of probative inference, while the general logic of evaluation describes the use of criterial inference to reach evaluative conclusions about the merit, worth, or significance of an evaluand (Scriven, 1980, 1994, 1995, 2012).

The general logic simplifies what is in reality a complex and contextually driven process to explicate an underlying logic for evaluative reasoning (House & Howe, 1999). It has been described as "philosophically strong, providing good conceptual structure for the field" (Shadish, Cook, & Leviton, 1991, p. 117) and, in the view of many evaluation theorists, underpins all evaluation—although the extent to which criterial inference might guide, complement, or supplant more intuitive judgment is debated (Stake et al., 1997).

In generic terms, Yarbrough, Shulha, Hopson, and Caruthers (2011) summarized the use of evaluative reasoning as involving a series of "if-then" statements (the "if" part describing a combination of assumptions, evidence, and criteria that must be met and the "then" part describing the conclusion that would be reached). Fournier (1995) described four steps in implementing the general logic of evaluation: (i) establishing the criteria of merit or worth, (ii) defining the performance standards (e.g., the difference between "excellent," "good," "acceptable," and "poor" performance), (iii) gathering and analyzing the evidence of performance against the standards, and (iv) synthesizing the results into an overall judgment. The steps described by Fournier (1995) should be regarded as a stylized description rather than a formulaic prescription; for example, they may be applied iteratively rather than sequentially (Scriven, 1992). Furthermore, there are other ways of implementing the general logic of evaluation—and the process of conducting a CBA is one example.

CBA and the General Logic of Evaluation

CBA can be conceptualized as a way of implementing the general logic of evaluation through quantitative valuing and synthesis, in that it involves identifying the things of value, including resource use and impacts; quantifying them; valuing them through a process of numerical weighting (with the weights usually, but not necessarily, being monetary); and synthesizing the evidence by aggregating the weighted values (positive and negative) to reach an overall determination of net value. Viewed in this way, CBA is a sophisticated example of a general approach that Scriven (1991, 1994) called *Numerical Weight and Sum*.

While monetization provides the principal basis for weighting the relative values of different costs and benefits in CBA, additional weights are also applied. Foremost among these, the values of costs and benefits are adjusted according to when they occur in time—a technique economists call *discounting* (Levin & McEwan, 2001). Additionally, probability weights are sometimes used as a way of adjusting estimates for uncertainty or risk. For example, the risk of a violent incident in the café could be incorporated in a CBA by multiplying the estimated cost of the incident if it did occur by the estimated probability of it occurring (with suitable quantitative estimates being informed by a literature review).

The weighting and synthesis of costs and benefits in this way is a systematic quantitative approach to applying the general logic of evaluation. It is not comprehensive, however, for the reasons already outlined. Therefore, something extra is needed to incorporate any additional

dimensions of the merit, worth, and significance of resource use. One set of options involves retaining the quantitative approach to valuing and synthesis, with the addition of normative weights. Another is to use qualitative valuing and synthesis, with an ordinal scale for weighting the relative importance of dimensions of merit, worth, or significance. These two approaches are summarized as follows.

Quantitative Valuing and Synthesis

The principal advantage of retaining quantitative valuing and synthesis as the overarching guide to evaluative reasoning is the ability to derive a net valuation through the use of a single valuation metric. These advantages could perhaps be retained if costs and benefits were reweighted to accommodate additional considerations. For example, CBA, by default, is agnostic about distributional values. Any allocation of resources or outcomes that maximizes aggregate value (for a particular purpose, at a particular time and place) represents the "preferred option" in a CBA. In any circumstances where one distribution would be considered more desirable than another on social justice grounds, this consideration would need to be balanced alongside efficiency. This is likely to be a requirement in most social investments—indeed, some argue that "ethical concerns (such as equity and fairness) are critical to any discussion of real-world resource allocation decision making" (Pinkerton et al., 2002, p. 80).

One way to address distributional issues in a CBA is quantitatively through the use of distributional weights. Returning to the scenario of the social enterprise café, it may be that Recovery House wants to improve equity of outcomes by working with hard-to-reach clients who have severe addiction issues. Perhaps working with such clients involves much more intensive work, but only achieves a slightly greater reduction in the social costs of addiction, compared to working with "easier" clients who have mild-to-moderate addiction. A basic CBA might therefore conclude that working with hard-to-reach clients was less efficient than working with easier clients. To accommodate the normative goal of improving equity of outcomes for hard-to-reach clients, the CBA could be reweighted, so that outcomes for hard-to-reach clients were valued at a higher rate than outcomes for other clients.

The use of this approach would require a sound basis for determining and justifying appropriate weights. Ideally, all of the values in a CBA, including distributional values, should be determined empirically. In the absence of direct measurement and curve fitting, the determination of weights becomes more arbitrary, leading Scriven (1991) to conclude that although quantitative valuing and synthesis is "sometimes approximately correct, and nearly always clarifying" (p. 380), it can lead to fallacious conclusions that are not always traceable. Adler and Posner (2006) noted "economists have as yet failed to develop readily implementable schemes for distributive weighting" (p. 188). Quantitative valuing and synthesis with normative weights might be feasible in some circumstances, but a wider set of options may be needed for other evaluation contexts. An alternative approach (that avoids the risk of bias arising from the use of inappropriate quantitative weights) is qualitative valuing and synthesis.

Qualitative Valuing and Synthesis

Qualitative approaches to valuing and synthesis represent a flexible set of approaches to implementing the general logic of evaluation. They involve ordinal scaling and can accommodate holistic or analytic evaluation, using absolute (grading, rating, scoring) or relative (ranking, apportioning) systems for determining merit, worth, and significance together with quantitative and/or qualitative synthesis. Scriven (1991) outlined an approach of this form, which he titled *Qualitative Weight and Sum.* Davidson (2005) proposed the use of rubrics as an ordinal framework to guide the process of valuing and synthesis.

If situated within this type of approach, economic methods would provide evidence (which might include ratio or interval measurements as well as qualitative insights gained through the process of economic analysis) contributing toward an overall judgment of the merit, worth, or significance of resource use. What might this look like in practice? A multitude of evaluation designs are possible, and it would be inappropriate to prescribe one approach here. Nevertheless, it is worth setting out an example.

For this example, consider once more the Recovery House social enterprise café. Suppose evaluators were able to determine, in consultation with key stakeholders, that the most important considerations in determining whether the café provides VFI are employment and well-being impacts for hard-to-reach clients, community perceptions of the café and Recovery House brand, reducing stigma associated with addiction, and the financial result (size of profit or loss). Further, while earning a profit is part of the Board's long-term expectations, they accept that costs may exceed income in the short term while establishing the café and that the café might even run at a modest loss for the medium term and still be worth operating if sufficiently valuable outcomes are being achieved for clients.

Taking these factors into consideration, perhaps it would be agreed that the café would represent *poor* VFI if it achieved little or no employment or well-being impacts for hard-to-reach clients, if it had any negative impact on community perceptions of Recovery House or its clients, or if it ran at an unaffordable financial loss with no reasonable prospect of revenues covering costs at any time in the future. Conversely, the café would represent *acceptable* VFI if it were found that the café supported hard-to-reach clients to improve their employment prospects and well-being and was generally viewed positively by community stakeholders—even if it made a modest and affordable financial loss, provided its financial performance indicated it could realistically earn a profit in the longer term.

Building on these concepts, perhaps the café would be considered *good* VFI if its effects on employment and well-being were at least as substantial as comparable work experience programs, if community feedback indicated it was contributing positively to perceptions about Recovery House and reducing stigma about addiction, and if it was earning a profit that was at least as good as the next-best alternative use of resources (e.g., the rental income that Recovery House could earn by leasing the floorspace to a commercial café operator). Finally, it might be determined that the café would represent *excellent* VFI if, in addition to achieving substantial impacts for hard-to-reach clients, being a recognized and valued contributor to community well-being and reducing stigma, it was also earning a profit of sufficient magnitude to meaningfully be reinvested in developing new programs for Recovery House clients.

If these standards were used as a guide to explicit evaluative reasoning, then in this scenario, financial performance would be analyzed using accounting methods, while other outcomes could be measured using a mix of other methods such as clinical assessments, surveys, interviews, focus groups, and possibly economic analysis. This type of approach offers a way to combine economic and other methods within an overarching framework of explicit evaluative reasoning. By positioning qualitative valuing and synthesis as a superprocedure, this approach can incorporate economic analysis when feasible and appropriate as well as accommodating other quantitative and/or qualitative methods. As a result, many of the advantages of economic evaluation are retained (e.g., systematic identification and valuation of both costs and consequences, scenario and sensitivity analysis, indicators of efficiency) while also allowing for more deliberative approaches to addressing wider considerations such as distributional impacts, conflicting perspectives of different groups, qualitative differences between things of value, ethical bottom lines, process issues, and the flexibility to value things that matter in monetary or other terms according to context.

Qualitative valuing and synthesis, therefore, offers a procedure for implementing the general logic of evaluation that is sufficiently flexible to guide evaluative reasoning in any setting where the objective is to evaluate the merit, worth, or significance of resource use. In certain contexts,

economic analysis could contribute some of the evidence needed to make a determination of merit, worth, or significance of resource use. Furthermore, the use of qualitative valuing and synthesis raises the possibility of using CBA in combination with other methods, as part of a mixed-methods approach to evaluating VFI. In the search for new ways of understanding the value of social investments, the answer may lie in the use of multiple methods, governed by explicit evaluative reasoning.

Implications for Practice

This evaluative model of VFI raises a number of implications for evaluation practice. First, the considerations above suggest a theoretical foundation for evaluating VFI in social programs. It is proposed that the minimum requirements for such an evaluation would be that it (a) poses, and answers, an evaluative question about the merit, worth, or significance of resource use; (b) uses explicit evaluative reasoning to reach an evaluative conclusion; (c) selects and tailors methods (economic and/or other) according to context; and (d) is conducted in keeping with program evaluation standards.

Second, the current separation of evaluation and economics is a missed opportunity. VFI links the evaluative concepts of merit, worth, and significance with the fundamental economic problem of resource scarcity. Yet the disciplines of evaluation and economics tend to operate as if competing or complementary disciplines. Few evaluators are trained in economic analysis, and resource use is rarely included in the scope of program evaluations (Herman, Avery, Schemp, & Walsh, 2009; Levin, 1987; Persaud, 2005; Yates, 2012). Conversely, economic evaluation is cost-inclusive but privileges economic efficiency and quantitative valuing, which may crowd out wider considerations (Julnes, 2012; Sinden et al., 2009). Both disciplines, and our capacity to evaluate VFI, are worse off as a result.

Third, there are definable circumstances in which economic methods are likely to enhance evaluation and circumstances where they are likely to be insufficient. Economic methods are applicable where economic efficiency is a relevant criterion and the warrant claim and values locus of the methods are compatible with the purpose and context of the evaluation. Additionally, it would be necessary, though perhaps not sufficient, that CBA provides a determinate estimate of economic efficiency (Adler & Posner, 2006; Sinden et al., 2009) and that the costs of data acquisition and analysis be justified (Levin & McEwan, 2001).

Economic methods of evaluation, on their own, cannot provide a determination of merit, worth, or significance when there are additional criteria to economic efficiency. Successive amendments to Executive Orders under the Clinton and Obama Administrations appear to acknowledge this, with provision now included for agencies to consider "values that are difficult or impossible to quantify, including equity, human dignity, fairness, and distributive impacts" (Executive Order No. 13563, 2011, p. 3821). Explicit evaluative reasoning offers a superprocedure to enable holistic conclusions to be reached taking all of these factors into account.

Fourth, if methods are to be matched to context in the evaluation of VFI, further guidance is needed to inform the selection of methods. Julnes (2012) sets out considerations for informing the selection of methods for "Assisted Valuation in the Public Interest" to support the goal of maximizing VFM in public sector initiatives. Key contextual factors are identified that might influence selection of methods of valuing—informed by "a pragmatic approach that acknowledges and defends the value of multiple approaches to valuing" (p. 109). These factors indicate that economic methods are better suited to some evaluation purposes, information needs, valuation needs, and social process needs, than others. A number of questions remain, particularly in regard to the selection and mixing of methods in an evaluation of the merit, worth, and significance of resource use. These areas are ripe for research and explication.

Conclusion

VFI poses an evaluative question about an economic problem. Addressing this question requires an understanding of resources invested, the value derived from their investment, and an evaluative basis for determining whether the investment represents good use of resources. Economic methods of evaluation offer a powerful set of techniques for conducting cost-inclusive evaluations, and these techniques are underutilized by evaluators.

The current separation of evaluation and economics is a missed opportunity to address the VFI question more comprehensively. As evaluators, we need to reach across disciplinary boundaries and make better use of economic methods. But more than that, we need to become more effective at addressing the VFI question through evaluative reasoning and deliberation, making use of economic and other methods.

CBA is often able to add to the knowledge base and information for making evaluative judgments—for example, by promoting systematic thinking and transparency about costs and consequences. In most evaluations of social investments, however, CBA alone offers too narrow a perspective for evaluating VFI. Other dimensions of value and approaches to valuing also need to be considered. These can be addressed through explicit evaluative reasoning, with methods (quantitative and/or qualitative) being matched to context. More guidance is needed on how this should be done.

The tensions between economic and evaluation-specific approaches to valuing have been likened to the "quant-qual" debate and the causal wars, in that both controversies involved opposing sets of worldviews in which one side maintained that a particular set of methods (quantitative data analysis and randomized controlled trials, respectively) represented a gold standard while the other argued that methods should be tailored to context (Davidson, 2006; Julnes, Schwandt, Davidson, & King, 2012; Scriven, 2008). In both cases, the latter sides' appeals to a higher order, overarching logic offered a basis for a set of principles framing the dominant methods as conditionally valid and sometimes appropriate contributors to evaluation, rather than being universally superior methods. Although these debates are not over, evaluators are at least armed with robust frameworks to design and defend context-appropriate methodologies.

There are no gold standards in evaluation, and CBA is no more a gold standard than is the randomized controlled trial. As with any evaluation method, it is important to consider when and how economic methods are constructed and applied in each case, as this can have a critical bearing on the conclusions reached. This article contends that a framework is needed to guide the judicious use of economic methods in evaluation, guided by explicit evaluative reasoning.

Author's Note

John Hattie, Janet Clinton, and Ghislain Arbour reviewed earlier drafts of this article and provided helpful feedback. Any errors or omissions are my own.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Adler, M. D., & Posner, E. A. (2006). *New foundations of cost-benefit analysis*. Cambridge, MA: Harvard University Press.

Arvidson, M., Lyon, F., McKay, S., & Moro, D. (2010). *The ambitions and challenges of SROI* (Working Paper 49). Birmingham, England: Third Sector Research Centre.

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

Davidson, E. J. (2006). The RCTs-only doctrine: Brakes on the acquisition of knowledge? *Journal of Multidisciplinary Evaluation*, *6*, ii−v.

Davis, K. E., & Frank, R. G. (1992). Integrating costs and outcomes. *New Directions for Evaluation*, *54*, 69−84.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddard, G. L. (2005). *Methods for economic evaluation of health care programs*. Oxford, England: Oxford University Press.

Dumaine, F. (2012). When one must go: The Canadian experience with strategic review and judging program value. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New Directions for Evaluation* (Vol. 133, pp. 65−75). San Francisco, CA: Jossey-Bass.

Exec. Order No. 13563, 3 C.F.R. Page 3821 (2011).

Fleming, F. (2013). *Evaluation methods for assessing value for money*. Retrieved from www.betterevaluation.org

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. In D. M. Fournier (Ed.), *Reasoning in evaluation: Inferential links and leaps. New directions for evaluation* (pp. 15−32). San Francisco, CA: Jossey-Bass.

Herman, P. M., Avery, D. J., Schemp, C. S., & Walsh, M. E. (2009). Are cost-inclusive evaluations worth the effort? *Evaluation and Program Planning*, *32*, 55−61.

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

Julnes, G. (2012). Promoting valuation in the public interest. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New directions for evaluation* (Vol. 133, pp. 109−129). San Francisco, CA: Jossey-Bass.

Julnes, G., Schwandt, T., Davidson, J., & King, J. (2012). *Valuing public programs and policies in complex contexts: Balancing multiple values, multiple cultures, and multiple needs*. Panel presentation, American Evaluation Association Conference, Minneapolis, MN.

King, J. (2015). Use of cost-benefit analysis in evaluation. Letter to the editor. *Evaluation Journal of Australasia*, *15*, 37–41.

Levin, H. (1987). Cost-benefit and cost-effectiveness analyses. In D. S. Cordray, H. S. Bloom, & R. J. Light (Eds.), *Evaluation practice in review. New directions for program evaluation* (Vol. 34, pp. 83–99). San Francisco, CA: Jossey-Bass.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Nicholls, J., Lawlor, E., Neitzert, E., & Goodspeed, T. (2012). *A guide to social return on investment*. Haddington, England: The SROI Network.

Persaud, N. (2005, February). Is cost analysis underutilized in decision making? *Journal of Multidisciplinary Evaluation*, *2*, 81–82.

Pinkerton, S. D., Johnson-Masotti, A. P., Derse, A., & Layde, P. M. (2002). Ethical issues in cost-effectiveness analysis. *Evaluation and Program Planning*, *25*, 71−83.

Rivkin, C. H. (2015, June 9). *The Global Economic Outlook—What it means for the future of legal practice*. Address to the International Bar Association 11th Annual Group Members' Leadership Summit, New York, NY. Retrieved March 2016, from http://www.state.gov/e/eb/rls/rm/2015/243318.htm

Scriven, M. (1980). *The logic of evaluation*. Inverness, FL: Edgepress.

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

Scriven, M. (1992). Evaluation and critical reasoning: Logic's last frontier? In R. Talaska (Ed.), *Critical reasoning in contemporary culture* (pp. 353−370). Albany: State University of New York.

Scriven, M. (1994). The final synthesis. *Evaluation Practice*, *15*, 367−382.

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

Scriven, M. (2008). A summative evaluation of RCT methodology: An alternative approach to causal research. *Journal of Multidisciplinary Evaluation*, *5*, 11−24.

Scriven, M. (2012). The logic of valuing. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New directions for evaluation* (Vol. 133, pp. 17−28). San Francisco, CA: Jossey-Bass.

Scriven, M. (2013, March 22). *Key evaluation checklist (KEC)*. Retrieved March 2016, from http://www.michaelscriven.info/images/KEC_3.22.2013.pdf

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

Sinden, A., Kysar, D. A., & Driesen, D. M. (2009). Cost-benefit analysis: New foundations on shifting sand. *Regulation & Governance*, *3*, 48−71.

Stake, R., Migotsky, C., Davis, R., Cisneros, E. J., Depaul, G., Dunbar, C., . . . Chaves, I. (1997). The evolving syntheses of program value. *American Journal of Evaluation*, *18*, 89–103.

Svistak, M., & Pritchard, D. (2014). *Economic evaluation: What is it good for? A guide for deciding whether to conduct an economic evaluation*. London, England: New Philanthropy Capital.

Tuan, M. T. (2008). *Measuring and/or estimating social value creation: Insights into eight integrated cost approaches*. Prepared for Bill & Melinda Gates Foundation Impact Planning and Improvement. Seattle, WA: Gates Foundation. Retrieved February 2015, from https://docs.gatesfoundation.org/Documents/wwl-report-measuring-estimating-social-value-creation.pdf

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behaviour* (2nd ed.). Princeton, NJ: Princeton University Press.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Yates, B. T. (2009). Cost-inclusive evaluation: A banquet of approaches for including costs, benefits, and cost-effectiveness and cost-benefit analyses in your next evaluation. *Evaluation and Program Planning*, *32*, 52−54.

Yates, B. T. (2012). Step arounds for common pitfalls when valuing resources used versus resources produced. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New Directions for Evaluation* (Vol. 133, pp. 43−52). San Francisco, CA: Jossey-Bass.

Discussion

The objective of this study was to address the question: what are the requirements of a model for evaluating VFM in social policies and programs? The requirements were identified through critical analysis of literature, and together represent a conceptual model for the evaluation of VFM.

The proposed conceptual model makes several novel and significant contributions to the field of evaluation. First, it proposes a definition for VFM as "an evaluative question about an economic problem". An evaluative question is a question about the merit, worth or significance of something – that is, a question about how good something is, and whether it is good enough (Davidson, 2005). An evaluative question about VFM focuses on the fundamental economic problem of resource use – that is, how well resources are used, and whether the resource use is justified. This definition positions VFM as a shared domain at the intersection between evaluation ("the process of determining the merit, worth or significance of something") (Scriven, 1991, p. 139) and economics ("the study of how people choose to use resources") (American Economic Association, 2013).

Definitions of 'merit', 'worth' and 'significance' vary in the evaluation literature (Schwandt, 2015). However, 'merit' is often used to refer to the intrinsic quality, value or virtues of a policy or program while 'worth' is often used to refer to its extrinsic value within a particular context (Coryn & Stufflebeam, 2014; Scriven, 1991; Davidson, 2005; Schwandt, 2015). For example, "the merit of researchers lies in their skill and originality, whereas their worth (to the institution that employs them) might include the income they generate through grants, fame, or bequests, attracting other good faculty and students" (Scriven, 1991, p. 227). Schwandt (2015) noted that some evaluators also equate merit with effectiveness and worth with costs, and that a further distinction is made between absolute merit and worth (the inherent quality of a policy or program) and relative merit and worth (the quality of a policy or program in comparison to alternatives). Significance relates to the concept of importance – and can variously represent a total synthesis of merit and worth (Scriven, 1991) taking into account the relative importance of different criteria of merit and worth, as well as additional considerations such as clinical, practical, statistical, historical, or cultural significance (Davidson, 2005; Scriven, 1991).

For the purposes of this research, and bearing in mind the diverse and overlapping ways in which the three terms have been used in evaluation literature, what is most important is the collective meaning of these terms rather than distinctions between them. For these purposes it is sufficient to note that VFM encompasses intrinsic (e.g., ethicality, affordability) and extrinsic (e.g., effectiveness, equity) dimensions of resource use, and that the particular dimensions at play, and their relative importance, may be pluralistic (a matter of context and perspective). The 'value' in 'value for

money' represents merit, worth and significance so defined, and 'money' represents all relevant resource use (financial and non-financial). Nevertheless, this thesis retains the terminology of 'merit, worth and significance' in order to explicitly emphasise a capacious definition of VFM, acknowledging that people value things on a variety of grounds. Evaluators should bring clarity to what is meant by merit, worth and/or significance when defining evaluation criteria for a specific context.

The second contribution of the model is to propose that evaluative questions about VFM should be addressed through evaluative reasoning. There are multiple approaches to evaluative reasoning, with a widely-used approach involving criterial inference, in which predetermined definitions of merit and worth are used to make explicitly evaluative judgements from empirical evidence. The conceptual model argues that the merit, worth and significance of resource use should be defined criterially. This can be defended on the basis that VFM is a 'cluster concept' (Scriven, 2007). That is, the VFM of a particular policy or program can be defined in terms of selected dimensions that provide necessary and sufficient definition for evaluation purposes. Criteria of VFM may vary for different programs; as Schwandt (2015) noted, "choices of criteria used in judging program value are contextually determined" (p. 49). The weighting and synthesis of evidence across multiple criteria can be supported numerically or qualitatively.

Third, the model notes that CBA is a form of evaluative reasoning, in that it exhibits the features of the general logic of evaluation. CBA involves the use of criteria (total welfare) and standards (opportunity cost, represented by the discount rate) together with evidence (monetary valuations of costs and benefits). The criteria, standards and evidence are synthesised (using a formula) to reach a judgement (net present value).

Fourth, the model argues that CBA can enhance an evaluation of VFM – for example, by promoting systematic and rational analysis of costs and consequences, by valuing both costs and consequences in the same units, and by transparently exploring the boundaries of uncertainty and risk through sensitivity analysis and scenario analysis.

Fifth, the model argues that CBA has limitations that mean it falls short of providing a full and comprehensive evaluation of VFM. Several conceptual and practical limitations have been identified in the literature which, though not exhaustive, illustrate that CBA is less than a gold standard for the evaluation of VFM. For example, although CBA is capable of capturing multiple values, it ultimately evaluates them against the primary criterion of economic efficiency. If the VFM of a policy or program is to be judged against two or more criteria (for example, efficiency and equity), where deliberation on trade-offs, power imbalances or other tensions is desirable, then CBA may not be the most appropriate evaluation method for the context.

Sixth, the model proposes that "a stronger approach [than CBA alone] would involve explicit evaluative reasoning, supported by judicious use of economic and other methods". It proposes that the minimum requirements for such an approach would be that it: pose, and answer, an evaluative question about VFM; use evaluative reasoning; match methods to context; and be conducted in keeping with program evaluation standards. These four requirements, and their rationale, are further explored in the next chapter.

The proposed conceptual model is a prototype and leaves much to be explored further. In particular, it does not fully elucidate the circumstances in which CBA meets the requirements of the conceptual model and is appropriate as a comprehensive approach to VFM assessment. Similarly, it doesn't identify the circumstances in which CBA might more appropriately be used as one method within a multiple or mixed methods evaluation. Neither does it identify whether there are circumstances in which an evaluator should consider *not* using CBA. These matters are addressed in the next study.

Chapter 5: Gap analysis

Introduction

This chapter addresses the second research question:

RQ2: To what extent and in what circumstances can cost-benefit analysis meet the requirements of the proposed model?

The purpose of this study is to better understand the strengths and limitations of CBA in the evaluation of VFM. From the research presented so far, it appears that there are reasons to argue for the inclusion of CBA in evaluation of VFM, and rationale for such evaluation not to rely on CBA alone.

A conceptual model has been developed, proposing requirements for an evaluation of VFM of social policies and programs. The proposed requirements are that an evaluation of VFM should: a) pose, and answer, an evaluative question about the merit, worth or significance of resource use; b) use explicit evaluative reasoning to reach an evaluative conclusion; c) select and tailor methods (economic and/or other) according to context; and d) be conducted in keeping with program evaluation standards. The rationale for each of these requirements, and the capacity of CBA to meet them, are systematically examined in this study.

CBA has distinctive strengths and is often assumed to be the best method for evaluating VFM – but how well can CBA meet the requirements of the conceptual model? Are there particular circumstances in which CBA doesn't meet these requirements? Are there circumstances in which CBA might make a worthwhile contribution to an evaluation of VFM in combination with other methods? How might an evaluation of VFM harness the strengths of CBA while compensating for some of its limitations? A model for evaluating VFM should take these considerations into account.

This chapter systematically assesses the potential for CBA to meet the four requirements proposed in the theoretical model that is, to: address an evaluative question about VFM; to use explicit evaluative reasoning; to select and tailor methods according to context; and to be conducted in adherence with program evaluation standards.

The model proposes that CBA can be used in two distinct ways in an evaluation: CBA can either be the overarching form of evaluative reasoning that guides the whole evaluation; or it can be used as one method (along with other methods) that contribute evidence to an evaluation guided by a different overarching approach to evaluative reasoning. This chapter will identify and propose a set of principles for determining: a) the circumstances in which CBA is suitable as the overarching form of evaluative reasoning (and

when it is not); and b) the circumstances in which CBA is suitable as a method within a wider evaluation (and when it is not).

The perspective taken in this analysis is one of *efficacy* (that is, it addresses a question of theoretical feasibility and validity – '*can* CBA meet the requirements of the model?') as distinct from real-world *effectiveness* ('*does* it meet the requirements?') – which would depend on a wider set of contextual considerations. Accordingly, the inherent capacity of CBA to fulfil the requirements is rated. Three types of rating are used: prescribed, permitted, and precluded. A *prescribed* rating means that standard model of CBA requires the method to fulfil one of the minimum requirements. *Precluded* means that adherence to the methodological standard would logically make adherence to the requirement impossible. *Permitted* means that meeting this requirement is neither prescribed nor precluded, so the requirement could be included as a customisation of the prescribed economic method.

Following the systematic assessment, a brief thematic summary of the main strengths and limitations of CBA is presented. As the feasibility and appropriateness of CBA, like any method, is context-specific (Drummond et al., 2005; Julnes, 2012c), the analysis considers the circumstances (e.g., characteristics and context of the policy or program) in which the method may be feasible and appropriate. Although economic evaluation is principally examined here from an efficacy perspective, real-world constraints such as limitations of time, resources, data and stakeholder worldviews (Bamberger, Rugh, & Mabry, 2011) are also touched upon, acknowledging that these constraints might limit the use or fitness-for-purpose of particular methods in certain contexts.

Findings

The following systematic assessment considers the four elements of the theoretical model in turn. First, it considers the extent to which CBA is able to pose, and answer, an evaluative question about the merit, worth or significance of resource use. It is argued that CBA addresses evaluative questions about efficiency, which is one criterion of VFM. The standard method of CBA cannot fully answer a broader evaluative question about the merit, worth or significance of resource use, though it can theoretically be modified to do so.

Second, the analysis considers the extent to which CBA is able to fulfil the requirement of using explicit evaluative reasoning to reach an evaluative conclusion. The analysis notes that CBA is an approach to evaluative reasoning (that is, it implements the general logic of evaluation), with strengths and limitations that make it more likely to be fit for purpose under some conditions than others. From a premise that selection of methods is

contextual and involves evaluator judgement, considerations are identified to guide judgements about when CBA may be suitable or unsuitable to use as the overarching approach to an evaluation of VFM.

Third, the analysis considers the principle of selecting methods according to context. Two scenarios are analysed. The first scenario relates to an evaluation in which CBA is selected as the approach to evaluative reasoning – that is, where the whole evaluation is a CBA. CBA involves the use of a subset of specific methods to measure and monetise costs and benefits. A decision to use CBA is also a decision to prioritise quantitative methods, and to use particular approaches to measurement (such as willingness to pay). The second scenario relates to an evaluation of VFM in which a different form of evaluative reasoning is selected, and where CBA is viewed as a method to supply some of the necessary evidence, in combination with evidence from other sources – that is, CBA is part of an evaluation. Strengths and limitations of such an approach are identified.

Fourth, the analysis assesses the extent to which CBA can be conducted in keeping with program evaluation standards. Economic evaluation standards tend to focus on technical aspects of applying the methods – for example, their validity and replicability. Program evaluation standards contain guidance on additional considerations in the design and conduct of program evaluations – for example, ethical matters such as ensuring attention to stakeholders, negotiation of evaluation purposes, meaningful processes and products, and concern for consequences and influence. Conducting CBA in adherence with program evaluation standards would include using the standards to inform decisions about when to use (or not use) CBA in an evaluation, and when to broaden an evaluation from CBA alone to CBA in combination with other methods.

VFM is an evaluative question

An evaluative question is a question about the merit, worth or significance of something – i.e., a question about how good something is, and whether it is good enough (Davidson, 2005). This thesis defines VFM as the merit, worth and significance of resource use. An evaluative question about VFM focuses on how well resources are used, and whether the resource use is justified.

The conceptual model for evaluation of VFM argues that merit, worth and significance of resource use should be defined criterially. This can be defended on the basis that VFM is a 'cluster concept' (Scriven, 2007). That is, the meaning of VFM can be disaggregated into dimensions that provide necessary and sufficient definition for evaluation purposes. Criteria are determined according to context (Schwandt, 2015). Accordingly, this thesis does not identify specific criteria of VFM but rather, argues that evaluators need to develop and define clear context-specific criteria for VFM.

Illustratively, however, it is worth emphasising that the merit, worth and significance of resource use is a capacious concept. As noted earlier, criteria of VFM may be conceptualised within the following three key dimensions: the resource use itself ('what did we put in?'); consequences of the resource use ('what did we get out?'); and whether the resource use is justified ('was it worth it?'). There may be multiple criteria within each category – for example, possible dimensions of resource use could include relevance (such as what needs are worth assigning resources to), affordability (being achievable within available resources), and frugality (minimising wastage). Similarly, consequences could include outputs and/or outcomes - and could focus on specific features of these such as intended and/or unintended consequences, their aggregate value and/or their incidence (for example, how benefits are distributed and the extent to which they reach the intended groups). This list is not exhaustive but serves to illustrate that VFM may have multiple dimensions. Further examples of criteria that may be relevant to the merit, worth or significance of resource use in certain contexts also include: ethical and lawful resource use; fidelity; efficiency; effectiveness; equity; sustainability; feasibility; scaleability; scientific value; environmental value; and cultural fit (Gargani, 2017; Goodwin, Sauni & Were, 2015; Stufflebeam, 2001b; Scriven; 2013; OECD, 2017). This view of VFM as a cluster concept has important implications for the role of CBA, as demonstrated below.

Cost-benefit analysis and VFM

While VFM is multi-criterial, CBA addresses a single criterion. Textbook CBA, as taught and used in economics, law and public policy, evaluates the efficiency of policies and programs (Drummond et al., 2005; Ellerman, 2014; Sunstein, 2018). In CBA, efficiency is equated with a program's impact on total welfare (changing the size of the pie), while distributional impacts (changing shares in the pie) are set to one side; they may be acknowledged or discussed, but are not explicitly evaluated within the core analysis (Ellerman, 2014; Sunstein, 2018). Furthermore, where the primary goal of a policy is redistribution, through transfers such as progressive income taxes or welfare benefits, it is widely understood that CBA is not a relevant method to use, precisely because it evaluates net benefits and not distributive impacts (Chapple, 2013; Sunstein, 2018).

VFM, as noted earlier, can have multiple criteria including, but not limited to efficiency. Therefore the standard formulation of CBA cannot fully address an evaluative question about VFM. Theorists have argued, however, that the standard CBA framework can be modified to incorporate valuations of wider criteria, and such adjustments have been made in CBAs (HM Treasury, 2018). Just as costs and benefits are weighted to adjust for differential timing, and sometimes for probability or risk, further weights can be applied such as distributive weights.

Two layers of distributional adjustment can be distinguished. The first layer is concerned with improving the accuracy of CBA in valuing costs and benefits, by accounting for differences in the marginal value of money (HM Treasury, 2018). At the margin, one extra dollar is more valuable to a poor person than a wealthy one, because of its relative scarcity. As an alternative to valuing all dollars equally (the default practice in CBA), valuations can be re-weighted to adjust for differences in wealth. This could materially affect results if, for example, a program creates winners and losers, and those winners and losers differ in regard to their wealth. However, "welfare economists have not proposed a practical way of determining the appropriate method of weighting" because "there does not seem to be a reliable way of determining people's marginal utility of money" (Adler & Posner, 2006, p. 24). This issue only arises because money is being used as a proxy for wellbeing (Sunstein, 2018). If wellbeing could reliably be measured directly, this problem would disappear. We return later to construct validity issues in monetary valuation.

A second layer of distributional adjustment involves weighting of valuations to incorporate distributive goals into the analysis. For example, a policy or program might be justified on distributive grounds, even if it decreases overall welfare (Sunstein, 2018). By valuing benefits according to their incidence (that is, who benefits), policies that benefit poor people could be assigned proportionately greater value than policies favouring rich people. Alternatively, the discount rate could be modified to adjust for intergenerational equity (Destremau & Wilson, 2017). Either approach would require a sound empirical basis for setting and justifying the weights used (Scriven, 1991).

Conceptually, empirical distributive weights could be derived by measuring and monetising people's altruism – a departure from the standard utilitarian construct of self-interested preferences (Adler & Posner, 2006). But what model of altruism should we measure? People's actual altruistic preferences, given their beliefs and life circumstances, or their hypothetical altruistic values under a Rawlsian 'veil of ignorance'? (Rawls, 1971). Is the aggregative 'voting' system inherent in CBA the way we should arrive at a weighting system for equity, or is a different process more appropriate, such as public dialogue and debate? (Damart & Roy, 2009; House & Howe, 1999). For now, distributive weighting in CBA should be more appropriately regarded as an intriguing possibility rather than a practical and feasible way to conduct an evaluation (Adler & Posner, 2006).

Notwithstanding the debate surrounding the possibility of distributive weighting, it is a point of fact that conventional CBA literature remains focused on a goal of maximising total welfare, irrespective of its impacts on equity (Ellerman, 2014). "Inequality features only as a peripheral concern in the world of equilibrium economics" (Raworth, 2017, p. 127). Moreover, contemporary evaluations of CBA by leading theorists have argued that this is precisely what CBA should do: play to its strength of providing a sound

estimate of total welfare while leaving wider considerations to other methods (Adler & Posner, 2006; Sunstein, 2018). These theorists acknowledged that it is important to address the distributive question of "who bears the costs and who obtains the benefits" (Sunstein, 2018, p. xiii), but argued that CBA is not the best method for doing so (Adler & Posner, 2006; Sunstein, 2018). This is not an objection to the use of CBA, though it may be a reason not to rely on CBA alone.

Summary: CBA addresses evaluative questions about efficiency

The function of CBA is to address evaluative questions about efficiency, which is one criterion of VFM. Therefore, conventional CBA offers a relevant approach to evaluative reasoning where efficiency is a relevant criterion – and may be insufficient on its own where there are additional criteria to address. For example, values such "equity, human dignity, fairness, and distributive impacts" (Executive Order No. 13563, 2011, p. 3821) need to be addressed outside a standard CBA and may conflict with efficiency goals, necessitating trade-offs.

Non-efficiency criteria may be included within a modified CBA to the extent that they can be adequately measured and monetised. The literature suggests that the inclusion of non-efficiency values in CBA is *permitted* in principle but remains *precluded* in practical terms.

In this light, CBA remains a strong candidate for evaluating efficiency, and a weak candidate for conducting a broader, multi-criterial assessment of VFM. In either case, CBA needs to pass additional tests, the next one being whether it is a legitimate and fit-for-purpose model of evaluative reasoning. This is explored in the next section.

Explicit evaluative reasoning

Evaluative reasoning is the process by which an evaluator gets "from scientifically supported premises to evaluative conclusions" as described by the general logic of evaluation (Scriven, 1995, p. 51). *Explicit evaluative reasoning* is evaluative reasoning that is clearly communicated, so that evaluative judgements and their rationale are transparent and traceable (Scriven, 1991; Yarbrough et al., 2011).

Evaluation is "a judgment-oriented practice" (Schwandt, 2015, p. 47) and the model of evaluative reasoning described by Scriven (1980; 1991; 1994; 1995; 2012) reflects the premise that judgements of the value of social programs "can and should be made on the basis of evidence and argument" and can be empirically warranted (Schwandt, 2015, p. 46).

Under this premise, evaluative reasoning is mandatory to address evaluative questions. Fournier (1995) argued that the general logic of evaluation:

> *is the basic reasoning that specifies what it means to evaluate something… it specifies the game and the rules of the game that one is playing when conducting an evaluation in any field… If someone says that he or she is doing an evaluation, then he or she must be setting criteria and standards, measuring the evaluand along these lines, and synthesizing the information into a final judgement about the merit or worth of the evaluand* (p. 17).

There are multiple approaches to evaluative reasoning (Schwandt, 2015), as detailed earlier. This thesis focuses primarily on technocratic approaches while also acknowledging the validity of tacit, all-things-considered, and deliberative approaches. Technocratic approaches are systematic and empirically-based, and span rule-governed, algorithmic and rule-based approaches, including both quantitative and qualitative approaches to weighting and synthesis (Schwandt, 2015). Although not formally aligned with the general logic of evaluation in economic texts, CBA can legitimately be viewed as one of the technocratic approaches to evaluative reasoning.

CBA as evaluative reasoning

CBA involves identifying things of value, including resource use and impacts; quantifying them; valuing them; and synthesising the evidence by aggregating the values to reach an overall determination of net value (King, 2017). By implementing these steps, CBA conforms to the general logic of evaluation. The general logic of evaluation is a logic shared by all evaluation approaches, though "what counts as criteria or evidence or how the evidence is weighed varies from one approach to another" (Fournier, 1995, p. 17). Table 3 compares and contrasts CBA with other forms of economic evaluation and with quantitative and qualitative approaches to valuation and synthesis.

Within the CBA model of evaluative reasoning, the most visible criteria (aspects of performance) that are explicitly evaluated are costs and consequences. Any cost or consequence can be identified, quantified, and valued within a CBA. There is an overarching criterion, however, which may be taken for granted and therefore not made explicit in the reasoning process that underpins a CBA. The overarching criterion is Kaldor-Hicks efficiency, in which any change that increases total welfare (that is, where the net present value is greater than zero, or benefits exceed costs) is deemed desirable (Adler & Posner, 2006).

If the overarching criterion is Kaldor-Hicks efficiency, the standard for that criterion is set by the discount rate. The discount rate represents the opportunity cost of the investment (that is, the value that could be obtained through alternative investments of comparable risk). The greater the discount rate, the higher the hurdle that must be met for the net present value to be greater than zero (Drummond et al., 2005).

Evidence of performance in CBA is measured quantitatively, through monetary valuations of gains and losses in utility, which can either be determined empirically (for example, through a survey) or by reference to proxy values (such as market prices) (Gargani, 2017). While monetary valuation provides the principal basis for weighting the relative values of costs and benefits, additional weights are also applied. For example, positive and negative signs are used to differentiate the value of benefits and costs respectively, the discounting process adjusts for differential timing of benefits and costs, and probability weights may be added to adjust for uncertainty or risk. As discussed earlier, normative weights can also be added (e.g., to rebalance the evaluation toward the achievement of additional objectives such as equity). Synthesis in CBA is achieved quantitatively, by aggregating the weighted values of costs and benefits (Adler & Posner, 2006).

Table 3: Economic methods as evaluative reasoning

Evaluation approach	Criteria	Performance Standards	Evidence of performance	Synthesis*
Cost-benefit analysis	Kaldor-Hicks efficiency. Costs and consequences, which are valued in monetary units and treated as compensatory.	Threshold for 'good enough' determined by the discount rate, reflecting the opportunity cost of capital for projects of comparable risk.	Monetary valuations of gains or losses in utility, determined empirically or by reference to proxy values. Quantitative measurement and/or modelling.	Quantitative reconciliation of discounted monetary valuations of costs and benefits, into a net present value or benefit:cost ratio.
Cost-effectiveness analysis	Efficiency of achieving a specified outcome. Resource costs, valued in monetary units, and a single quantitative outcome variable.	No inherent performance standard built into the method; requires external 'decision rules' to guide judgements of what is a 'good' level of cost-effectiveness.	Relationship of resource consumption to dependent outcome variables. Quantitative measurement and/or modelling.	Quantitative analysis, producing an incremental cost-effectiveness ratio: an indicator of efficiency.
Cost-utility analysis	Efficiency of achieving a gain in utility. Resource costs, valued in monetary units, and a quantitative measure of utility.	No inherent performance standard built into the method; requires external 'decision rules' to guide judgements of what is a 'good' level of cost-utility.	Relationship of resource consumption to utility measures. Quantitative measurement and/or modelling.	Quantitative analysis, producing an incremental cost-utility ratio: an indicator of efficiency.
Quantitative valuation and synthesis/ multi-criteria	Capable of supporting any	Cardinal importance weighting through decision	Cardinal importance scoring through	Ranking of alternatives on the basis of

Evaluation approach	Criteria	Performance Standards	Evidence of performance	Synthesis*
decision analysis	criteria of merit or worth.	makers' judgments. Can support ranking.	decision makers' judgments.	summation of weighted scores.
Qualitative valuation and synthesis	Capable of supporting any criteria of merit or worth.	Ordinal importance weighting through definitions of different levels of performance, e.g., 'poor', 'adequate', 'good' and 'excellent'. Can support grading or ranking.	Capable of accommodating qualitative, quantitative or mixed evidence	Rating or ranking of alternatives on the basis of judgement by evaluators and/or stakeholders, guided by criteria and standards.

* The principal approach to synthesis is given in the table. However, any of these approaches may also be supplemented with other forms of evaluative reasoning. For example, Schwandt (2015) described tacit, all-things-considered, or deliberative approaches, any of which apply here.

The table demonstrates that while each of the compared approaches to evaluative reasoning conform to the general logic of evaluation, they each vary in terms of "what or how criteria are identified, what or how standards are constructed, how performance is measured, and how data are synthesized" (Fournier, 1995, p. 18). In other words, each evaluation approach has its own working logic. Understanding the working logics of CBA can help to clarify the circumstances in which it may be fit for purpose, as shall be demonstrated below.

Working logic of CBA

"Working logic" is a term drawn from Toulmin's (1964, 1972) characterisation of disciplines as rational enterprises. Using Kaplan's (1964) distinction between reconstructed logic and logic-in-use, Fournier (1995) argued that "working logic is the logic-in-use found in everyday practice to establish and justify evaluative claims" and is "the variation in detail in which the general logic is followed when conducting an evaluation" (p.18).

While all evaluation follows the general logic, the working logic varies between evaluation approaches. For example, a principal source of criteria in a consumer approach to product evaluation relates to the properties inherent in the product whereas criteria for evaluating social programs can include stakeholder values or the values of an expert (Fournier, 1995).

Fournier (1995) identified two ways of conceptualising the working logic in evaluation. In the first, a set of four parameters defined the scope or boundaries in which the reasoning process occurs: *problem* (the nature of the

evaluation, for example, whether it is evaluating causality or unique features of the evaluand), *phenomenon* (the nature of the evaluand including its parts, organisation, structure, and relationship to wider context), *question* (the nature of the information need that the judgement should fulfil), and *claim* (the nature of the conclusion that will be reached). These four parameters are interwoven. In combination, they provide the foundation for building an argument that establishes and supports conclusions.

For example, in a consumer approach, the problem is "the extent of performance", the phenomenon is "a functional product", the questions are "is X a good one of its kind?" and "is X good/less good than other Xs?" and the claim is "performance/value". In contrast, for a causal approach, the problem is "intervention effectiveness", the phenomenon is "program defined as treatment-outcome relationships", questions are "what is the outcome of intervention A" and "is A more effective than B in producing X?" and the claim is "causal/value" (Fournier, 1995). The various approaches are not mutually exclusive – e.g., a product evaluation could include the employment of a causal approach (Fournier, 1995).

Applying this framework to CBA, the problem is *intervention efficiency* – that is, CBA evaluates the efficiency of an intervention by valuing and comparing its costs and consequences (Drummond et al., 2005). The phenomenon is *program defined as opportunity cost of resource use and consequences* – that is, CBA evaluates the net present value of an intervention, using the discount rate as a benchmark for the next-best alternative use of resources (Drummond et al., 2005). The questions are *what is the net value of intervention A?* and *is the net value of A greater than B?* The claim is *cost/value.* Defining the working logic of CBA in this way serves to highlight the applicability of CBA, not as a gold standard, but as a valid approach to evaluative reasoning in circumstances where the problem, phenomenon, questions, and claim are defined in this way.

The second formulation of the working logic describes the argument structure – the pattern of reasoning used to justify conclusions within the four parameters of each approach. Toulmin (1964) "examined different types of inquiry and found that inquiry is best characterised as the building of a defensible argument" (Fournier, 1995, p. 24). Fournier (1995) explained that six logical features, common to all inquiry, work together to support and justify evaluative conclusions: claims (that conclude what is to be taken as acceptable or legitimate); evidence (the facts that form the basis or foundation for the claim); warrants (that legitimate the inferences from the evidence to the claim, by appeal to some authority); backings (added authority of a more general form that support the warrant); conditions of exception (that identify circumstances when the warrant may not hold); and qualifiers (that identify the strength or forcefulness of the claim).

Taking selected features of the two concepts of working logic, Fournier (1995) compared the working logics of various approaches to evaluation. Building on that analysis, Table 4 adds the economic methods and outlines key features of their working logics.

This analysis identifies the source of values as a critical factor affecting the pattern of reasoning:

> *How the phenomenon is defined (that is, socially constructed) is important because it influences the source or locus of values from which criteria are selected (step 1 of the general logic). In turn, criteria selection affects the validity of conclusions because it influences the reasoning used in establishing them. The reasoning is affected because the source of the criteria commits us to look for certain kinds of evidence and to appeal to certain kinds of warrants in order to justify resulting claims. In other words, how evaluators reason toward evaluative judgements depends on how value (criteria) is defined* (Fournier, 1995, p. 22).

In CBA, the phenomenon of interest is a program defined in terms of the opportunity cost of resource use and its consequences (Drummond et al., 2005). The locus of values is gains or losses in individual utility (Adler & Posner, 2006). Evidence is gathered in the form of monetary valuations of those gains or losses in utility, either through empirical measurement or with reference to secondary data (Gargani, 2017). The warrant claim is that total utility (represented by the aggregation of gains and losses in utility) should be maximised (Ellerman, 2014).

The warrant is backed by the authority of the theory and practice of CBA which is, for the most part, uncontested and generally accepted as credible – in a similar manner to which the warrantability of causal claims in an experimental study is backed by its grounding in sampling theory (Fournier, 1995).

Table 4: Working logics of economic methods

Evaluation approach	Phenomenon of interest	Source of criteria (locus of values)	Evidence (foundation for a claim)	Warrant (authorises inference)
Connoisseurial/ critic approach to program evaluation	Program defined as set of qualities identifiable by an expert	Personally held values of an expert	Expert values	Expert is reliable and credible
Pluralistic approach to program evaluation	Program defined as set of values held by stakeholders	Stakeholder values	Stakeholder values and their connection to impact	Stakeholder values reflect what is desirable and important
Consumer approach to	Functional product	Properties inherent in the	Properties and their connection	Accepted meaning of the

Evaluation approach	Phenomenon of interest	Source of criteria (locus of values)	Evidence (foundation for a claim)	Warrant (authorises inference)
product evaluation		product and consumer use	to extent of performance	word (such as car or watch)
Goal-free approach to program evaluation	Program defined as a means of meeting needs	Consumer needs	Needs and their connection to program effects	Needs accepted as necessary requirements for existence
Causal approach to program evaluation	Program defined as set of treatment-outcome relationships	Dependent variables in goals or research literature	Relationships among variables	Relationships were identified under reliable methods
Cost-benefit analysis	Program defined as opportunity costs of resource use and consequences (causal impact of resource allocation)	Gains or losses in utility brought about by the program, valued in monetary terms by pecuniary (real) or non-pecuniary (hypothetical) markets	Individual values, discounted and aggregated	Total (net) utility should be maximised
Cost-effectiveness analysis	Program defined as incremental costs and effects (causal impact of resource allocation) relative to next-best alternative	Relationship of resource consumption to dependent variables	Relationship between costs and outcome variable	Programs should be ranked according to cost-effectiveness
Cost-utility analysis	Program defined as incremental costs and utility (causal impact of resource allocation) relative to next-best alternative	Relationship of resource consumption to utility-adjusted dependent variables	Relationship between costs and utility measure	Programs should be ranked according to cost-utility
Quantitative valuing and synthesis/ multi-criteria decision analysis	Program defined as set of values held by decision makers	Decision makers' values	Decision makers' values	Decision makers are reliable and credible judges of the public good; numerical weights are valid
Qualitative valuing and synthesis	Program defined by criteria of merit or worth, with ordinal standards Can accommodate any of the phenomena from	Pluralistic; drawing on multiple sources such as stakeholder values, program documentation, economic	Pluralistic; can accommodate any mix of evidence	Multiple values and perspectives need to be balanced in reaching an overall judgement of merit, worth or significance

Evaluation approach	Phenomenon of interest	Source of criteria (locus of values)	Evidence (foundation for a claim)	Warrant (authorises inference)
	other evaluation approaches, including combinations of these, and/or other phenomena	efficiency, and others Can accommodate any locus or loci of values		Ordinal valuing is more reliable than quantitative valuing, in the absence of direct measurement

Toulmin (1964) distinguished warrant-using arguments (resting on warrants that are largely uncontested) from warrant-establishing arguments (where the warrants themselves may be contested). The findings of a CBA may be considered warrant-using arguments in the sense that the findings are derived from the use of a highly developed, widely accepted method of economic evaluation. Nevertheless, warrant claims may not hold in certain circumstances and it is worthwhile scrutinising the warrant claims of CBA to determine whether, and under what conditions, they can be justified.

The overarching warrant claim in CBA is that total utility should be maximised. This claim rests on a set of premises about economic efficiency, from welfare economics (Drummond et al., 2005). This model of welfare asserts that individuals are the best judges of their own welfare, that individuals have preferences that are stable over time, that utility increases when those preferences are satisfied, and that total utility should comprise the aggregate utilities of each individual member of society (Adler & Posner, 2006).

In referencing Pareto optimality and Kaldor-Hicks efficiency, CBA rests on further warrant claims that any net gain in overall welfare is worthwhile (even if it creates winners and losers), and that competing alternatives should be compared and ranked according to their net gain in overall value (Adler & Posner, 2006).

In order to derive monetary valuations of gains and losses in utility, further warrant claims are necessary: that gains and losses in utility should be valued in commensurable units; that the actual or hypothetical behaviour of markets adequately represents the value of things. This requires an assumption that resource allocation is determined by a competitive market, which is in equilibrium (Drummond et al., 2005). Additionally, in order to adjust costs and benefits for differential timing, it needs to be accepted that social opportunity cost is adequately represented by the discount rate (Creedy & Poassi, 2017).

In order for CBA to be justified as a fit-for-purpose approach to evaluative reasoning, these warrant claims need to be justified in the evaluation

context. Where these warrant claims do not apply, CBA may not be suitable as an overarching method of evaluative reasoning.

As it turns out, the warrant claim linking CBA to welfare economics is tenuous. The Kaldor-Hicks criteria provide a moral backing for CBA on the basis that they are consistent with the utilitarian goal of maximising overall wellbeing. However, the Kaldor-Hicks test is morally inapposite because of the fact that the compensation from winners to losers is hypothetical and does not actually take place. In effect, the Kaldor-Hicks test would accept a policy that unfairly favours the wealthy over the poor (Adler & Posner, 2006).

Kaldor-Hicks with compensation is equivalent to the Pareto test. CBA cannot be justified on these grounds either, because the Pareto test requires assumptions that are inconsistent with the underlying concept of welfare that it seeks to maximise. For example, any utility-enhancing government-funded projects would be likely to violate the Pareto standard because they involve taxing individuals beyond the limits of their altruistic preferences. The Pareto test is also impractical for real-world application because it isn't feasible to measure and administer compensation for every new project or regulation (Adler & Posner, 2006).

CBA, moreover, only approximates the Kaldor-Hicks test because it relies on additional assumptions about the distribution of preferences in a population. A stronger moral defence of CBA would be to decouple it from the Kaldor-Hicks test and link it directly to wellbeing. However, attempts to ground CBA in a social welfare function have, to date, failed because it has proven complex to define and measure (Adler & Posner, 2006; Sunstein, 2018). Monetised costs and benefits have an ambiguous relationship to actual welfare (Sunstein, 2018), therefore CBA as currently constructed is not sufficiently related to wellbeing to constitute an iron-clad warrant claim.

In sum: "Although most economists continue to support CBA, it is fair to say that most economists also think that the practice does not have a firm theoretical basis" (Adler & Posner, 2006, p. 24).

These problems are not fatal to CBA as an approach to evaluative reasoning, however. The fundamental idea that costs and benefits can be quantified, valued in commensurable units, and reconciled remains sound. Adler & Posner (2006, p. 25) concluded that the best defence of the method is that CBA is an imperfect but useful "decision procedure" and "a rough-and-ready proxy for overall well-being". Sunstein (2018) took a similar view, arguing that CBA provides essential information to assist in decision-making, even if that information is incomplete.

Adler & Posner (2016) proposed "new foundations" to move CBA closer to a defensible approach by invoking a moral principle they called "weak welfarism", in which "overall welfare has moral relevance but. . . other considerations, such as distributive or rights-based considerations, may have

moral relevance as well" (p. 26). They also argued that CBA can move closer to tracking overall welfare if some of the standard assumptions underpinning utility are relaxed – for example, if narrow, self-interested preferences are "laundered" (p. 150) to address distortions such as poorly-informed and "objectively bad" preferences (p. 129).

Even under such new foundations, however, there remain conditions of exception (Toulmin, 1964) where the basic warrant of estimating and aggregating costs and benefits may not hold. For example, use of CBA rests on the validity of the premises: that values should be treated as commensurable and fungible; that aggregation of values should be prioritised over deliberation on differences; and of the consequentialist view that the means can be judged by the ends (Julnes, 2012b).

As an example of an evaluation circumstance where the warrant claims for CBA might not apply, consider a social program targeting a minority indigenous group and aimed at addressing some long-term effects of colonisation (perhaps, for instance, the indigenous group has lower socioeconomic and health status than prevailing ethnic groups in the population). The program is developed by the government of the day, in consultation with the indigenous group, but based heavily on the values of the ruling political party – emphasising personal responsibility, whereas the indigenous group places greater value on collective responsibility. In such a case, aggregating the values of a ruling majority and a disadvantaged minority might not only obscure the divergence of values at play, but serve to reinforce power imbalances and perpetuate the status quo.

Another example of a context where the warrant claims for CBA may not hold is an evaluation focused on program processes. Consequentialism may be too restrictive in some evaluations. For example, VFM in a process evaluation may include the value stakeholders place on the quality of program delivery (for example, whether program staff behave ethically and are culturally competent) independently from the value of the outcomes the program achieves (Goodwin et al., 2015). Furthermore, during program implementation, it may simply be too early to evaluate outcomes. CBA may be used in *ex-ante* scenario analysis (to appraise potential future value and inform investment decisions), and in *ex-post* summative evaluation (to measure realised value on the basis of actual costs and consequences), but may have limited applicability during an intermediary period where a programme has commenced but outcomes are yet to emerge.

Summary: CBA is evaluative reasoning but is not universally fit for purpose

On the basis of these considerations, it is concluded firstly that CBA is an approach to evaluative reasoning: it uses quantitative valuing and synthesis *(prescribed)*, with the distinct advantage that costs and consequences are

valued in commensurable units and can be combined in the synthesis step to derive a net valuation.

Secondly it is concluded that CBA, as an approach to evaluative reasoning, can only provide a comprehensive evaluation of VFM in specific circumstances: in particular: where maximising aggregate value is the sole criterion; where all values should (and can) be measured in commensurable units; where aggregation of diverse values is valid; and where only costs and consequences (and not processes) matter. Outside of these conditions of exception, the use of CBA as a method of evaluative reasoning is *precluded.* In short, one of the unique strengths of CBA – its ability to reduce a complex value construct to a single criterion – also turns out to be a limitation. "We give up too much by idealizing a thermometer that simplifies decision making. The complexity of multiple contexts, cultures, and criteria should moderate expectations that a single, pre-established, more-is-better evaluative criterion can be consistently established" (Gargani, 2017, p. 117).

These findings do not diminish the power of CBA as a method for addressing evaluative questions about efficiency, in circumstances where the warrant claims hold. Rather, they clarify the circumstances in which CBA can be regarded as the whole evaluation, and when it should more appropriately be seen as providing part of an evaluation, requiring supplementation with other methods. It is to the selection of appropriate methods that our attention must now turn.

Match methods to context

The preceding section addressed approaches to evaluative reasoning, using criteria and standards to reach evaluative conclusions from evidence. This section deals with the choice of methods to obtain the necessary evidence. As characterised by Christie and Alkin (2013), the "methods branch" of evaluation is concerned with "obtaining the most rigorous knowledge possible given the contextual constraints" (p. 12).

What constitutes credible evidence is a matter of debate (Donaldson, Christie & Mark, 2015) with key points of contention including the respective roles of qualitative and quantitative evidence, and the relative merits of different methods to support causal inference in particular. As Schwandt (2015, p. 22) noted, "method debates are generally proxies for deeper differences surrounding what evaluation should be examining" – that is, philosophical orientations toward the nature of reality (ontology) and knowledge (epistemology) (Kuhn & Hacking, 2012). It is not necessary to replicate those debates in this research. Here, it is held that no methods are perfect, that multiple methods may be used within an evaluation, and that the validity and feasibility of methods is contextually determined (Donaldson, Christie & Mark, 2015; Patton, 2011; Schwandt, 2015).

Moreover, some theorists argue that different methods and data sources can be combined in intentional and planned ways (Schwandt, 2015), enabling evaluations to reach a richer and more nuanced understanding than could be achieved through the use of a single method. For example, Bamberger (2012) described "mixed methods evaluation" as follows:

> *Mixed methods evaluations seek to integrate social science disciplines with predominantly quantitative and qualitative approaches to theory, data collection, data analysis and interpretation. The purpose is to strengthen the reliability of data, validity of the findings and recommendations, and to broaden and deepen our understanding of the processes through which program outcomes and impacts are achieved, and how these are affected by the context within which the program is implemented.* (p. 1).

Greene (2005; 2007) argued that mixed methods designs enable: triangulation of evaluation findings (enhancing their credibility and validity by comparing different types of evidence from different sources); development (using results from one method to inform the design of another); complementarity (harnessing the relative strengths of different methods to broaden and deepen understanding); initiation (generating new insights by examining points where findings from different sources diverge or conflict); and value diversity (as different methods advance different values).

"Methodological decisions", Greene argued, "are always made in service to substance" (2007, p. 114). In this thesis, the substance is the process of reasoning used to establish evaluative conclusions, and the selection of methods is influenced by the type of evaluative reasoning used (Fournier, 1995; House, 1980).

It is at this point that we encounter a fork in the road – and we must take both forks in order to properly assess the efficacy of CBA for evaluating the merit, worth or significance of resource use. CBA, as noted earlier, can either be used as a form of evaluative reasoning (to address economic criteria), or as a source of evidence within an evaluation (to address multiple and diverse criteria). The path chosen fundamentally affects the choice of evaluation methods, as we shall see.

First, the scenario is considered in which a decision has been made to use CBA as the overarching approach to evaluative reasoning. How does this influence the selection of methods? Second, the scenario is considered in which CBA is a method for obtaining part of the evidence to support a broader evaluation. When might economic evaluation be fit for this purpose? In what circumstances would it not be fit for this purpose?

Scenario 1: CBA as a whole evaluation

Suppose it has been decided that an evaluation of VFM is to be undertaken, and CBA is the chosen approach to evaluative reasoning (as noted above, this would be a situation where we are solely interested in evaluating the efficiency of a program, where we accept the legitimacy of net present value as an appropriate construct, and accept its warrant claims for the purpose at hand). This decision to use CBA will determine the methods we have to use in the evaluation.

Economic methods of evaluation are principally quantitative – they involve identifying, measuring and comparing numerically the costs and consequences of alternatives (Drummond et al., 2005). Qualitative methods may be used to support this endeavour – for example, logic modelling may be carried out in consultation with stakeholders in order to inform the specification of a quantitative model (Nicholls et al., 2012). Fundamentally, however, a decision to undertake a CBA is a decision to measure, analyse and synthesise values quantitatively.

CBA is underpinned by a specific suite of methods for gathering and analysing the evidence required. When a choice is made to use CBA as the approach to evaluative reasoning, certain methods come 'packaged' with that reasoning approach. In particular, CBA involves discounted cashflow analysis and utilises specific theories and practices of measurement. Key considerations are discussed here.

Discounted Cashflow Analysis

Discounted Cashflow analysis is a form of time series analysis, common to CBA, CEA and CUA, that adjusts costs and consequences for differential timing. Economic evaluation is comparative – it involves comparing the costs and consequences of alternatives – and that comparison must be made at one point in time (typically the present). Therefore, the timing of costs and consequences that occur in the past and future needs to be taken into account, by converting them to present values (Drummond et al., 2005).

The timing of costs and consequences also varies within a program. For example, a five-year investment aimed at changing social norms (e.g., male attitudes to female economic participation in a patriarchal society) may take considerably longer than five years to achieve its full impacts. In such a case, costs will be incurred over the near term while benefits may be achieved over a longer time horizon.

Discounting, as explained earlier, makes allowance for the variation in timing of costs and consequences, between and within programs. This allowance is based on the economic concept of *time preference:* "even in a world with zero inflation and no bank interest, it would be an advantage to receive a

benefit earlier or to incur a cost later – it gives you more options" (Drummond et al., 2005, p. 72).

Discounted cashflow analysis disaggregates costs and benefits according to their timing, and allocates them to a defined series of time intervals (e.g., monthly or annual) over a defined time horizon (e.g., a certain number of years). It then adjusts for this differential timing by applying a *discount rate* – a (usually) fixed percentage amount that progressively reduces the value of costs and consequences, the later they occur (Levy & Sarnat, 1994).

The formula for estimating the present value of a future cost outlay is given by:

$$PV = \sum_{n=1}^{n} F_n(1+r)^{-n}$$

Where:

PV = Present Value

F_n = Future cost, at year n

r = discount rate (Drummond et al., 2005).

The discount rate in a CBA should reflect the value of the alternative use of resources (Levy & Sarnat, 1994). For example, the net present value represents the incremental value of a project, taking into account its costs, its consequences, and the next-best alternative use of resources. A positive net present value indicates that a program is worth investing in because its net benefit is greater than that of the next-best alternative program. In this sense, as an aid to evaluative reasoning, the discount rate sets the threshold for determining whether a program is worth investing in. The greater the discount rate, the higher the threshold (Drummond et al., 2005).

In a financial investment, setting the discount rate is relatively straightforward and represents the financial opportunity cost of the investment. This can be benchmarked against market rates such as the cost of borrowing funds, returns from investments with similar risk characteristics, or a threshold rate of return determined by management (Levy & Sarnat, 1994). In public investments targeting social change, determining an appropriate discount rate is more complex because, firstly, it cannot be directly benchmarked and must be estimated, and secondly, because it has implications for inter-generational equity (Destremau & Wilson, 2017).

The debate on setting social discount rates has been dominated by two competing theories (Drummond et al., 2005), both of which are used in practice (Creedy & Passi, 2017): the *social opportunity cost of capital* approach, and the *social rate of time preference* (Drummond et al., 2005). In

theory (under ideal market conditions) both methods should arrive at the same number. Under real world conditions, they don't (Creedy & Poassi, 2017).

The social opportunity cost of capital defines the discount rate as "the rate of return that a decision-maker could earn on a hypothetical 'next-best alternative' to a public investment" (Creedy & Passi, 2017, p. ii). A proxy for the social opportunity cost is the rate of return that can be expected from private sector investments with similar risk characteristics. This can be estimated using a model that estimates a risk-free rate of return plus a risk-based premium, with appropriate comparator investments being determined by context and judgement (Levy & Sarnat, 1994).

The alternative approach, the social rate of time preference, defines the discount rate as "the rate of return that a decision-maker requires in order to divert resources from use in the present, to a public investment" (Creedy & Passi, 2017, p. ii). This method involves the use of a welfare function model in which a number of assumptions and judgements must be made to set values for input variables representing future economic conditions, the decision-maker's pure rate of time preference, and other factors (Creedy & Passi, 2017).

Choosing which approach to use, and using the chosen approach to estimate a discount rate, requires multiple analyst judgements and attention to the purpose of the CBA and what is being evaluated. If the costs and benefits are principally financial and relatively short-term, then the government's cost of funds may be an appropriate rate. If, however, the CBA involves weighing the incidence of costs and benefits across different groups of people at different times, then more deliberation may be required as to the appropriate method and parameters (Destremau & Wilson, 2017). Despite its positivist roots and empirical ambitions, CBA, like any other approach to evaluation, is a judgement-oriented practice.

Aside from discounted cashflow analysis, the other family of methods distinctive to CBA are the methods used to value costs and benefits monetarily. These are summarised next.

Cost analysis and monetary valuation

In CBA, a critical methodological issue is the valuation of both costs and consequences in monetary terms (Drummond et al., 2005). Crucially, though costs and consequences are represented by monetary units, the underlying value is understood to be non-financial (Nicholls et al., 2012; Svistak & Pritchard, 2014).

In economic evaluation, the term "cost" refers to the opportunity cost of resource use. In other words: "The cost of a specific intervention will be defined as the value of all the resources that it utilizes had they been

assigned to their most valuable alternative use" (Levin & McEwan, 2001, p. 44). The costs that should be included in a CBA are not limited to the financial resources contributed by a funding body, such as a government or philanthropist. For example, the opportunity cost of unpaid volunteer time should be converted to a monetary value (Levin & McEwan, 2001). Additionally, there may be costs borne by participants, which may be monetary or intangible – for example, psychological costs (Yates, 1996).

Conceptually, costs and benefits are flipsides, representing losses and gains in utility respectively. For a person who stands to gain from a project, the *compensating variation* is the maximum amount of money that must be taken from that person to maintain them at the pre-project level of utility, based on their preferences (Adler & Posner, 2006). This represents the maximum they would be willing to pay for the project to proceed (Drummond et al., 2005). Conversely, an *equivalent variation* can be measured, being the minimum amount that must be paid to a person to forego the gain and maintain them at the pre-project level of utility (Adler & Posner, 2006). This represents the minimum they would be willing to accept in compensation for the project not going ahead (Drummond et al., 2005). CBA assumes that each person's compensating variation or equivalent variation is an adequate representation of the difference in their utility between the two states (Adler & Posner, 2006). There are alternatives to this approach as a basis for valuing individual preferences (Nussbaum, 2000). However, willingness to pay is the generally accepted and used approach in CBA (Drummond et al., 2005; Levin & McEwan, 2001).

There is a considerable body of literature devoted to techniques for estimating the monetary value of changes in utility. A detailed critique of these is beyond the scope of this research. However, it is relevant to provide a general overview in order to illustrate some of the challenges and limitations involved. Conceptually, these techniques fall into three broad categories, based on the sources of information used: real money, real markets, and notional markets (Gargani, 2017).

The first and most straightforward category of monetary measurement is applied to costs and benefits whose natural unit of measurement is already monetary (Drummond et al., 2005) – for example, funds invested in a project, fiscal savings stemming from the impacts of a project (such as reduced utilisation of hospital services due to improvements in health status), or income earned by a project. These costs and benefits can either be measured directly from financial records or can be inferred from experimental, quasi-experimental or correlational studies (Levin & McEwan, 2001). Adjustments are sometimes necessary to convert financially accounted costs into estimates of opportunity cost. For example, the opportunity cost of capital equipment is the portion of the equipment's useful lifespan consumed by the investment, which differs from the accountancy treatment of depreciated value (Drummond et al., 2005).

The second family of techniques involves the observation of actual behaviours in real markets to determine the value of changes in utility, based on the prices of relevant goods or services. For example, the "human capital approach" measures the value of labour productivity in terms of market wages (Drummond et al., 2005). Using this method, volunteer labour (which incurs no financial cost) can be valued at a representative wage rate for the type of labour being provided. Similarly, other intangible values can be inferred from prices set in real markets. So-called 'revealed preference' studies use differences in market prices to determine the value people place on things (Levin & McEwan, 2001). For example, by comparing the difference between real estate values inside and outside the boundaries of a desirable school catchment, the intangible value to consumers of being inside that catchment is revealed. Similarly, the statistical value of a human life can be inferred by comparing market wages of risky jobs with those of less-risky jobs at similar skill levels (Adler & Posner, 2006; Sunstein, 2018).

The third category of valuation techniques relates to intangible costs or benefits that do not have an observable market value. The methodological solution to this problem falls under the general heading of 'contingent valuation' studies, in which a hypothetical market or set of trade-offs is constructed to elicit measurements of willingness to pay (Adler & Posner, 2006; Sunstein, 2018). For example, the monetary value of intangible benefits (like retaining access to a public library, being reunited with family members, or gaining a sense of hope) can be determined empirically by setting up a trade-off in which survey respondents reveal what they would be willing to pay to access the benefit.

Within each of these families of methods are a number of analytical choices. Each of these techniques has particular strengths and its problems, and the choice of technique can affect the estimated value (Levin & McEwan, 2001). Drummond et al. (2005) noted that there is disagreement in which technique should be used under what circumstances. For example, reviews of willingness to pay studies in health care have found wide variation in the design and methods used (Drummond et al., 2005). Some of the commonly cited challenges are outlined below.

Construct and measurement challenges

In conventional CBA, welfare (or wellbeing) is the phenomenon of interest (Sunstein, 2018). The construct used to represent wellbeing is utility, and the measure representing the construct is willingness to pay. Challenges arise because willingness to pay imperfectly measures utility, and utility imperfectly represents wellbeing.

The construct validity of real-world market valuations to represent willingness to pay is open to criticism on the basis that markets are assumed to be in equilibrium, whereas in reality they are subject to distortions such as

information asymmetries (Drummond et al., 2005). Furthermore, it is known that human behaviour is poorly represented by the assumptions embedded in *homo economicus*, the model of a "rational economic man" which has been described as the atom of economic theory (Raworth, 2017, p. 82). Homo economicus is narrowly self-interested, has fixed preferences, is isolated from others, makes decisions based on perfect knowledge by calculating the costs and benefits of everything, and has dominion over nature. Research into real humans, in contrast, has revealed that people are social and reciprocating, have fluid values, are interdependent, make estimations based on partial information and flawed heuristics, and are "embedded in the living world" (Raworth, 2017, p. 98).

Given these discrepancies between the model and reality, some revealed preferences may not adequately represent the value people truly place on something. For example, preferences may be distorted by poor information (Adler & Posner, 2000). Short-term preferences can undermine the achievement of longer-term objectives (Boston, 2017). Preferences can reflect various cognitive and motivational distortions with respect to both values and facts; perceptions of risk, for instance, may be influenced by media reporting over more objective data (Sunstein, 2000). People may prefer the status quo because they have adjusted to it, even though some other set of conditions would objectively serve them better (Adler & Posner, 2000). Preferences, as noted earlier, are also distorted by access to resources; since the marginal utility of money depends on people's wealth (and therefore ability to pay), those with more resources may systematically value things more highly than those with fewer resources (Sen, 2000).

Adler and Posner (2000) noted that the preferences that determine utility under traditional models of welfare are self-interested preferences, and that these preferences incompletely represent welfare. For example, preferences can be altruistic (wanting to benefit other people for no personal gain), morally motivated (for example, wanting to preserve an endangered species for no personal gain), or morally repugnant (such as addiction or murder). The authors subsequently advocated for "laundering" preferences to adjust for these shortcomings (Adler & Posner, 2006). The basis upon which this should be done, however, has not been elucidated (Sinden et al., 2009).

Related to the shortcomings of using individual, self-interested preferences as the basis for measuring welfare, are further objections to willingness to pay that focus on potential differences between individual and collective or social values (Destremau & Wilson, 2017). For example, people may make different choices as consumers than they do as citizens, and the sum of individual compensating variations may not represent the value a community would place on a public good in a collective sense. People's willingness to pay may, for instance, be modified by knowledge of what other people are willing to contribute (Julnes, 2012c). It was for reasons such as these that Sen

(2000) argued that CBA should look beyond mainstream approaches to willingness to pay and invoke explicit social choice judgments.

The use of hypothetical market models to price non-market goods brings additional problems. For example, people may not reveal what they are truly willing to pay for public goods unless the question is backed by an actual demand for payment (Sen, 2000). Individual choices can have perverse effects, disadvantaging the collective (as predicted by the classic problem of free-riding with pure public goods) (Carlton & Perloff, 1994). Additionally, willingness to pay and willingness to accept can yield very different answers to the same question, because of the endowment effect – the fact that people often require greater compensation to give up something they already have than they would be willing to pay for the same thing if they don't already have it. It is not always clear which valuation is the more correct one to use (Adler & Posner, 2006).

There are particular difficulties determining existence values under willingness to pay – for example, the value of some feature of the natural environment that does not have instrumental value to human beings but is prized for its own sake. In such cases, the use of contingent valuation approaches has produced results that contradict commonly accepted rational choice – for example, Sen (2000) cited the "embedding effect" where people express the same willingness to pay to save 2,000 endangered birds to saving 20,000 birds, even though which option is selected may have a material bearing on the sustainability of the species.

In practical terms, there are limits to the costs and benefits that are feasible to include in a CBA. "Many of the costs and benefits of particular policy interventions are hard to quantify, let alone monetise" (Boston & Gill, 2017, p. 26). Consequently, it has been found that "CBA analysts, in practice, ignore welfare dimensions [that] are just too hard to estimate given current techniques" (Adler & Posner, 2006, p. 78). "This is critical because the exclusion of particular benefits or costs can affect whether the net valuation obtained through CBA is positive or negative, and therefore whether the investment is considered worthwhile or not". (King, 2017, p. 105).

Summary: efficacy of CBA as a whole evaluation

The examples provided above are sufficient to illustrate that CBA imperfectly measures welfare. Nevertheless, CBA should not be abandoned. As Sunstein (2018) argued, "if we have a reliable welfare meter, we should use it. We should not use cost-benefit analysis. Unfortunately, none of us has a welfare meter" (p. xiv). Despite its construct and measurement challenges, CBA is able to provide acceptable proxies of welfare to inform decision-making and there are no options that can provide a better estimate at present (Adler & Posner, 2006; Sunstein, 2018).

Economic evaluation excels in dealing transparently and informatively with unknown variables, through modelling and forecasting, with scenario and sensitivity analysis (King, 2015). These approaches support robust evaluation of the circumstances in which a project would return a net present value, even when the precise values of input variables are not known. Breakeven analysis, for example, offers a solution to the problem of missing information (Sunstein, 2018; Svistak & Pritchard, 2014). Breakeven analysis enables assessment of the prospect of benefits equalling or exceeding costs, when the costs are measurable but the benefits are not. For example, if the value of one unit of a program outcome can be estimated (for instance, the value of one life saved), then breakeven analysis can determine the threshold number of outcomes to balance the costs of the program (Svistak & Pritchard, 2014).

In summing up, the methodological strategies and challenges in CBA were explored in order to appraise the efficacy of CBA as an overarching approach to evaluative reasoning. It has been demonstrated that a decision to use CBA is a decision to use quantitative methods and, in particular, discounted cashflow analysis together with a menu of approaches to monetary valuation. The methods used in a CBA come bundled in with the evaluation approach and, if using CBA, the evaluator accepts the whole bundle. CBA measures welfare imperfectly but is the least-bad method we have for this purpose.

If CBA were viewed only as an approach to evaluative reasoning, this would make CBA an all-or-nothing, take-it-or-leave-it proposition. One would have to conclude, on the basis of its limited scope and conditions of exception to its warrant claims as an approach to evaluative reasoning, that CBA is unsuitable for use in social policies and programs where equity and social justice matter, where processes matter, and where intangible values are hard to credibly value in monetary terms.

There is another option, however: CBA can alternatively be viewed as a method supplying part of the evidence for an evaluation. If viewed in this way, the evaluator may consider combining the findings from CBA with additional evidence – using a different form of evaluative reasoning to make overall judgements of the merit, worth and significance of resource use.

Scenario 2: CBA as part of an evaluation

Taken together, the considerations canvassed so far suggest that CBA often should not be used as the overarching framework for evaluative reasoning in an evaluation of VFM in social policies and programs. This accords with the arguments of Adler and Posner (2006) and Sunstein (2018) that CBA should stick to its core strength of providing a sound estimate of aggregate welfare.

VFM, as already noted, can be a multi-criterial construct – and the set of criteria in the value construct can have consequences for the methods used to gather and analyse evidence. Where the criteria include economic criteria

(such as efficiency or net present value) and additional criteria (such as relevance, equity or sustainability), economic methods and other methods of evaluation could be used in combination. This would require an overarching method of evaluative reasoning, such as Fournier's (1995) depiction of the general logic of evaluation, in which criteria and standards are defined prior to determining what evidence must be gathered or what methods should be employed.

The prospect of addressing a VFM question using CBA and other methods together, introduces the possibility that these methods could be used together in a coordinated way. For example, rather than attempting to re-weight a CBA to take equity into account, the CBA component of an evaluation could focus on estimating efficiency, while another set of methods could be used to analyse equity. Rather than attempting to place a monetary valuation on intangible benefits, a CBA could be tailored to focus on readily monetisable costs and benefits, with alternative methods being used to assess social and cultural values (or alternatively, the analysis of intangibles could involve multiple methods including CBA, providing richer evidence to inform deliberation). Rather than focusing exclusively on aggregate value, a range of methods could be used to elicit and understand the interests and worldviews of affected groups. This understanding could then inform the development of CBAs exploring costs and benefits from the perspectives of different subgroups. Rather than taking an exclusively consequentialist view of program value, an evaluation could balance outcome efficiency with process-related considerations that may not be measured in the valuation of the program's consequences.

If used in this way, many of the advantages of CBA could be retained, including systematic identification of costs and consequences, synthesis of costs and consequences to produce summative indicators of efficiency, sensitivity and scenario analysis to understand the boundaries of a model under uncertainty, and discounting to take care of differential timing and opportunity cost. At the same time, limitations of CBA could be compensated for, through the use of complementary methods – for example, by making explicit any trade-offs between total welfare and distributional impacts, and allowing for deliberation on conflicting perspectives or qualitative notions of value.

The evaluator must then determine when it is appropriate to use CBA as one of the methods. Of primary concern here, given the focus of this thesis, is the question of when to use, or not to use, CBA. Selecting between different economic methods (CBA, CEA, and CUA) is already well covered in the literature (Levin & McEwan, 2001; Drummond et al., 2005; Persaud, 2007; Yates, 2008). It is beyond the scope of this analysis to canvass considerations informing the selection of any or all social science methods that could be used to gather and analyse evidence of performance and VFM.

Julnes (2012b) proposed a framework for informing the selection of methods for assisted valuing in the public interest. The framework endeavoured to identify and organise key contextual factors that might influence selection of methods of valuing – informed by "a pragmatic approach that acknowledges and defends the value of multiple approaches to valuing" (p. 109). The framework reflects the view that individual evaluation contexts are too complex to warrant a formal algorithm but that a set of considerations might nevertheless be defined that would help to guide judicious selection of appropriate methods.

Table 5 summarises a series of considerations, building on those proposed by Julnes (2012b), that might influence the selection of methods. The table distinguishes three groups of methods. The first group includes CBA, with costs and consequences valued in commensurable units. The second group encompasses CEA and CUA, which quantify costs and consequences in differing units. The third group covers all other methods of valuing, examples of which include "surveys, focus groups, case studies, and evaluator judgment" (Julnes, 2012c, p. 111).

Grouping the methods in this way is, of course, a simplification. Nevertheless, it serves to draw attention to key considerations that could affect the selection of economic and/or other valuing methods. This is an area ripe for further research. What is clear from this analysis is that making an informed decision to incorporate economic analysis within an evaluation requires clarity about how economic analysis will contribute evidence toward evaluative judgements, including consideration of information needs, valuation needs, social process needs, and feasibility.

Table 5: Factors influencing selection of valuing methods

	CBA	CEA and CUA	Other models of valuing
Examples of relevant evaluation questions	Is the evaluand worth investing in, based on a net valuation of its costs and its consequences?	How does the cost per unit of outcome or utility compare to that of another option?	Is the evaluand worth investing in, based on the balance of evidence about relevant factors?
Examples of criteria	Economic efficiency (total welfare)	Comparative outcome efficiency (cost-effectiveness or cost-utility)	Equity, human dignity, fairness, distributive impacts, sustainability, cultural value.
Information needs: Evaluation purposes, and the types of practical decisions stakeholders need to make, influence complexity of valuing	Well suited where the evaluation purpose spans oversight and accountability (e.g., decisions about which programs or policies need more attention), program or organisational improvement (e.g., decisions about what incremental changes might increase VFM), and where relevant factors can adequately be reduced to a subset amenable to rational valuing. If the evaluation purpose is more complex, e.g., to assess overall merit, worth or significance, economic methods may provide supporting evidence in combination with other sources.		More complex situations, e.g., where there are diverse stakeholders, long-term outcome indicators, predominantly intangible values, and/or where the evaluation purpose is to assess value criteria beyond efficiency.
Information needs: Needed precision influences levels of measurement	Ratio measurements for single or multiple evaluands (may also contribute toward a categorical or ordinal judgment).	Interval measurements for the purpose of ranking evaluands (may also contribute toward a categorical or ordinal judgment).	Categorical or ordinal judgments about single or multiple evaluands.
Valuation needs: Balancing individual and collective values influences use of individual and/or group based methods	Individual judgements in valuation where total value is the sum of individual valuations (and where individual outcomes have little impact on communities, e.g., private consumption). Quantitative algorithm for aggregation of values.	Total value within the analysis is the sum of individual outcomes and/or valuation (however, final valuation occurs post-analysis with collective decision making). Quantitative algorithm for aggregation of values.	Can accommodate individual valuing as well as collective or group valuing (e.g., based on stakeholder interactions, as in a focus group). Aggregation of values via qualitative algorithm (e.g., rubric) and/or holistic synthesis.
Social process needs:	'Social betterment' defined for purpose at hand by Kaldor-	'Social betterment' defined for purpose at hand by a single,	'Social betterment' may include outcomes and/or processes. The

Development as a process goal, influences mechanistic and organic models	Hicks efficiency criteria; processes are only of interest to the extent that they achieve outcomes and so are fully and adequately represented by outcomes.	quantifiable outcome indicator or by a multi-attribute utility function; processes are only of interest to the extent that they achieve outcomes and so are fully and adequately represented by outcomes.	ultimate outcome may even be a process (such as ongoing development). Can accommodate multiple criteria and associated trade-offs, e.g., efficiency and equity.
Social process needs: Paradigms privileging social problems; influences strategic use of multiple paradigms	Addressing the social problem of inefficiency via Kaldor-Hicks efficiency analysis. Privileged (elite) source of values (with some stakeholder-driven valuation feasible within the confines of the method). Consensus seen as possible and desirable. Value claims are aggregated.	Addressing the social problem of inefficiency via analysis to assist decision makers. Privileged (elite) source of values. Consensus seen as possible and desirable. Value claims are aggregated.	Can accommodate a range of social problems including domination, conflict suppression and illegitimacy, as well as inefficiency. Can accommodate the possibilities of divergent worldviews and/or consensus. Balance of stakeholder-driven and elite valuing. Value claims can be aggregated, or not.
Practical prerequisites for applying the methods: Choice of method is influenced by available data, time, money, and political pressures (Bamberger et al., 2011). The cost of an evaluation should not exceed the value that can be derived from it (Levin & McEwan, 2001).	It is feasible to derive credible quantitative estimates of impacts. These impacts can be expressed or converted into monetary values in a valid, reliable, acceptable way. These valuations can be meaningfully aggregated (Svistak & Pritchard, 2014).	Two or more alternatives are being compared (and, in the case of CEA, these alternatives have equivalent outcomes which can be adequately expressed by one indicator). It is feasible to derive credible quantitative estimates of impacts (and, in the case of CUA, impacts can be converted into reliable utility weights).	Offers a more flexible suite of options in any situation where the pre-conditions for CBA, CEA or CUA are not met.

Summary: efficacy of CBA as part of an evaluation

This analysis has examined the efficacy of CBA under two scenarios: The use of CBA as a form of evaluative reasoning (that is, CBA as the whole evaluation), and the use of CBA as one of multiple methods within an evaluation, to support a different form of evaluative reasoning (CBA as part of an evaluation).

When CBA is used as a form of evaluative reasoning, the methods used to support the approach (in particular, discounted cashflow analysis and monetary valuation) are mandatory. Additional methods may be matched to context provided they support the structural requirements of a CBA. For example, the analyst may obtain valuations through the use of methods such as surveys or literature review. Ultimately, however, the forms of evidence required are quantitative estimates of modelling parameters. In practical terms this *precludes* the use of qualitative methods to serve the core of evaluative reasoning, though such methods may be *permitted* in an ancillary role. When evaluating VFM in social policies and programs, the limitations of scope, warrant claims and methods indicate that conducting a full evaluation using CBA alone should be the exception rather than the rule.

On the other hand, if CBA is used to produce evidence supporting a wider evaluation of VFM, the selection of appropriate methods according to context is *permitted.* For example, economic analysis can provide a quantitative estimate of efficiency and in so doing can enhance an evaluation. Where CBA would be insufficient on its own to evaluate VFM, evidence from CBA can be combined with evidence from other sources to provide a richer understanding of the program, and better-informed evaluative judgements, than could be achieved with either economic or non-economic methods alone.

This appraisal of CBA provides a strong argument for the use of evaluative thinking (Patton, 2018; Vo & Archibald, 2018). Understanding the strengths and limitations of the method, interpreting findings in light of those strengths and limitations, and bringing wider considerations into evaluative judgements, should always feature in safe and effective use of the method. Such deliberations should be guided not just by technical considerations, but also by wider features of good evaluation practice such as ethical considerations. For CBA to be accepted into the field of practice of program evaluation, it should be subjected to the same norms of practice as any other evaluation method. This has implications for the way CBA is used including the nature and extent of stakeholder involvement, the use of CBA in conjunction with other methods, and decisions about when not to use CBA. These implications are explored next.

Program evaluation standards

Meta-evaluation, according to Scriven (1991, p. 228) is "the evaluation of evaluations. . . and represents an ethical as well as a scientific obligation where the welfare of others is involved". An evaluation of VFM, like any other evaluation, should itself be open to scrutiny to determine whether it exhibits expected features of a 'good' evaluation. There is no definitive, universally agreed checklist for this purpose. Nonetheless, a number of existing and widely used standards for program evaluation offer a basis for judging the quality of an evaluation. Many countries and organisations have developed evaluation standards or principles that identify the features of good evaluation and/or good evaluation practice (AEA, 2004; AES, 2013; OECD, 2012; Patel, 2013; Scriven, 2013; ANZEA & Superu, 2015; Yarbrough et al., 2011; UNEG, 2016). Such standards are the culmination of debate and formalise some degree of consensus about evaluation as a field of practice. Commonly agreed features of high quality evaluations are that they should be useful, practical, ethical, accurate, and accountable (Patton, 2017a).

New Zealand's evaluation standards, for example, are conceptualised within an overarching principle of "evaluation with integrity" (ANZEA & Superu, 2015). The standards argue that evaluators have an ethical commitment to contribute to the wellbeing of society and, accordingly, evaluation practices, processes and products should "assure trust and confidence in the information, findings, judgements and conclusions" (ANZEA & Superu, 2015, p. 9).

Similarly, the *Program Evaluation Standards* of the Joint Committee on Standards for Educational Evaluation (Yarbrough et al., 2011) argue that evaluation should take an explicit interest in its effects on people's lives, which in turn demands not only valid evaluative reasoning and careful selection of methods, but also attention to stakeholders, concern for consequences and influence, responsive and inclusive orientation, protecting human rights and dignity, and a range of related considerations.

The existence of a multitude of standards and ethical guidelines for evaluations reflects the fact that there is no universal position on the matter of ethics in evaluation. Values are time, place and culture-bound; any set of evaluation standards necessarily reflects particular axiological and epistemological orientations, and evaluation theories that stem from them (Mabry, 2010). Nevertheless, such standards provide a point of reference for canvassing relevant ethical issues.

Economic evaluation, like any other method, should be used in ways that uphold evaluation standards. Accordingly, deliberation about whether, when, and how economic methods are used in an evaluation is necessary. For example, careful selection of methods would include checking whether the assumptions, criteria, metrics, and processes of reaching conclusions in a CBA are explicitly justified in the cultures and contexts where the

evaluation has consequences (Yarbrough et al., 2011) – and if they are not, then alternative methods should be considered.

There are also important ethical implications associated with some seemingly technical decisions within the design of an economic analysis, such as the perspective taken and the choice of discount rate. For example, taking a funder perspective in the identification of costs and benefits excludes the values of affected communities (Pinkerton et al., 2002). The choice of discount rate affects the valuation of long-term costs and benefits affecting future generations who are not able to have a say at the time of the decision (Destremau & Wilson, 2017). The conduct of evaluation with integrity demands attention to these sorts of decisions.

Systematic analysis of CBA against program evaluation standards finds that some ethical principles espoused in program evaluation standards are not explicitly prescribed in economic evaluation. Nevertheless, CBA is generally capable of being conducted in accordance with the standards, provided the options of not using CBA, or combining CBA with other methods, are available. The analysis is presented as follows.

Assessment of CBA against program evaluation standards

The capacity of CBA to be conducted in adherence with program evaluation standards is assessed in Table 6. Notwithstanding the multitude of program evaluation standards in use globally, the *Program Evaluation Standards* of the Joint Committee on Standards for Educational Evaluation (Yarbrough et al., 2011) were used as a checklist to systematically assess CBA, reflecting their longstanding and wide usage in program evaluation, their influence on other evaluation standards internationally (Schwandt, 2015), and the degree of commonality they share with other evaluation standards (Coryn & Stufflebeam, 2014; Deane & Harré, 2016).

The methodological prescription for CBA has been widely published in numerous texts, three of which were used in this analysis. Firstly, a textbook commonly used in teaching economic evaluation to program evaluators (Levin & McEwan, 2001) was consulted. Secondly, additional detail was sourced from a seminal text for health economists (Drummond et al., 2005). Thirdly, the *Green Book* (HM Treasury, 2018) was reviewed. The *Green Book* is the principal document setting out government guidance on the appraisal of public investments in the United Kingdom. It is cited widely in international guidance including Australian and New Zealand governmental guidance on CBA (Argyrous, 2013).

A literature search was unable to identify a directly comparable set of standards in economics that would correspond with program evaluation standards. The American Economic Association first adopted a code of professional conduct in April, 2018, and this code was reviewed for the purposes of this appraisal (American Economic Association, 2018).

For each program evaluation standard, CBA was rated by applying the efficacy rating system that is used throughout this chapter. Ratings address the question of whether program evaluation standards *can* be followed in CBA, not the extent to which economists do so in practice. *Prescribed* means there is an explicit expectation that CBA should fulfil the relevant program evaluation standard – it is explicitly stated in any one of the texts consulted (American Economic Association, 2018; Drummond et al., 2005; Levin & McEwan, 2001; HM Treasury, 2018). *Precluded* means that adherence to CBA would logically or practically make adherence to the program evaluation standard impossible. *Permitted* means that meeting the program evaluation standard is neither prescribed nor precluded.

The methodological prescription for CBA focuses primarily on technical aspects of economic methods to ensure precision, accuracy and reliability. In particular, textbooks and guidelines focus on fidelity to methods for the valuation and discounting of costs and benefits (HM Treasury, 2018; Drummond et al., 2005; Levin & McEwan, 2001). Standards for reporting findings of economic studies emphasise transparency and replicability. For example, reports from health economic evaluations should include a description of relevant context such as setting and location, the study scope and perspective, justification for methodological decisions such as choice of outcome measures, time horizon and discount rate, and discussion of study findings, limitations, generalisability, and how the findings fit with current knowledge (Husereau et al., 2013). Absent from economic texts is guidance on ethical aspects of evaluation such as concern for consequences and influence, contextual viability, human rights and respect (Yarbrough, et al., 2011).

Generally there is nothing in the methodological prescriptions for CBA that would preclude CBA from being conducted in adherence to program evaluation standards – provided there is freedom to choose not to conduct a CBA, or to conduct a CBA in combination with other methods. If, however, a decision has already been made to conduct a CBA as the whole evaluation, prior to the evaluation commencing, then there are significant risks that the evaluation will fall short of standards concerned with negotiated purposes, explicit values, meaningful purposes and products, concern for consequences and influence, contextual viability, and responsive and inclusive orientation. These risks are summarised below.

The **negotiated purposes** standard states: "Evaluation purposes should be identified and continually negotiated based on the needs of stakeholders" (Yarbrough et al., 2011, p. 29). Evaluation purposes guide decisions about the design and implementation of an evaluation. Program evaluation stakeholders will have different needs which may not always be aligned with the needs of those responsible for decision-making or resourcing the program and its evaluation. Agreeing to implement any particular method prior to clarifying the evaluation purpose or purposes is a hazard to meeting this standard. This does not point to any inherent

shortcoming in CBA but indicates that, as with all methods, its use should be negotiated and not preordained.

The **explicit values** standard is that "Evaluations should clarify and specify the individual and cultural values underpinning purposes, processes, and judgments" (Yarbrough et al., 2011, p. 37). This standard recognises that valuing is central to the decisions and judgements made in evaluation, and stakeholders are more likely to find an evaluation credible and useful if they see their own values reflected in it, and if they understand the value perspectives of others. This standard encourages evaluators to work in open and inclusive ways, ensuring the values guiding the evaluation are not only the values of those in power. Evaluators should avoid imposing their own values, including the values they might place on different methods and evidence, and should not purport to be objective or values-free. CBA alone would be too restrictive a framework in which to meet this standard, unless this evaluation design was negotiated and agreed in advance with all stakeholders.

The **meaningful processes and products** standard states: "Evaluations should construct activities, descriptions, and judgments in ways that encourage participants to rediscover, reinterpret, or revise their understandings and behaviors" (Yarbrough et al., 2011, p. 51). It is argued that this standard is a necessary part of ensuring stakeholders' needs are met in an evaluation. Hazards to meeting the standard include using the initial contract to stipulate and impose upon stakeholders the approach that will be taken, and proceeding without regard to stakeholders' reactions to the evaluation. This again does not indicate any inherent shortcoming in CBA but requires the flexibility to engage stakeholders in determining whether or not to use the method, or to use it in combination with other methods.

Concern for consequences and influence encourages those conducting evaluations to "promote responsible and adaptive use while guarding against unintended negative consequences and misuse" (Yarbrough et al., 2011, p. 65). This standard seeks to ensure evaluation contributes to social betterment by catalysing improvements in policies, programs and contexts, and recognises that evaluations can have potential to do harm – for example, by jeopardising democratic participation, equity, social justice or truth. The standard states that it is important not to assume "that a technically excellent evaluation is sufficient for positive use and effective influence" (p. 67). This again underscores the risk of affording CBA 'gold standard' status and argues for a more flexible and responsive approach to evaluation design and methods.

The **contextual viability** standard is that "Evaluations should recognize, monitor, and balance the cultural and political interests and needs of individuals and groups" (Yarbrough et al., 2011, p. 93). This standard recognises the potential power imbalances between stakeholders and the need to understand the different cultural, political and economic interests held by different stakeholder groups. It is important to take care to

respond to all stakeholder needs in a balanced way and not, for example, to be perceived as placing the needs of one group (such as decision-makers) ahead of others. Adherence to this standard requires appropriate mechanisms for stakeholders to have input, which again reinforces the need for flexibility to determine an appropriate method or mix of methods.

Responsive and inclusive orientation demands that "Evaluations should be responsive to stakeholders and their communities" (Yarbrough et al., 2011, p. 113). Evaluators have a "moral professional duty" to ensure stakeholders are included and attended to in a proportionate, systematic and transparent way. Meeting this standard requires evaluators to build meaningful relationships and seek stakeholder contributions to the evaluation. It requires an openness to contradictory views and interests. It may, on occasion, involve deliberative and democratic processes. Hazards to this standard include "always favouring a specific evaluation method or approach without proper regard for the needs of the actual stakeholders in the current setting and the purposes of the evaluation", "not attending adequately to context or culture in evaluation designs and practices" and "ignoring the political vibrancy and inherent value of stakeholder positions and value judgments" (p. 116). This standard would be difficult to meet using CBA alone.

Collectively, the analysis summarised in Table 6 lends further support to the proposition that CBA should be regarded as one tool in an evaluator's toolbox, to be used in contextually responsive ways, and in combination with other methods.

Table 6: Assessment of CBA against program evaluation standards

Program Evaluation Standards (Yarbrough et al., 2011)	**Efficacy of CBA as whole evaluation method**	**Efficacy of CBA as part of a mixed methods evaluation**
Utility standards		
U1: Evaluator credibility Evaluations should be conducted by qualified people who establish and maintain credibility in the evaluation context	Prescribed	Prescribed
U2: Attention to stakeholders Evaluations should devote attention to the full range of individuals and groups invested in the program and affected by its evaluation	Prescribed	Prescribed
U3: Negotiated purposes Evaluation purposes should be identified and continually negotiated based on the needs of stakeholders	Precluded	Permitted
U4: Explicit values Evaluations should clarify and specify the individual and cultural values underpinning purposes, processes, and judgments	Precluded	Permitted
U5: Relevant information Evaluation information should serve the identified and emergent needs of stakeholders	Permitted	Permitted
U6: Meaningful processes and products Evaluations should construct activities, descriptions, and judgments in ways that encourage participants to rediscover, reinterpret, or revise their understandings and behaviors	Precluded	Permitted
U7: Timely and appropriate communicating and reporting Evaluations should attend to the continuing information needs of their multiple audiences	Prescribed	Prescribed
U8: Concern for consequences and influence Evaluations should promote responsible and adaptive use while guarding against unintended negative consequences and misuse	Precluded	Permitted

Feasibility standards		
F1: Project management Evaluations should use effective project management strategies	Permitted	Permitted
F2: Practical procedures Evaluation procedures should be practical and responsive to the way the program operates	Permitted	Permitted
F3: Contextual viability Evaluations should recognise, monitor, and balance the cultural and political interests and needs of individuals and groups	Precluded	Permitted
F4: Resource use Evaluations should use resources effectively and efficiently	Prescribed	Prescribed
Propriety standards		
P1: Responsive and inclusive orientation Evaluations should be responsive to stakeholders and their communities	Precluded	Permitted
P2: Formal agreements Evaluation agreements should be negotiated to make obligations explicit and take into account the needs, expectations, and cultural contexts of clients and other stakeholders	Permitted	Permitted
P3: Human rights and respect Evaluations should be designed and conducted to protect human and legal rights and maintain the dignity of participants and other stakeholders	Permitted	Permitted
P4: Clarity and fairness Evaluations should be understandable and fair in addressing stakeholder needs and purposes	Permitted	Permitted
P5: Transparency and disclosure Evaluations should provide complete descriptions of findings, limitations, and conclusions to all stakeholders, unless doing so would violate legal and propriety obligations	Prescribed	Prescribed
P6: Conflicts of interest	Prescribed	Prescribed

Evaluations should openly and honestly identify and address real or perceived conflicts of interests that may compromise the evaluation.		
P7: Fiscal responsibility Evaluations should account for all expended resources and comply with sound fiscal procedures and processes	Permitted	Permitted
Accuracy standards		
A1: Justified conclusions and decisions Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences	Permitted	Permitted
A2: Valid information Evaluation information should serve the intended purposes and support valid interpretations	Prescribed	Prescribed
A3: Reliable information Evaluation procedures should yield sufficiently dependable and consistent information for the intended uses	Prescribed	Prescribed
A4: Explicit program and context descriptions Evaluations should document programs and their contexts with appropriate detail and scope for the evaluation purposes	Prescribed	Prescribed
A5: Information management Evaluations should employ systematic information collection, review, verification, and storage methods	Permitted	Permitted
A6: Sound designs and analyses Evaluations should employ technically adequate designs and analyses that are appropriate for the evaluation purposes	Prescribed	Prescribed
A7: Explicit evaluation reasoning Evaluation reasoning leading from information and analyses to findings, interpretations, conclusions, and judgments should be clearly and completely documented	Permitted	Permitted
A8: Communication and reporting	Prescribed	Prescribed

Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors		
Evaluation accountability standards		
E1: Evaluation documentation Evaluations should fully document their negotiated purposes and implemented designs, procedures, data, and outcomes	Prescribed	Prescribed
E2: Internal metaevaluation Internal: Evaluators should use these and other applicable standards to examine the accountability of the evaluation design, procedures employed, information collected, and outcomes.	Permitted	Permitted
E3: External metaevaluation Program evaluation sponsors, clients, evaluators, and other stakeholders should encourage the conduct of external metaevaluations using these and other applicable standards	Permitted	Permitted

Summary: Program evaluation standards should guide the use of CBA

For CBA to be accepted into the field of practice of program evaluation, it should adhere to program evaluation standards and be subjected to meta-evaluation against such standards. Although such standards are open to ongoing debate and interpretation, they nevertheless provide a widely accepted point of reference for assessing the legitimacy and quality of evaluation designs, practices and products.

From this analysis it is concluded that CBA can be conducted in adherence to program evaluation standards, provided its use is negotiated with stakeholders. If CBA is used in combination with other methods, it is possible to work in an inclusive and responsive way, with the full range of stakeholder values, and conduct to conduct evaluations that are contextually viable and meaningful for stakeholders.

As demonstrated by the majority of standards that are 'permitted' in economic evaluation (that is, CBA is capable of adhering to the standards though there is currently no recommendation to do so), the use of program evaluation standards may enhance the use of economic methods of evaluation. While many of these principles may already be well accepted and followed by people conducting economic evaluations, there is value in making them explicit (Yarbrough et al., 2011).

There are significant risks that the standards could not be adhered to, however, in situations where CBA is chosen in advance as the sole evaluation method. To do so runs the risk that the evaluation will fall short of standards for negotiated purposes, explicit values, meaningful purposes and products, concern for consequences and influence, contextual viability, and responsive and inclusive orientation. Use of program evaluation standards should help to clarify when and how to use economic methods in evaluation.

Summary

In this study, CBA has been systematically assessed against the four requirements proposed in the conceptual model for evaluating VFM in social policies and programs: a) pose, and answer, an evaluative question about the merit, worth or significance of resource use; b) use explicit evaluative reasoning to reach an evaluative conclusion; c) select and tailor methods (economic and/or other) according to context; and d) be conducted in keeping with program evaluation standards. A perspective of efficacy (*can* CBA meet the requirements?) has been applied in this analysis, with a rating system that identifies whether each requirement is *prescribed, permitted,* or *precluded* by the methodological prescription for CBA.

The analysis reveals the extent to which, and the circumstances in which, CBA can meet the requirements of the proposed model. In doing so, the

analysis reveals that CBA should not be regarded as the gold standard for evaluating VFM – rather, it is a valid option for evaluating VFM, with strengths and limitations that make it fit for purpose in some contexts and not in others.

The first requirement of the model is to **pose and answer an evaluative question about VFM**. CBA addresses evaluative questions about efficiency – but efficiency is only part of the broader construct of the merit, worth or significance of resource use. Where efficiency is a relevant criterion in an evaluation of VFM, CBA is very likely to be a valid and useful method.

Non-efficiency values are peripheral to a standard CBA. Distributional values can be assigned monetary weights and included within a CBA – but there are technical challenges and conceptual objections to doing so (Adler & Posner, 2006). The literature suggests that the inclusion of non-efficiency values in CBA may be *permitted* in principle but in practical terms, is often *precluded*. For most social programs and policies, CBA is too narrow in scope to answer an evaluative question about the merit, worth and significance of resource use.

The second requirement is to **use explicit evaluative reasoning.** CBA is a form of evaluative reasoning. When CBA is used, a specific form of evaluative reasoning is *prescribed.* This form, however, is not universally fit for purpose. CBA can only provide a comprehensive evaluation under specific circumstances: in particular where maximising aggregate value is the primary objective; where all values should be – and can accurately be – measured in commensurable units; where aggregation of diverse values is valid (and qualitative differences between values are unimportant); and where only costs and consequences (and not processes) matter. For most social programs and policies, the warrant claims for CBA are too limiting for a complete evaluation of VFM.

The third requirement is **to match methods to context.** When CBA is selected as the form of evaluative reasoning – that is, when CBA is the whole evaluation – core methods of discounted cashflow analysis and monetary valuation come bundled with the approach. The use of additional methods such as a survey or literature review are *permitted* in an ancillary role. However, contextually-responsive selection of methods is, to a significant extent, *precluded* by the structural requirements of a CBA.

When CBA is instead used in combination with other methods, providing part of the evidence to support an overarching process of evaluative reasoning, the selection of appropriate methods according to context is *permitted*. CBA can, for instance, estimate efficiency while another set of methods could be used to gather evidence of a program's effects on equity. This brings a possibility that evidence from economic evaluation, combined with evidence from other sources, might provide a deeper understanding of the program or policy, and therefore better-informed

evaluative judgements, than could be achieved with either economic or non-economic methods alone.

Despite construct and measurement challenges, CBA is the preferred method to provide an estimate of total welfare or net benefit (Adler & Posner, 2006; Sunstein, 2018). The centrality of these criteria to VFM, and the strengths of CBA as a method, provide a forceful argument for the use of CBA in evaluations of VFM.

The fourth requirement is to **adhere to program evaluation standards.** Program evaluation standards can guide the judicious use of CBA as a potential method within an evaluation. If the use of CBA is preordained or imposed, however, there are significant risks that the evaluation will not be able to adhere to standards concerned with the evaluators' ethical duties to meaningfully involve stakeholders, engage with their values and meet their evaluation needs.

Together, these considerations suggest that in certain contexts, CBA may be neither necessary nor appropriate to include in an evaluation of VFM. For example, CBA would have limited applicability where non-efficiency values, intangible values, qualitative evidence, program processes, comparison or deliberation on diverging interests have greater importance than efficiency values, tangible values, quantitative evidence, program outcomes, aggregation and consensus.

The domain for this research is the evaluation of VFM in social policies and programs. Given the centrality of issues of equity, intangible (e.g., social, cultural) values, the common usage of multiple forms of evidence (qualitative and quantitative) to understand complex social issues, and the context of diverse values and differences in power and privilege that are often relevant in social programs, these findings support the proposal that, though CBA can enhance an evaluation of VFM, it will usually be insufficient on its own and a stronger approach would involve a different form of explicit evaluative reasoning, supported by judicious use of economic and other methods.

In the next section, the cumulative findings from the research presented thus far are parsed into a series of formal theoretical propositions which will guide the remaining steps in the research.

Theoretical propositions for a 'Value for Investment' model

Critical analysis of literature, development of a conceptual model, and gap analysis of CBA against the requirements of the model, have found that CBA should be regarded not as a gold standard for evaluating VFM, but as a potential method that can enhance evaluation in combination with other methods, at the service of explicit evaluative reasoning. The core requirements for the model are as follows (Box 1). It is convenient at this point to label the model so that we have an abbreviated name for it in the remaining analysis. For the remainder of this research, the term "Value

for Investment" (VFI) will refer to the conceptual model and its application in practice, for the purpose of evaluating a construct called VFM.

Box 1: Core requirements of a Value for Investment evaluation

Pose, and answer, an evaluative question about the merit, worth and/or significance of resource use.
Use explicit evaluative reasoning to provide a logical and transparent basis for making sound judgements. Usually criteria should be weighted using an ordinal scale, supported by qualitative weighting and synthesis. Exceptions are possible where the conditions for CBA or numerical weighting and synthesis are met. Numerical weighting and synthesis is a viable option where there is an empirical basis to justify weights *and* sound mathematical logic is possible (for example, sufficiently few criteria to avoid swamping and mutually exclusive criteria to avoid interactions between criteria). Use CBA as the whole evaluation only if: maximising aggregate welfare is the sole objective; *and* all relevant and material values should be and can accurately be represented by monetary valuation; *and* qualitative distinctions not reflected in monetary valuations are unimportant; *and* aggregation of values is appropriate; *and* costs and consequences (and not processes) matter.
Match methods to context, including consideration of the validity of methods to address the construct of VFM as defined by the criteria determined above, and feasibility of methods within the social and practical context of the evaluation.
Incorporate economic evaluation where feasible and appropriate, to contribute evidence toward judgements about VFM. Economic methods of evaluation are candidates for inclusion where: efficiency is a relevant criterion; *and* the requisite data, resources, time and skills are available. CBA is a candidate for inclusion where, in addition to the general requirements above: total welfare or net benefit is a relevant subcriterion; *and* a summation of monetisable costs and consequences is likely to give a sufficiently robust estimate of total welfare; *and* impacts can meaningfully be monetised and aggregated. CEA or CUA are candidates for inclusion where, in addition to the general requirements above: two or more alternatives are being compared; *and* comparative outcome efficiency (cost-effectiveness or cost-utility) is a relevant subcriterion; *and* it is feasible to derive sufficiently robust estimates of impacts (and, in the case of CUA, impacts can be converted into reliable utility weights). Disaggregated economic analysis should be considered to investigate differences in the incidence of costs and consequences between subgroups.
Application of this model should be guided by program evaluation standards, ensuring that the evaluation not only meets technical standards but is conducted with due consideration of relevant norms of practice such as utility, feasibility, propriety, accuracy, and accountability to evaluation stakeholders.

The VFI model is backed by six core propositions. The first two propositions provide underpinning rationale for the model. The third, fourth and fifth propositions concern the validity, strengths and limitations of CBA, as the rival approach to VFM evaluation. The sixth proposition sets out the potential strengths of the proposed VFI model.

First, it is proposed that **VFM is an evaluative question (concerned with merit, worth or significance) about an economic problem (resource allocation)**. Because resource use has an opportunity cost (of foregone alternatives), choices need to be made in resource allocation – with a 'good' allocation being one that compares favourably to its next-best alternative. The centrality of opportunity cost suggests economic thinking should feature in VFM evaluation. A 'good' resource allocation may also, however, be subject to additional considerations. For example, ethical concerns (such as equity and fairness) may also apply. Balancing multiple criteria can involve trade-offs, suggesting that good resource allocation is a matter of context and perspective.

Second, it is proposed that **evaluative questions about VFM should be addressed through explicit evaluative reasoning** (Fournier, 1995; Schwandt, 2015; Scriven, 1991; Yarbrough et al., 2011). As a corollary to the first proposition, this may seem tautological. However, it is an important precursor to the third proposition.

The third proposition is that **CBA is an approach to evaluative reasoning**. Cost-benefit analysis can be conceptualised as a way of implementing the general logic of evaluation through quantitative valuing and synthesis. CBA, therefore, is potentially a valid approach to the evaluation of VFM.

The fourth proposition is that **CBA can enhance an evaluation of the merit, worth and significance of resource use by yielding insights that would otherwise be difficult to gain**. The following five sub-propositions, based on critical analysis of literature, support the core proposition:

a) **CBA promotes systematic and rational analysis of costs and consequences:** CBA provides a structure and set of analytical rules that promote systematic and rational analysis – identifying, measuring, valuing, and comparing both the costs and the consequences of alternatives.

b) **Valuing and reconciling costs and consequences in commensurable units can provide unique insights**: CBA values costs and consequences in the same units, and therefore costs and consequences can be reconciled in the final synthesis, as a single indicator of net present value, representing the impact of the program on total welfare. This feature of CBA can provide insights that would be difficult to intuit by looking at either costs or consequences in isolation from the other.

b) **Discounting is a strength of CBA**: Discounting takes the time value of money into account, with the discount rate reflecting the opportunity cost of the investment.

c) **Sensitivity and scenario analysis are strengths of CBA**: Sensitivity and scenario analysis facilitate transparency and robust thinking about relationships between benefits and costs, taking uncertainty and risk into account. Breakeven analysis, for example, can inform judgements about the prospect of benefits exceeding costs in the face of missing data.

d) **CBA can, in principle, accurately measure values**: CBA is capable of meeting a key condition for numerical weighting and synthesis to be valid – namely that weights can be determined empirically.

There may be additional advantages of CBA. The five features listed here, however, are sufficient to demonstrate that CBA can add knowledge in various ways that could not be reliably gained without using this method. Other methods may have some of these features, but CBA brings all of these features together.

The fifth proposition is that, when it comes to evaluating VFM in social policies and programs, **CBA is usually insufficient to fully answer an evaluative question about the merit, worth and significance of resource use**. The following sub-propositions provided the rationale for this proposition. Although not exhaustive, these sub-propositions are sufficient to demonstrate that textbook CBA is not a gold standard for evaluating VFM:

e) **CBA may not capture all criteria**: No single method can address all possible VFM criteria. CBA provides an estimate of efficiency. Efficiency is an important criterion of VFM. Some VFM questions may be solely concerned with efficiency. VFM, however, can involve multiple criteria such as equity, relevance, sustainability, and ethical considerations.

f) **CBA (in its standard form) does not consider equity separately from efficiency**: The Kaldor-Hicks model of efficiency, implicit in CBA, incorporates a normative position on equity that any net gain in overall value is worthwhile regardless of distributive impacts. Assessment of VFM may involve alternative positions on equity which are not reflected in CBA.

g) **CBA treats all values as fungible**: CBA reflects a normative position that all values should be valued in commensurable units and can be traded against each other; this is valid but is too restrictive to apply as a gold standard for VFM evaluation – for example; some criteria might represent ethical bottom-lines that should not be traded off; some trade-offs may be non-linear or too complex to validly be represented in a model of dollar-for-dollar commensuration.

h) **Commensuration may obscure diversity in people's values**: Aggregation of values using a common metric may reduce the visibility of qualitative differences in the perspectives of different groups (e.g., power

imbalances, diverging worldviews, values, or interests). Such differences may be better accommodated by other approaches to evaluative reasoning, as alternatives to CBA.

i) **Commensuration may obscure diversity in things of value**: Aggregation of values using a common metric may reduce the visibility of qualitative differences between things of value (e.g., economic, social, cultural, and environmental value). Such differences may be better accommodated by other approaches to evaluative reasoning, as alternatives to CBA.

j) **CBA is concerned with costs and consequences, not processes**: CBA reflects a consequentialist perspective; a practical upshot of this is that it has limited utility to evaluate the value of program processes, independently of their consequences. CBA has application for *ex-ante* scenario analysis (to understand potential future value), and for *ex-post* summative evaluation (to understand realised value). There may, however, be a period of program implementation in between these periods where VFM evaluation of 'process value' is important and where CBA is a poor fit for purpose.

k) **The scope of a CBA may be constrained by what is measurable**: In practice, important values may be excluded from a CBA because they are too hard to estimate with available techniques. This criticism could be levelled at any method, which is why we should not rely on any one method as gold standard.

l) **CBA is not explicitly required to adhere to program evaluation standards:** The VFI conceptual model proposes that evaluation of VFM should be guided by program evaluation standards. If CBA is imposed as a gold standard, there are risks that the evaluation could fall short of standards concerned with negotiated purposes, explicit values, meaningful purposes and products, concern for consequences and influence, contextual viability, and responsive and inclusive orientation. Using program evaluation standards to guide economic evaluation would require that the evaluator remain open to the possibility of *not* conducting an economic evaluation.

The sixth proposition is that, bearing in mind the preceding propositions, **"a stronger approach would involve explicit evaluative reasoning, supported by judicious use of economic and other methods"** (that is, a Value for Investment approach). It was suggested that the minimum requirements for such an approach would be that it: pose, and answer, an evaluative question about the merit, worth and significance of resource use; use explicit evaluative reasoning; match methods to context; and be conducted in keeping with program evaluation standards.

Explicit evaluative reasoning follows the general logic of evaluation and the four steps of: defining criteria of merit, worth or significance; setting performance standards; gathering and analysing evidence of performance; and synthesising the evidence to make judgements. This

can be done through numerical or qualitative approaches to weighting and synthesis of criteria and performance. Numerical approaches are more particular in regard to the conditions that must be met (such as empirical weight determination) in order to support valid evaluative reasoning. Where these conditions are not met, qualitative weighting and synthesis enhances validity.

While the fourth and fifth propositions canvass strengths and limitations of CBA that can be found in the literature, the sixth proposition makes the following novel sub-propositions that warrant empirical investigation. These sub-propositions will be the primary focus of the case studies:

m) **VFI allows multiple and diverse criteria.** Criteria of VFM should be contextually determined. Whatever the notion of VFM is in a given context, the set of criteria should validly represent that construct. Whereas CBA evaluates costs and consequences against an overarching criterion of Kaldor-Hicks efficiency, VFI does not specify overarching criteria, thereby enabling the merit, worth and significance of resource use to be contextually defined.

n) **VFI allows equity to be considered separately from efficiency** – for example, it can take the Kaldor-Hicks test into account without being limited to this position exclusively. It allows for efficiency and equity to be analysed as distinct and separate criteria, and for trade-offs between them to be made explicit.

o) **VFI does not require commensuration**: it can involve commensuration, but also has the ability to contrast and support deliberation on qualitative differences between criteria, groups of people and/or things of value.

p) **VFI does not take an exclusively consequentialist perspective**; it permits evaluation of design and implementation separately from consequences. A practical upshot of this is it can evaluate the merit, worth and significance of resource use during design (for developmental purposes), during implementation and delivery (for learning and improvement, or to assess the quality and value of processes), and/or in terms of outcomes (for summative purposes).

q) **VFI can incorporate a broader range of analytical options about value than CBA**: VFI has flexibility to accommodate holistic or analytic evaluation, using absolute (grading, rating, scoring) or relative (ranking, apportioning) systems for determining merit, worth and significance, together with quantitative and/or qualitative synthesis.

r) **VFI does not prescribe the methods to be used to gather evidence**; rather, it is flexible to enable study design and methods to be matched to context – recognising that no single method can address all VFM criteria and that no single method should be regarded as gold standard.

s) **VFI has the ability to accommodate mixed methods evidence** including quantitative and qualitative evidence – combining methods in planned and deliberate ways to strengthen data collection, analysis and interpretation through strategies such as triangulation and complementarity.

t) **VFI has the ability to incorporate economic evidence**, including the results of economic analysis, without being limited to economic criteria, methods and metrics alone.

u) **VFI and CBA are compatible**: Because VFI can incorporate economic analysis, VFI and CBA are not mutually exclusive or competing approaches – they can be combined, and doing so may strengthen evaluation of VFM compared to using either approach without the other.

v) **VFI can be conducted in full adherence with program evaluation standards**. Program evaluation standards should be used to guide any evaluation. In VFI their use should include guiding decisions about selection of methods, including when to use CBA and when not to. Of course, this does not guarantee that it will happen – but, unlike the case where CBA is used as the overarching method of evaluative reasoning, at least it is feasible for a VFI evaluation to fully adhere with such standards.

In the next chapter, these requirements are developed into a prototype for an operational model, providing a sequence of steps for implementing the theoretical model in evaluation practice.

Chapter 6: Process model

Introduction

Evaluation is a practical endeavour (Julnes, 2012c). To have utility in evaluation practice, the conceptual model described in the previous chapter needs to be implementable – and for this purpose it needs additional specification to translate the conceptual requirements into a practical process. Accordingly, this chapter addresses the third research question:

RQ3: How should the model be operationalised?

This chapter describes a prototype for a process model, setting out a series of steps and principles to guide the design and implementation of VFM evaluations.

This prototype is designed for a specific context: international development programs, where VFM assessment is often mandated but typically has not been carried out using evaluation methods. Specifically, the model applies the criteria of VFM used by the United Kingdom Government's Department for International Development (DFID) – though it would be a straightforward matter to substitute alternative criteria when applying the model. The model reflects the primary evaluation purposes driving VFM assessment in DFID programs (accountability and learning).

First, the international development context is introduced – with a particular focus on the DFID context and common challenges in evaluating VFM in this setting. Second, the procedures employed in developing the VFI model are summarised. Finally the process model is presented.

Context: VFM in international development programs

In international development, because of limited aid budgets and political pressures to be accountable for the use of taxpayers' funds, it is accepted that aid should be well targeted and managed effectively (Adou, 2016). These drivers have led to an increased interest in VFM (Fleming, 2013). VFM assessment, however, is challenging for a number of reasons, which are summarised here. Firstly, there is no universal definition of VFM (King, 2017), nor are there any internationally standardised guidelines on how to approach VFM in programming, hence how to apply VFM methods continues to be a field of debate (Renard & Lister, 2015).

Economic evaluation – principally, cost-benefit analysis (CBA) – may be applied *ex-ante* to inform a business case for a new program, applying scenario analysis to assess potential worth and inform funding decisions. However, the assumptions behind the original economic case are not routinely checked during the life of the program (ICAI, 2018). Furthermore, CBA is often not feasible for *ex-post* VFM appraisal, for reasons well-covered in the literature. For example, some outcomes are hard to monetise, and it is difficult to incorporate equity in CBA.

VFM in international development has typically not been assessed using evaluation methods. Donor requirements tend to emphasise the use of linear project management, monitoring and accountability frameworks such as the Logical Framework or logframe (Schiere, 2016) and "a persisting bias towards quantitative methods and 'hard' data" (Emmi, Ozlem, Maja, Ilan & Florian, 2011, p. 16), despite an interest in the social or 'soft' impacts of aid investments.

VFM assessments have similarly tended to follow a linear, indicator-based approach. There is a risk of making invalid assessments of VFM if the approach used relies on a narrow set of indicators drawn from a logframe, in the absence of evaluative reasoning. For example, such an assessment could focus on activities that are easy to measure but relatively unimportant. Similarly, there is a risk that such an assessment could focus on quantification of outputs and outcomes at the expense of more nuanced consideration of their quality and value (King & Guimaraes, 2016).

DFID has established itself as "a global champion on VFM" (ICAI, 2018, p. i), advocating for VFM with multilateral agencies and partner organisations, and embedding routine VFM assessment into its performance management processes for all DFID-funded programs (DFID, 2011). Other UK-based organisations, including the Independent Commission for Aid Impact (ICAI, 2011) and Itad (Barr & Christie, 2014), have produced further guidance expanding on DFID's approach. These various guides set out high level principles for VFM assessment but fall short of providing a blueprint approach for evaluating VFM. In particular, although both DFID's (2011) and ICAI's (2011) documents acknowledge that VFM assessment involves making judgements from evidence, they do not set out an explicit process for making those judgements.

A review of DFID's VFM approach (ICAI, 2018) found that the approach in practice tended to emphasise "controlling costs and holding implementers to account for efficient delivery" and that annual reviews focused principally on progress at output level, and not to achievement of outcomes or VFM as a whole. DFID's VFM criteria relate not only to efficiency but also to equity. Without clear criteria and standards for equity however, there is a risk that funding allocation decisions will systematically disadvantage fragile states where change can cost more to achieve (Jackson, 2012).

Moreover, the approaches typically taken to assess VFM have struggled to accommodate the adaptive and incremental approaches that are commonly used on reform programs to gain traction (Andrews, 2013). This is a particular challenge in complex development programs, as "those development programs that are most precisely and easily measured are the least transformational, and those programs that are the most transformational are the least measurable" (Andrew Natsios, former head of US AID, as quoted in Emmi et al., 2011, p. 16). ICAI (2018) similarly noted that despite DFID's support for adaptive programming, its approach

to VFM had yet to reflect the importance of experimenting and adapting in achieving VFM.

VFM assessment has tended to serve oversight and accountability purposes, though its deeper potential is recognised. VFM assessment can help not only to promote transparency and accountability, but can also serve as a helpful communication tool, project management tool, and can help increase learning within projects (Jackson, 2012; Schiere 2016).

VFM assessment has tended to focus on the donor's perspective, with a focus on donor resources spent and the achievement of outputs and outcomes specified by the donor. Jackson (2012) pointed out that when assessing VFM it is important to understand for whom VFM is being assessed. Is VFM primarily a donor concern as they are likely to be held accountable by their funders (e.g. taxpayers)? And will taxpayers' understanding of VFM necessarily correlate with the intended beneficiaries' understanding of VFM and do they share the same time perspective? DFID's (2016) Bilateral Review stated that "beneficiary feedback has become part of the fabric of good program delivery" (p. 46) – but this expectation has not yet filtered through to VFM assessments.

Comparisons of VFM achievements across projects remain challenging due to the broad and diverse way in which the concept is applied (Barr & Christie, 2014; Jackson, 2012; Schiere 2016). Moreover, a number of limitations hinder the assessment of VFM in development contexts – including the lack of reliable statistical data, lack of capacity of development aid beneficiaries, non-harmonisation of methods and criteria, and the absence of clear guidelines (Adou, 2016).

The conceptual model offers potential to address these challenges. Theoretical propositions suggest that explicit evaluative reasoning, with mixed methods, can accommodate qualitative and quantitative evidence, and provide a transparent basis for making explicit evaluative judgements. With these considerations in mind, a process model was developed to translate the requirements of the conceptual model into a process that could be applied to evaluate VFM in DFID-funded development programs. The methods employed are described as follows.

Methods

A planned action model (Nilsen, 2015) was developed, setting out a sequence of steps for planning and undertaking an evaluation of VFM, while acknowledging that the process may not be purely sequential and may involve some iteration between the steps.

The starting point for developing the model was to identify the broad sequence of steps that should be involved in developing and undertaking a VFM evaluation. As the conceptual model proposes that evaluative reasoning should be applied, the general sequence of steps associated with the general logic of evaluation (Fournier, 1995) provided a foundation for developing the stepped model. The initial 'concept sketch'

of the process model remained open to revision during the process that followed.

Having identified an appropriate sequence of steps, the second stage of the model development process was to provide guidance within each step, summarising the purpose and rationale for each step (and, where germane, the rationale for its position within the overall sequence) together with a description of the tasks involved in completing each step. In alignment with the theorising strategy of engaged scholarship (Shepherd & Suddaby, 2017), development of the process model and its components involved amalgamating evaluation literature with practitioner experience. Theory-building strategies of blending, bricolage, thought experiments, and problematising (Shepherd & Suddaby, 2017) supported the model-building process, informed by literature on evaluative reasoning, program evaluation, and economic evaluation.

Model development was guided by the principle of parsimony, seeking to specify as few parameters as are necessary and sufficient to provide an adequate framework for implementing the requirements of the conceptual model. Within the broad confines of these parameters, the intent was to develop a model that would be sufficiently broad and flexible to accommodate a range of evaluation worldviews, methods and tools, in the expectation that these should vary according to program context as well as evaluator orientations and skills (Schwandt, 2015).

The use of this model requires, and assumes, that the evaluator brings the requisite skills and experience to design and conduct an evaluation. The model does not purport to be a full evaluation manual or instructional document – rather, it charts a process that should be followed for implementing the requirements of the theoretical model. The process model draws attention to the use of evaluative reasoning and contextualised method selection to evaluate VFM, and contributes a series of steps to guide methodological decisions by a professional evaluation team well-versed in evaluation theory and practice, including economic methods of evaluation.

Results

The VFI model focuses on the implementation of the four core theoretical requirements: pose and answer an evaluative question about the merit, worth or significance of resource use; use explicit evaluative reasoning; use mixed methods including economic evaluation where feasible and appropriate; and adhere with evaluation standards.

A flexible approach is adopted, seeking to accommodate (without prescribing) the diverse methodologies and tools found in program evaluation, which may be used in situationally determined ways to complement and enhance the overarching process of evaluative reasoning. A stepped approach is described as follows.

A stepped approach

The use of stepped models in evaluation dates back to Tyler's work in the 1930s (Gargani, 2014) and the work of Suchman in the 1960s (Coryn & Stufflebeam, 2014) so has lineage to the evaluation literature as well as implementation science (Nilsen, 2015). This process model builds on Fournier's (1995) four steps describing the general logic of evaluation (Scriven, 1980; 1991; 1994; 1995; 2012). These four steps are: establishing criteria of merit; constructing standards; measuring performance and comparing with standards; and synthesising and integrating data into a judgement of merit or worth. Additional steps are included by disaggregating Fournier's model to describe a more detailed process of evaluation design and implementation (Davidson, 2005; 2014; King et al., 2013; McKegg et al., 2018; Wehipeihana et al., 2018).

The resultant VFI model comprises eight steps: understand the program; identify VFM criteria; set VFM standards; determine evidence requirements and associated methods; gather evidence; analyse each stream of evidence separately; synthesise the streams of evidence to make an overall judgement about VFM; and communicate findings (Figure 4).

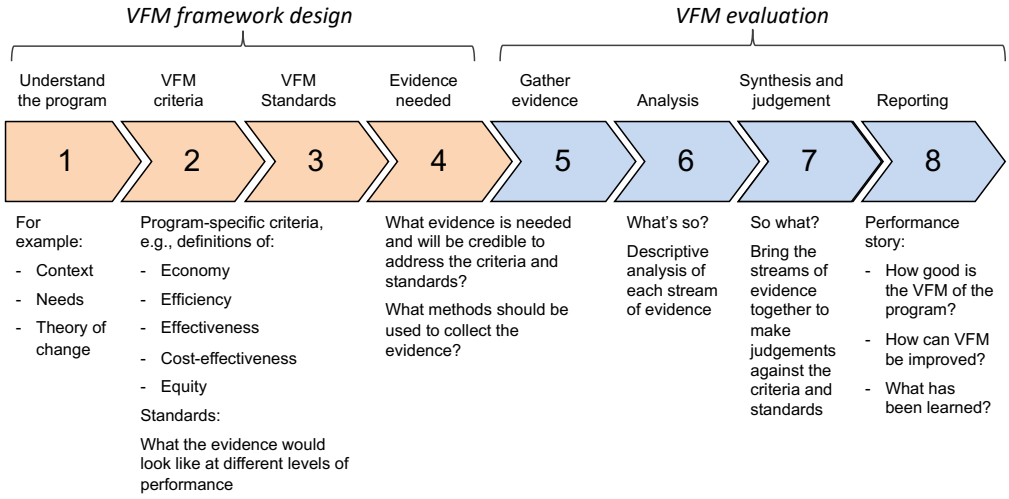

Figure 4: Overview of prototype process model

The following paragraphs summarise each step of the model including the purpose and rationale of each step, and the tasks involved. These steps may involve some iteration – for example, insights that emerge during analysis, synthesis and reporting might prompt revisions to criteria and standards.

Step 1: Understand the program

An important early step in any evaluation is to understand the program or policy being evaluated, including its context, stakeholders, and information needs (Scriven, 2013; Patton, 2008). This is prerequisite to identifying the criteria against which it should be evaluated (Scriven, 1980; 1981). How the program or policy is viewed, furthermore, "leads to

a position on what evaluation approach should be taken" (Fournier, 1995, p.23).

For example, a theory of change can be used as a tool to articulate a clear and shared understanding of what needs the program is intended to meet and how it is intended to function. A theory of change "explains how activities are understood to produce a series of results that contribute to achieving the final intended impacts" (Rogers, 2014, p. 1). A theory of change can be developed in a participatory manner (Funnell & Rogers, 2011), with this process of development being intentionally used as an opportunity to build an evaluation-focused relationship among the relevant stakeholders, reach an agreed understanding of the program or policy, and to foster evaluation ownership and use. A theory of change may also be used as a tool when assessing causality or contribution (Funnell & Rogers, 2011).

Step 2: VFM criteria

Criteria are determined on the basis of values (Schwandt, 2015). The way these criteria are defined has an important bearing on "how evaluators reason toward evaluative judgments" – and therefore, "the burden of justification rests on the criteria" (Fournier, 1995, p. 22). Accordingly, the model emphasises careful specification of criteria, as the second step in the VFM design process.

VFM criteria represent the dimensions of merit, worth or significance that are relevant to the VFM of a particular program in a particular context. These VFM criteria, as already noted, may include aspects of resource use (such as relevance, affordability, ethicality, frugality); consequences of the resource use (such as delivery of outputs, achievement of objectives, meeting identified needs, unintended consequences); and whether the resource use is justified (such as economic efficiency, how benefits are distributed and whether they reach the intended groups, sustainability, scientific value, environmental value, or cultural fit). This list is neither prescriptive nor exhaustive; as in program evaluation generally, criteria should be contextually determined (Schwandt, 2015).

In the specific context of DFID-funded development programs, the required criteria of VFM are defined as economy, efficiency, effectiveness, equity, and cost-effectiveness (outcome efficiency) (DFID, 2011).[2] DFID's definitions of these terms were summarised earlier (see *Background*). The VFI model entails the translation of DFID's generic criteria into program- and context-specific definitions. These definitions may include sub-criteria, reflecting the content of the policy or program and focusing on relevant aspects of performance.

Criteria can be developed with reference to multiple sources, including needs and stakeholder values. A theory of change can be used as a point

[2] DFID's definition of cost-effectiveness refers to the general notion of outcome efficiency (e.g., a comparison of inputs to impacts) and does not mandate application of the specific economic method of cost-effectiveness analysis.

of reference to facilitate the identification of a logical and complete set of criteria that are aligned with the intended functioning of the program or policy (Davidson, 2005). In the DFID context, for which this process model is tailored, the use of a theory of change enables definitions of the program to be explicitly aligned with DFIDs five criteria of VFM – which, as shown earlier (Figure 2), are conceptualised as discrete sections of a simplified results chain. Specifying inputs, outputs, and outcomes in a theory of change facilitates the definition of a coherent set of VFM criteria.

Step 3: VFM standards

Standards define levels of performance for each VFM criterion, which may be labelled 'excellent', 'good', 'adequate', and 'poor' or in various other ways (Dickinson & Adams, 2017; Martens, 2018b). There can be any number of levels; however, Davidson (2005) suggested that up to six levels is workable, with few gains in precision beyond that point.

At this point it is also necessary to determine what overarching process of evaluative reasoning should be used to synthesise the criteria, standards and evidence to reach an evaluative judgement. As indicated in the conceptual model, qualitative weighting and synthesis will usually be the preferred approach to evaluative reasoning in a VFM evaluation of a social policy or program. Exceptions are possible where the conditions are met for numerical weighting and synthesis (NWS) to be valid approaches to evaluative reasoning. NWS may be a viable option where there is an empirical basis to justify weights and sound mathematical logic is possible – for example, sufficiently few criteria to avoid swamping and mutually exclusive criteria to avoid interactions between criteria. CBA, viewed here as a sophisticated, cost-inclusive form of NWS, should only be used as the approach to evaluative reasoning if: maximising aggregate welfare is the sole criterion of VFM; all relevant and material values can be fairly and accurately weighted monetarily; qualitative distinctions not reflected in monetary valuations are unimportant; aggregation of values is appropriate; and only costs and consequences (not processes) matter.

Step 4: Identifying evidence required and selection of methods

Criteria and standards represent a specification of "what the mix of evidence will look like at different levels of performance" (Davidson, 2014, p. 6). Accordingly, criteria should be systematically analysed to determine what evidence is needed and will be credible in the evaluation (King et al., 2013). In turn, these evidence needs provide the basis for selecting suitable methods to gather the necessary evidence. From a practical standpoint, these steps also involve considering constraints such as time frames, budgets and existing data sources to ensure the methods selected are feasible and proportionate with the value and importance of the evaluation (Bamberger et al., 2011; Levin & McEwan, 2001).

By following the sequence described in this model – starting with understanding the program before moving on to develop criteria and standards, and then using the criteria and standards to inform the

selection of metrics and methods – the intent is to support validity in constructs and methods: the values embedded in the criteria and standards should cohere in a logical way with the program, and the evidence gathered should cohere with the criteria and standards (King et al., 2013).

Method selection should be contextual and negotiated (Montrosse-Moorhead, Griffith, & Pokorny, 2014; Patton, 2018; Yarbrough et al., 2011). The VFI model is flexible to accommodate the best combination of methods given the circumstances of each evaluation, to support validity, utility and credibility.

Among the methods that may be considered for an evaluation of VFM, the model emphasises the potential to incorporate economic evaluation. Economic methods of evaluation are candidates for inclusion where efficiency is a relevant criterion. CBA is a candidate for inclusion where: total welfare or net benefit is a relevant subcriterion; a summation of monetisable costs and consequences is likely to give a sufficiently robust estimate of total welfare; and impacts can meaningfully be monetised and aggregated. CEA or CUA are candidates for inclusion where: two or more alternatives are being compared; comparative outcome efficiency (cost-effectiveness or cost-utility) is a relevant subcriterion; and it is feasible to derive sufficiently robust estimates of impacts (and, in the case of CUA, impacts can be converted into reliable utility weights). If economic methods are included, the possibility of disaggregated economic analysis should also be considered to investigate differences in the incidence of costs and consequences between subgroups.

Collectively, the first four steps of the model are the steps involved in VFM framework design. Upon completion of step 4, a written VFM framework should be prepared. As with any evaluation framework, the VFM framework should describe: the program and context (including the theory of change); the criteria and standards that will be used to evaluate VFM; the design and methods that will be used to gather evidence; and a project plan and timetable for conducting VFM evaluations (McKegg et al., 2018).

It is an important feature of the model that the criteria and standards are determined in advance of the VFM assessment. The criteria and standards should represent an agreed and transparent basis for making evaluative judgements. Accordingly, stakeholder approval of a written evaluation framework is an important prerequisite before commencing steps 5-8: conducting a VFM assessment. These steps are described as follows.

Step 5: Gathering evidence

Gathering the evidence needed for a VFM evaluation involves following the technical and ethical principles of good social science research that apply to the selected study design and methods (Coryn & Stufflebeam, 2014; Donaldson et al., 2015; Drummond et al., 2005; Schwandt, 2015; Tolich & Davidson, 2018). This model is intended to accommodate any

methods or mix of methods determined to be appropriate by the evaluator and stakeholders. Within this model, the evaluator may employ mixed methods (Greene, 2007) and therefore the evaluation may involve some iteration between steps 4 and 5 – for example, where the results of one method inform the development of another.

Step 6: Analysis

Analysis of each individual stream of evidence should be carried out to separately identify findings and themes from each method or source. This step again makes use of accepted principles and practices in social science research (Coryn & Stufflebeam, 2014; Donaldson et al., 2005; Drummond et al., 2005; Schwandt, 2015; Tolich & Davidson, 2018). Illustratively, and without prescribing or limiting evaluation scope, the analysis step might include: financial analysis of program costs; thematic analysis of interview feedback; documentary analysis to summarise information from monitoring reports; quantitative analysis of survey results; observations from field visits; impact estimates from a randomised controlled trial; and economic analysis of the value of outcomes relative to costs.

Step 7: Synthesis and judgements

The synthesis step is an explicitly evaluative step because it goes beyond description to making inferences of merit, worth and significance (Davidson, 2005). It involves considering the totality of evidence collected and the results of each stream of analysis, including any areas of corroboration or contradiction between sources. Judgements should then be made by comparing the evidence of performance to the criteria and standards to determine the level of performance indicated by the evidence (Oakden & King, 2018).

The synthesis step can be broken into two sub-steps. First, synthesis and judgement-making should be carried out for each of DFID's five VFM criteria individually – resulting in judgements of performance for economy, efficiency, effectiveness, cost-effectiveness, and equity respectively. Second, a judgement should be made for VFM overall by synthesising the individual judgements made for the five criteria.

Such synthesis could be carried out using predetermined weights. For example, a numerical or qualitative weighting system could be devised to determine the relative importance of the criteria to the overall judgement of VFM – or, in the absence of clear rationale for differentially weighting the criteria, a straight average could be taken. In this model, however, it is argued that the overall determination of VFM should be a matter for evaluator and stakeholder judgement, on the basis of well-reasoned argument (Patton, 2018; Schwandt, 2015). This is suggested on the basis that the relative weighting of different criteria will vary according to the maturity of the program and evolving context. For example, economy and efficiency may receive all of the weight in the early stages of a program when it is too soon to evaluate outcomes. Later, as evidence of

effectiveness, cost-effectiveness and equity becomes available, these factors will progressively become more important. The VFM assessment should explain the rationale for the overall determination of VFM, including which criteria were emphasised in the assessment and why.

This model enables judgements to be made about both the VFM of the program overall, and opportunities to improve VFM. Systematic comparison of different aspects of performance against the criteria and standards can expose strengths, weaknesses, and opportunities to improve.

In this model, judgement-making is not necessarily the exclusive domain of the professional evaluator; the use of criteria and standards to guide interpretation of the evidence can facilitate processes of judgement-making in collaboration with stakeholders such as program management, internal monitoring and evaluation teams, donors, beneficiaries, or some combination of these. Such processes can enhance validity, credibility, understanding, ownership and use of evaluation findings (King et al., 2013; Wehipeihana et al., 2018).

Step 8: Reporting

This model is built on an overarching requirement of explicit evaluative reasoning. The preceding steps describe an approach to evaluative reasoning in which clear definitions of criteria (dimensions of performance and VFM) and standards (levels of performance and VFM) inform decisions about the collection and analysis of evidence, and are subsequently used as a framework for making judgements about performance and VFM from the evidence. The 'explicit' part of 'explicit evaluative reasoning' demands that findings – including evidence, evaluative judgements, and the reasoning linking the two – be clearly communicated.

Findings from the VFM evaluation should be reported to the relevant audiences. The point of reporting is "to communicate effectively and accurately the evaluation's findings in a timely manner to interested and right-to-know audiences and to foster effective uses of evaluation findings" (Coryn & Stufflebeam, 2014, p. 15). The report should provide a clear answer to the evaluation question (Davidson, 2014). Judgements should be transparently presented by setting out the evidence and reasoning that leads to each judgement, in a logical sequence. The model proposes accordingly that findings should be structured around the overarching VFM criteria, addressing each criterion in turn, starting with the judgement against the rubric, and then summarising the evidence and providing any further discussion of reasoning, additional evidence, or context that is necessary and sufficient to support the conclusions reached (Wehipeihana et al., 2018).

Summary

A process model has been developed from literature and practice-based knowledge, setting out a series of steps for designing and conducting a VFM evaluation in such a way that the requirements of the conceptual model can be met. The model presented here is a prototype. It sets out the operational elements needed for an experienced evaluator to conduct a VFM evaluation that adheres to the requirements of the conceptual model.

The first of these requirements was that an evaluation of VFM should pose, and answer, an evaluative question, such as: *To what extent does the program represent value for the resources invested, and how can its value be improved?* Addressing such a question requires evaluative judgements based on a transparent, logical set of values and evidence.

The second requirement is that explicit evaluative reasoning be followed, including the use of criteria and standards to provide a transparent basis for making the evaluative judgements. The stepped model builds and expands on the four steps described by Fournier (1995) which describe a way to implement the general logic of evaluation.

The third requirement is that the evaluation match methods to context. The model facilitates the selection and use of an appropriate combination of methods by using the criteria and standards as the basis for identifying what evidence is needed and will be credible to support evaluative judgements, and by providing a structure for analysing and synthesising diverse streams of evidence. Economic evaluation should be included, where feasible and appropriate, recognising the value of this set of methods in estimating measures of efficiency.

The fourth requirement is that an evaluation of VFM, like any program evaluation, should be conducted in adherence with program evaluation standards, in order to ensure that the evaluation is conducted with due consideration of utility, feasibility, propriety, accuracy, and accountability to evaluation stakeholders.

An evaluation of VFM conducted with fidelity to this sequence of eight steps (which may be applied iteratively) and in fulfilment of the requirements of the conceptual model, should result in an evaluation of VFM that reflects stakeholder values, uses an appropriate mix of credible evidence, and provides robust, traceable judgements about VFM to support evaluation use in decision making.

In the next chapter, this operational model is put to the test through two case studies. The conceptual quality of the theoretical propositions are systematically evaluated, and potential refinements to the model are identified.

Chapter 7: Case studies

Introduction

This chapter investigates two instances where evaluations of VFM were conducted with fidelity to the conceptual and process models, to learn whether this model of VFM assessment worked in practice and whether its theoretical propositions were applicable – that is, conceptually valid, relevant and appropriate. This chapter uses case studies to intensively and rigorously address the fourth research question:

RQ4: To what extent are the model's theoretical propositions applicable in evaluation in real-world contexts?

In this chapter, two value for money (VFM) evaluations are presented as case studies. These evaluations applied the conceptual and process models described earlier (a 'Value for Investment' or VFI approach), to assess VFM. The first application of the approach, in an international development context, was in the MUVA female economic empowerment program in Mozambique (case study 1). The second was the Pakistan Sub-National Governance (SNG) Program (case study 2).

The purpose of the case studies is threefold. First, they provide concrete illustrations of the use of the VFI model in practice. Second, each case study is analysed to examine the applicability of the theoretical propositions, and the two case studies are analysed together to assess replication of findings. Third, observations from the case studies are used to refine the VFI model.

Methods

The approach to the case studies followed the general sequence of steps described by Yin (2009). First, theory was developed based on critical review of the literature. The theoretical propositions from the first two studies provide the blueprint for the study. Second, cases were selected. The two cases are revelatory in that they are the first two instances where the process model was applied. OPM consented to the research and obtained DFID consent to use the two programs as case studies.

Third, documentary analysis was undertaken through collation and thematic review of the source documents. A predetermined structure for the case studies was prepared, observing Miller's (2010) recommendations that cases for testing evaluation theories should include relevant details of evaluation purpose, the rationale for applying the theoretical approach, how the theory was applied, chronology of the evaluation, what outcomes were expected and how they were substantiated. Copies of all relevant documents were obtained. Content analysis of documents was carried out, with information extracted from the documents being coded thematically according to the case study structure and theoretical propositions.

Fourth, case study analysis and writing were undertaken in an iterative manner. Written vignettes of the two cases were prepared, to a common structure, summarising the application of the process model. Cases were systematically analysed against the theoretical propositions. Cross-case conclusions were drawn from content analysis of the two case studies together. For each sub-proposition, commentary was prepared on the extent to which the proposition was applicable (or not) within the case, the limitations of the case in demonstrating the proposition, and a summative rating of whether, and to what extent, the case corroborates or does not corroborate the sub-proposition. Draft case studies were prepared, including summaries of the programs, evaluation design and VFM evaluations, and analysis against the theoretical propositions.

Fifth, draft case studies were independently reviewed by one member from each program team who had been involved in the design and implementation of the respective VFM evaluations. The review focused on validating the fidelity of VFM design and conduct to the model, and the accuracy of the coding of case studies against the theoretical propositions. Revisions were made as appropriate, responding to reviewer feedback.

Sixth, systematic analysis of the two case studies was undertaken to triangulate and determine the extent to which findings were replicated across the two assessments, in relation to the theoretical propositions.

Seventh, reflective workshops were facilitated with evaluators and program leaders from the MUVA and SNG programs to consider what had been learned from the application of this model of VFM evaluation, what had worked well, what had not worked well or been challenging, and how the model could be refined to support future practice. From these reflections, opportunities for refinements to the model were identified.

The remainder of this chapter presents each case study in turn. Each case study starts with a summary of the program and context. Next, an overview is provided of the VFM design and the conduct of the evaluation. After summarising these matters, each case study sets out a systematic analysis of the theoretical propositions. Following the sequential presentation and analysis of the two case studies, replication is appraised and overall conclusions are drawn.

Case study 1: MUVA female economic empowerment program

This case study centres on MUVA (formerly called Ligada), a female economic empowerment program in urban Mozambique. MUVA is six-year program funded by the UK Department for International Development (DFID), with the aims of improving female economic empowerment in urban Mozambique and identifying, testing, and supporting the uptake of solutions to barriers that exclude women from decent work opportunities.

A VFM framework was developed during May-June 2016, at the end of the program's inception phase (which ran from August 2015 to April 2016). In

accordance with requirements set by DFID, the first VFM assessment covered the first eight months of implementation, from May to December 2016. The second VFM assessment covered the four months from January to April 2017, aligning the VFM reporting date with DFID's annual review cycle for the MUVA program. The third VFM assessment covered the annual period May 2016 to April 2018 and contributed part of the evidence toward DFID's annual review. Subsequent VFM assessments will be conducted on an annual basis.

Program and context

In Mozambique, women have poorer prospects than men in formal employment and in setting up and running businesses. Women also have lower rates of school completion and university enrolment than men. Girls and women also face discrimination in access to work and are the main victims of urban and domestic violence.

The inception report for the program (OPM, 2016) noted that, although DFID was committed to investing in female economic empowerment, there was a lack of evidence, and a lack of innovation, to address barriers to female economic empowerment. Across the developing world, girls and women bear a disproportionate burden of poverty while evidence shows that investing in girls and women is transformational not just for their own lives but for the families, communities, and societies. MUVA was designed to innovate, work with the private and public sectors and identify successful pathways to female economic empowerment that could be scaled up through partners. By working in this way, MUVA was intended to contribute evidence on successful pathways for female economic empowerment.

MUVA has three overarching objectives: to identify, test and ensure adoption of sustainable solutions to urban female economic empowerment, taking account of both supply side and demand side constraints; to generate evidence and capacity on gender, to drive improvements in policy, programming and budgets (including among donors, the government of Mozambique and the private sector); and to provide a delivery platform for adaptive programs that includes an embedded monitoring, evaluation and learning component, both for the female economic empowerment approaches, and for the program management and delivery approach.

A logframe was used as a key management, reporting and accountability tool, as is generally the case in DFID-funded programs. The logframe was prepared in matrix form, specifying outputs, outcomes and impacts aligned with the program theory of change, together with indicators and milestone targets for tracking progress at specified points in time. MUVA's operational model involved developing and testing projects (a set of activities labelled "output 1" in the program's logframe), providing credible evidence about what works in female economic empowerment in an urban Mozambique context ("output 2") and influencing other organisations to take successful approaches to scale ("output 3").

In delivering these three outputs, the program was designed to support innovation, using a lean and agile approach to design and test new ideas in short cycles of testing and reflection. It took an adaptive approach to program management to enable it to be flexible in response to the context and the results of evaluations.

A Decision Unit, comprising DFID and OPM program leaders, was established to make resource allocation decisions within the program's overall budgetary envelope for output 1, including approving project proposals and deciding when to stop projects (e.g., if they were not gaining traction, showing insufficient impact relative to costs, or if, despite showing some success, did not appear to be scalable).

The design and intent of the MUVA program presented some challenges for VFM assessment, as follows – and these challenges led to the program management team's adoption of the VFM approach described here.

First, MUVA's objectives included trying new approaches to female economic empowerment that had not been tried before, and to provide evidence about what was successful, what was unsuccessful, and why. While some approaches would build on an existing evidence base, it was expected that others would push new boundaries and it was accepted that some approaches might not gain traction in the communities where they were being tested, or might be successfully implemented but fail to achieve their intended outcomes. In an economic evaluation, outcome evaluation, or standard logframe reporting system, it was perceived that such approaches may be judged as 'failures'. However, as MUVA's objectives explicitly included learning from both successful and unsuccessful trials, the VFM assessment needed a way of recognising the value of 'failure' in a learning program.

Second, and also related to MUVA's learning objectives, the program made a substantial investment in a bespoke monitoring, evaluation and learning system, intended to support management through internal program monitoring, reflection and lesson learning to improve the program across all dimensions of implementation. The monitoring and evaluation team was expected to engage with a wide range of stakeholders and reflect on program progress and performance – including adolescent girls and young women from urban Mozambique in a way that increases their voice and empowerment. The information provided by the monitoring and evaluation team was regarded as crucial, both for accountability to key stakeholders and to generate lessons in program implementation and outcomes.

In keeping with the importance and intended function of monitoring, evaluation and learning within the program, the investment in these activities was larger than would generally be considered the norm in a DFID program, and the VFM assessment needed to examine the value derived from this investment. MUVA's intended function as a learning program extended to learning how much it would cost to achieve outcomes in different interventions, and it was expected that the VFM

assessment would capture this learning. It was also anticipated that cost analysis in MUVA would start to create benchmarks for similar social norm change programs.

Third, MUVA was designed to be an adaptive program, explicitly building on the Problem Driven Iterative Adaptation (PDIA) model of Andrews and colleagues (Andrews, 2010; Andrews, 2013; Andrews, Pritchett & Woolcock, 2017). The PDIA framework puts forward four key principles for undertaking development work in the face of complexity: i) focus on particular problems in local contexts, nominated and prioritised by local actors; ii) foster an interactive approach to experimenting, learning and problem solving; iii) foster an 'authorising environment' for decision making, which encourages experimentation; and iv) engage broadly to facilitate the viability, legitimacy and relevance of reforms (Andrews, Pritchett & Woolcock, 2017). This approach to adaptive management was expected to improve the effectiveness and efficiency of the program, through short cycles of testing, reflection and adaptation that would help to incrementally improve the interventions that were being developed and tested. However, it also presented a potential challenge for VFM assessment: cost:output or cost:outcome ratios, being averages over a defined time period, may not fully accommodate the dynamic nature of an adaptive program.

Fourth, MUVA's long-term intended outcome was to influence local organisations, stakeholders and social norms in urban Mozambique. It was accepted that these long-term outcomes would not become fully apparent during the six-year duration of program delivery (the period during which VFM assessments were required). They may be subject to longer-term outcome evaluation – but in the meantime, the VFM assessment would need to assess intermediate actions being taken and short-term results that may be expected to contribute to MUVA's long term success as an influencing program.

Fifth, female economic empowerment interventions were explicitly targeting a range of outcomes, some of which could readily be valued monetarily (e.g., increased earnings associated with enhanced job skills and employability of young women in urban Mozambique who participated in MUVA's interventions), and some of which were described as 'softer' outcomes (e.g., quality of life, agency, and self-worth). It was anticipated that these outcomes, though inherently valuable for beneficiaries, would be difficult to credibly monetise for inclusion in a cost-benefit analysis.

OPM engaged the author as an external consultant to develop a VFM framework for the MUVA program. The VFM framework followed the VFI process model and conceptual model as already described. The following paragraphs describe the sequence of steps that were followed in developing the VFM framework, and the features of the resulting framework.

Developing the VFM framework

The VFM framework was designed and developed in a participatory manner with the MUVA team, in Maputo, Mozambique, during May-June 2016. The VFM design and assessment followed the 8-step process specified in the process model. First, in order to ensure a clear and shared understanding of the program, its theory of change was reviewed. Then, criteria and standards were developed which were aligned with the theory of change and provided an agreed basis for making judgements about performance and VFM. Subsequently, sources of evidence were identified that were needed to address the criteria and standards. Following these evaluation design steps, the VFM assessment was carried out by gathering evidence, analysing each stream of evidence separately, and then synthesising the findings to make judgements against the criteria and standards. These steps are detailed in the following paragraphs.

Understand the program

The evaluation design process commenced with a review of program documentation including its business case, inception report and related documents, as well as a briefing for the evaluators from the program leaders. In this instance, the theory of change provided an important point of reference for defining the program and linking key features of its design and delivery to the VFM evaluation design.

In the MUVA program, a comprehensive theory of change had already been developed by an external evaluator, as one of the core deliverables of the inception phase. The MUVA theory of change had been developed prior to VFM framework development, through a participatory process, with the local MUVA team and partners (including organisations that would be involved in implementing approaches to female economic empowerment and organisations providing technical assistance to MUVA), to clarify the outcomes sought and the processes that would contribute to those outcomes.

This investment in a robust theory of change was advantageous for VFM framework development, because it provided a clear summary of the program theory and its underlying assumptions, developed with, understood and endorsed by the many international and local actors who had a stake in designing and implementing the MUVA program. The theory of change had also underpinned the development and alignment of the logframe, the program's monitoring, evaluation and learning system, and individual project proposals and theories of change under the MUVA program umbrella.

The Ligada theory of change is illustrated in Figure 5 (the program's name was subsequently changed to MUVA in 2017, at the conclusion of a brand development exercise). The theory of change articulated the program's intended long-term impact as: 'Urban disadvantaged adolescent girls and young women have enhanced capacity and agency to make use of decent economic opportunities in an increasingly supportive environment'. As an

intermediary outcome towards this long-term impact, it was intended that: 'Stakeholders take up and champion approaches and Ligada's vision to improve female economic empowerment, sustainably and at scale, based on evidence and learning from Ligada'.

In order to achieve this outcome, the program had three key areas of delivery: testing locally driven approaches to female economic empowerment; learning (generating and disseminating credible evidence about the tested approaches); and brokering and influencing (with a view to fostering local adoption and scale-up of successful approaches as well as social norms with regard to gender roles in urban Mozambique). The inception report was clear that, while the program would directly contribute to the long-term impact by enhancing adolescent girls' and young women's skills and supporting them to enter the labour force (as part of testing locally-driven approaches to female economic empowerment), the main pathway to the impact would involve stakeholders taking up and/or championing the approaches and vision that emerge through Ligada – and that this would require approaches to female economic empowerment that stakeholders would be aware of, comprehend and believe are effective, scalable and sustainable, based on credible evidence.

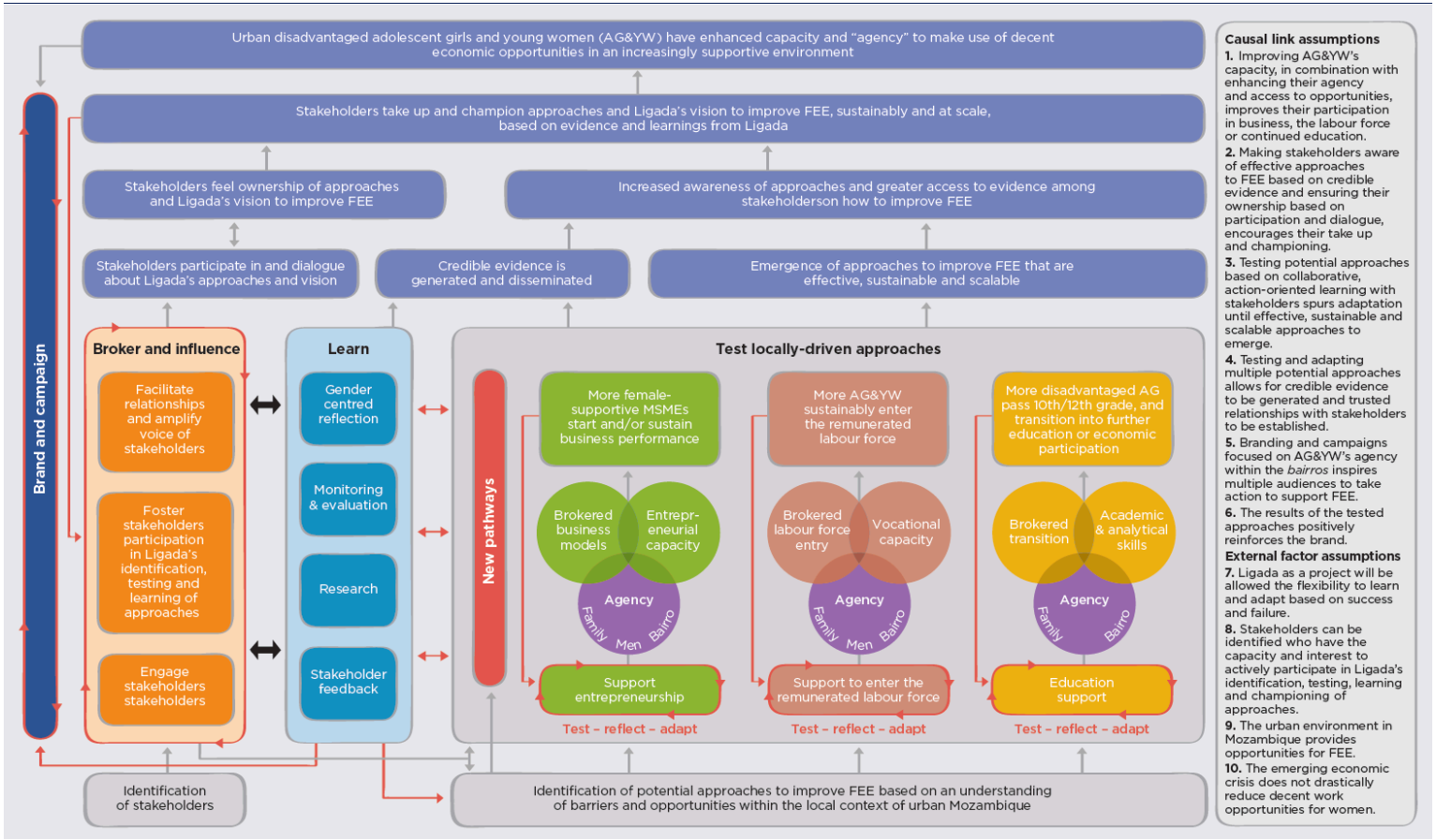

Figure 5: Theory of change for MUVA program. Reprinted from OPM (2016). Reprinted with permission.

Criteria and standards

VFM criteria, as described in the process model, are dimensions of program performance that are relevant to the overall determination of VFM. Standards are levels of performance. The model involves developing criteria and standards for each of DFID's VFM criteria (DFID, 2011) – i.e., economy, efficiency, effectiveness, cost-effectiveness, and equity.

Criteria and standards for the MUVA VFM framework were developed by the author, in alignment with the theory of change and in consultation with the Program Manager, Program Director, and workstream leaders including the Monitoring, Evaluation and Learning leader, and those team members responsible for designing and implementing the first round of interventions to be tested under the MUVA program umbrella.

In particular, the draft of the initial VFM framework went through several iterations between the author and the Program Manager, during the one-week duration of a visit to the MUVA offices in Maputo. Subsequently, the framework went through further revisions and refinements which were facilitated by email and teleconferencing with key MUVA staff.

Additionally, the author met with key DFID officials responsible for the management of the MUVA contract, to present the intended approach to VFM design and invite DFID's input into the design of the criteria and standards.

A draft of the VFM criteria and standards was submitted to, and reviewed by DFID, and a revised framework was produced taking DFID feedback into account. The final VFM criteria and standards were signed off by OPM and DFID prior to the first VFM assessment being carried out.

For each criterion, a program-specific definition was developed, as shown in the first column of Table 7. Effectiveness was differentiated across the three contracted outputs of the program, i.e., "effectiveness as an urban female economic empowerment program", "effectiveness as a learning program", and "effectiveness as an influencing program", with explicit definitions for each. This was considered important to ensure program performance would be explicitly assessed along each dimension, to assess the value that may be derived from each.

Each criterion was further supported by a set of identified indicators (a subset of the program's logframe indicators agreed to be relevant to the VFM assessment, with performance targets and milestones aligned with the logframe) and narrative evidence intended to supplement the indicators to support a nuanced assessment of performance and VFM.

A generic set of standards (Table 8) was developed which specified the levels of performance that would be applied to each criterion in turn.

Table 7: MUVA VFM Criteria from original VFM framework

Criteria	Indicators & narrative
Economy: The Ligada team manages project resources economically, buying inputs at the appropriate quality at the right price.	• % days worked national/international consultants • % management costs compared to the original contract • Average flight expenditure compared to the most up-to-date agreed targets (national/international) • Average DSA compared to budget • Narrative evidence of one-off savings through procurement process as stated in office manual
Efficiency: The Ligada team produces the intended quality and quantity of outputs, within the available resources.	• Number of stakeholders engaged through the influencing strategy • Length of time from proposal to signing of contracts for different approaches • Number of DFID programs and policy events informed and engaged • Workplan delivery as contracted or better: Narrative providing evidence of the projects being implemented as planned (on time, milestones met, targets met, within budget – includes all Logframe output indicators) • Direct engagement with target group: Numbers of adolescent girls and young women engaged (signed up and participating meaningfully) in each project (where measurable) • Indirect engagement: Total reach beyond direct engagement (subjective estimate with rationale/ examples) • Beneficiary feedback from MEL system on the quality of outputs • Actual costs per beneficiary are in line with expected costs per beneficiary
Equity: Ligada targets resources to the intended groups, including: • Clearly defining the intended target group for each intervention, with supporting rationale; and • Reaching the identified target groups.	• Narrative describing the target groups for each project and rationale for target groups (e.g., adolescent girls, young women, boys, men, and level of disadvantage/ marginalisation appropriate to intervention) • Evidence of projects reaching their intended target groups • Proportion of approaches that directly address vulnerable groups • Percentage of beneficiaries who are women. *Note: equity also means achieving outcomes with the identified target groups that improve equity. In this framework, these outcomes are considered as part of the 'effectiveness' criterion, with the rationale that, having identified and reached the appropriate target groups, Ligada's projects will improve equity when they are effective in achieving their intended outcomes.*
Effectiveness as an urban female economic empowerment program: • Ligada's low-risk approaches make their intended contributions to	• Narrative summaries of project-level outcomes for low-risk approaches. *Note: These outcomes are defined in project-specific Theories of Change. Indicators for these outcomes are currently being developed. The performance of*

Criteria	Indicators & narrative
capacity and/or agency and/or opportunity as defined in project-level Theories of Change, and are scalable.	*each project will be monitored and reported individually. The standards below will be used to evaluate each project's outcomes and support a judgement about Ligada's effectiveness overall as an urban female empowerment program. As a general principle, projects would be considered to performing at the 'good VFM' level if they are making a difference for around one-third of participants.*
Effectiveness as a learning program: • Local participation, relationships and knowledge contribute to project development; • Reflective learning processes are occurring as intended; and • Ligada provides credible evidence about the effectiveness of every project, including evidence to enable decisions about which projects deliver better results.	• Narrative summaries of local participation, relationships and knowledge contributing to project development. • Number of reports on reflective meetings completed in a timely manner with evidence of MEL feedback positive impact loops. • Narrative summaries demonstrating that there is credible evidence about the effectiveness of every project (i.e., whether they are effective and why/why not). *Note: Credibility of evidence is context-specific and includes considerations such as study design (e.g., whether there is a control group with randomisation or statistical matching), the practical and statistical significance of results, the extent to which findings are supported by multiple indicators or sources of evidence, and the extent to which different indicators or sources of evidence corroborate. Qualitative evidence can strengthen quantitative findings – e.g., by providing insights into participants' and stakeholders' perceptions of program effectiveness and by exploring potential reasons why projects may be effective for some people/contexts more than others.*
Effectiveness as an influencing program: • Ligada is a recognised brand in the communities in which it is working; • Ligada influences DFID's programs in the wider Mozambique context; • Effective approaches are taken up and implemented by partners; and • Stakeholders become champions/agents of change.	• Brand recognition: % increase in people sampled who recognise the Ligada brand • Number of partners that adopt and/or scale up approaches based on influencing through Ligada engagement (Logframe outcome indicator 1) • Number of DFID policies and programs that have changed as a result of evidence produced by Ligada (cumulative) (Logframe outcome indicator 2) • % of DFID programs referencing evidence from Ligada • Narrative (anecdotal evidence, examples) of stakeholders becoming champions or agents of change.
Cost-effectiveness (outcome efficiency) • The value created by low-risk projects with monetisable outcomes exceeds the resources they consume.	• Breakeven analysis for selected projects *Note: Initially (in 2016) the modelling will be based on budgeted costs and expected beneficiary numbers, rather than actuals. The models will be updated over time as actual results come in. A number of caveats apply to the cost-effectiveness criterion. First, agency, capacity and opportunity are complex multi-dimensional concepts that may not be fully represented by the monetised value of outcomes. Second, the analysis will only be carried out for low-risk projects and not the whole program, because outcomes such as learning and influencing are not readily monetisable. Third, there are no*

Criteria	Indicators & narrative
	existing benchmarks against which to evaluate meaningful cost-effectiveness ratios for the program as a whole. As a fixed quantum of funding has been allocated to Ligada based on an anticipated set of results, VFM overall is a function of how well the allocated resources were used to achieve the expected results, and can be sufficiently evaluated against the effectiveness criteria.

Table 8: MUVA VFM standards from original VFM framework

Excellent VFM	Performance targets generally met or exceeded Excellent performance on all aspects; no weaknesses of any real consequence
Good VFM	Performance targets generally all met Reasonably good performance overall; might have a few slight weaknesses but nothing serious
Adequate VFM	Performance targets generally or nearly met Fair performance, some serious but non-fatal weaknesses on a few aspects
Poor VFM	Performance targets generally not met Clear evidence of unsatisfactory functioning; serious weaknesses across the board on crucial aspects

Evidence sources and methods

The criteria and standards provided a foundation for determining what evidence was needed and would be credible to support evaluative judgements. This sequence of evaluation design helped ensure the selected evaluation methods, and evidence gathered, would be aligned with the program design and the values embedded in the VFM assessment criteria. It was determined that evidence needed to address these criteria would be drawn from: financial and administrative data compiled by the MUVA management team; review of program documentation including intervention proposals, work plans, budgets, quarterly performance monitoring reports prepared by the MUVA team for DFID; outcome evaluation findings presented in MUVA evaluation reports; and narrative evidence gathered through interviews with key members of the MUVA team.

Additionally, it was planned that economic modelling would be conducted to investigate the costs, potential benefits, and break-even point of selected projects with monetisable outcomes. These models would be developed in a spreadsheet, structured according to the rules and conventions of CBA (Drummond et al., 2005). The unit of analysis for each model would be an individual project under the MUVA umbrella. An example is discussed under the first VFM assessment, as follows.

First VFM assessment

MUVA's inception phase commenced in August 2015 and was completed in April 2016. The first VFM assessment covered the eight months from May to December 2016, alongside an evaluation of the inception phase. The VFM assessment contributed to, and was cited in DFID's annual review of the MUVA program.

The first VFM assessment addressed two key evaluation questions: "To what extent has MUVA delivered value for money during the first eight months of its implementation?" and "How can value for money [of MUVA] be improved?".

Given the early stage of the project's implementation, only a limited subset of the VFM indicators were applicable. In particular, it was too early to assess MUVA's VFM performance against the medium to long-term indicators in the VFM framework.

Gathering evidence: The evidence needed to address the criteria was collated and analysed by the author, with assistance from the MUVA team to locate evidence within existing documents. Key sources of evidence were as described above.

Making judgements about performance and VFM: Analysis, synthesis and judgements were undertaken by the author and reviewed by the MUVA program leaders. First, each source of evidence was analysed individually to identify key pieces of information that were necessary to address the VFM criteria. Second, the evidence was synthesised thematically to triangulate evidence relating to each criterion. Third, judgements were made by comparing the evidence of performance to the criteria and standards.

The purpose of this case study is to describe the use of the approach to evaluative reasoning in VFM assessment, and not to present the full results of the MUVA VFM assessment. Nevertheless, it is worth briefly outlining key findings. The assessment found that MUVA was "on track to deliver VFM", with evidence supporting judgements of 'excellent' against economy and efficiency criteria, and 'good' against criteria for effectiveness as a learning program. It was too early to report on performance in terms of equity, effectiveness as an urban female economic empowerment program, effectiveness as an influencing program, or cost-effectiveness.

Reporting: A report was prepared and submitted to DFID in advance of the annual review process (which involved an independent review of the program by DFID staff who had not been involved in the business case nor in managing the program contract). The report summarised the evaluation approach and methods used, the criteria and standards, evaluative judgements, and supporting evidence.

Break-even analysis

Although the first VFM assessment did not provide an evaluative judgement on outcome efficiency, break-even analysis was conducted for one project, to test and illustrate the use of the approach. This project was selected because it was in a sufficiently advanced stage of implementation to provide cost data. The break-even analysis was a prospective analysis, based on budgeted costs and targeted beneficiary numbers. The analysis included ongoing costs of delivery, while one-off implementation costs were excluded. Using published monthly wage data, the potential value of an employment outcome for one beneficiary was estimated, based on the notion of one person securing and sustaining employment at the minimum wage, who would not have done so in the absence of the program.

In order to break even (i.e., for the total value of benefits to equal delivery costs) it was estimated that fifteen percent of program participants would need to secure and sustain employment, over and above those who would have done so without the intervention. It was concluded that a positive return on investment was achievable for the project.

In presenting this conclusion, a number of caveats were made. The estimates were based on budgeted costs and targeted participant numbers, not actuals. The scope of the analysis only included DFID's investment as costs and market wages as benefits. Additional costs may include partner contributions to some projects, and additional benefits include intangible (hard to monetise) benefits such as increases in self-confidence and self-advocacy (at work, in education, at home, and in the community). Moreover, the analysis ignored displacement effects – i.e., it estimated value to the target group rather than the Mozambique population overall. If a member of the target group gaining employment displaces another Mozambican who otherwise would have been employed, net benefit to the Mozambique population would be zero. Finally, it was noted that where a successful intervention is adopted for scale-up by a local partner, there may be potential to reduce ongoing costs per beneficiary over time through economies of scale and productivity improvements.

Second VFM assessment

The second VFM assessment covered a four-month period from January to April 2017. This shortened time period was used because DFID had determined that MUVA should be placed on an annual review cycle with reports to be provided each April (whereas initially a review cycle of 6 months was stipulated).

The second VFM assessment provided a light-touch update on the previous assessment. It tracked performance against the economy and efficiency indicators and provided an update for effectiveness as a learning program. For the first time, the assessment was also able to

include emerging evidence of performance as an influencing program, and for the equity criterion. It was still too early to make conclusive judgements about effectiveness as an urban female economic empowerment program and cost-effectiveness.

It was concluded that MUVA continued to demonstrate good evidence of VFM at the economy and efficiency levels and there were indications that MUVA would be able to achieve strong VFM at the outcome level in the medium and long-term.

Third VFM assessment

In late 2017, following the second VFM assessment, DFID provided feedback on the first two VFM assessments and the author, together with the MUVA team, revised the VFM framework in preparation for the 2018 VFM assessment.

The updated criteria (Table 9) and standards (Table 10) are shown below. Several key changes are notable. First, the second column of the table was changed from "indicators and narrative" to "sub-criteria". This reflected a growing acceptance by DFID and the MUVA team of the role of criteria in evaluative reasoning, supported by indicators and narrative, rather than the use of indicators directly. In reality, the "indicators and narrative" from the first iteration of the framework were more akin to sub-criteria so the re-labelling of the table did not signify a material change in their use.

Second, the VFM framework made a stronger departure from the logframe, reducing the use of logframe indicators as a source of evidence. This was done at DFID's request, to reduce duplication with other reporting and accountability processes, but also provided an opportunity to review the conceptual alignment of the VFM framework with MUVA's theory of change and to strengthen the use of qualitative evidence in the VFM assessment. This change signified, and was enabled by, growing acceptance by DFID and the MUVA team of the validity of narrative forms of evidence to support evaluative judgements.

Third, the efficiency criteria were revised, to better reflect the intended functioning of the program. Three key sub-criteria were defined: allocative efficiency (appropriate allocation of resources to activities); dynamic efficiency (timely implementation, adaptive management and exits from trials); and delivery efficiency (the right quality and quantity of outputs delivered within available resources).

Fourth, the cost-effectiveness criterion was reviewed (though it was still not applied in the third VFM assessment). DFID's definition of cost-effectiveness refers to the general notion of outcome efficiency (e.g., a comparison of value created with value consumed) and not to the specific economic method of cost-effectiveness analysis. It was reasoned that in the long run, MUVA's outcome efficiency should be judged by the value of successfully trialled, sustainable approaches being adopted and scaled up

in urban Mozambique. In this light, the net value of successful approaches taken to scale is a relevant indicator of the outcome efficiency of MUVA. Break-even analysis, it was argued, should therefore model the future costs of scale-up scenarios (as distinct from the pilots conducted by MUVA), using cost and outcome data, and other learning from the pilots, together with assumptions and reasoning about what might be expected to change when the approaches were adopted by local pilots and taken to scale. These future costs and benefits could be explored through scenario analysis, using a discounted cashflow model (King & Wate, 2018).

There was a fifth in-depth revision: the equity criteria. The new framework made a more explicit distinction between equity-related activities at input and output level, and the achievement of equity-related outcomes, as shown in Table 9.

Table 9: MUVA VFM Criteria, Mk 2

Criteria	Sub-criteria
Economy: The MUVA team manages project resources economically, buying inputs at the appropriate quality at the right price.	• Management costs as a percentage of total program costs, compared to target & trend from previous periods • Average flight expenditure compared to agreed targets (national/international) • Average daily subsistence allowance (DSA – for lodging, meals and daily living expenses of staff and consultants when travelling) compared to agreed target • Narrative evidence of good procurement practices (as stated in office manual) leading to savings, value-added services, contributions from other sources, and/or minimisation of cost increases.
Efficiency: The MUVA team produces the intended quality and quantity of outputs, within the available resources – including: • Allocative efficiency – appropriate allocation of resources to activities • Dynamic efficiency – timely implementation, adaptive management and exits • Delivery efficiency – the right quality and quantity of outputs delivered within available resources	Allocative efficiency: • Spend between output areas is in line with intended split over the life of the program • Spend within output areas supports delivery to work plan and strategy Dynamic efficiency: • Timely implementation – work plan delivery as contracted or better • Timely adaptation – MEL feedback loops inform project adaptation • Timely exits – efficient exit of trials (cutting losses) Delivery efficiency: • Meets logframe output targets • Actual costs per beneficiary are in line with expected costs • Stakeholder feedback indicates high levels of satisfaction with quality of outputs
Effectiveness as an urban female economic empowerment program: • MUVA's approaches make their intended contributions to female economic empowerment as defined in	• Systematic assessment of outcomes and scalability using a rubric (Table 11). *Note: These outcomes were defined in project-specific theories of change, which related explicitly to the program theory of change. Evaluations of each project were carried out by the MUVA MEL team and*

Criteria	Sub-criteria
project-level theories of change and are scalable.	*provided the evidence that was used to make evaluative judgements for the VFM assessment.*
Effectiveness as a learning program: • Local participation, relationships and knowledge contribute to project development; • Reflective learning processes are identifying learning that influences adaptive management; and • MUVA provides credible evidence about the effectiveness of every project, including evidence to enable decisions about which projects deliver better results.	• Local participation, relationships and knowledge contributed to the program from the outset and to the development of every project; • Narrative examples demonstrate learning influencing adaptive management; • Systematic assessment of evidence provided by the MEL team for every project that has been running long enough to warrant outcome evaluation (i.e., evidence of the extent to which they are/aren't effective, and why) • MEL team developing innovative new research approaches or methods that deliver credible evidence on FEE approaches.
Effectiveness as an influencing program: • Effective approaches are taken up and implemented by partners; • MUVA influences DFID and non-DFID programs; and • Stakeholders become champions/agents of change for FEE.	• Number of partners that adopt and/or scale up approaches tested through MUVA, compared to target (Logframe outcome indicator 1) • Number of DFID policies and programs that have changed as a result of evidence produced by MUVA, compared to target (Logframe outcome indicator 2) • Narrative on the significance of MUVA's influence • Narrative (anecdotal evidence, examples) of stakeholders becoming champions or agents of change in ways that increase the likelihood of MUVA approaches being adopted and scaled up.
Cost-effectiveness (outcome efficiency) • The value created by projects with monetisable outcomes exceeds the resources they consume.	• Cost-benefit analysis of potential break-even for projects taken to scale.
Equity: MUVA targets resources to the intended groups, including: • Clearly defining the intended target group for each intervention, with supporting rationale (input equity); • Reaching the identified target groups (output equity); and • Achieving the intended changes in the identified target groups (outcome equity).	• Systematic assessment verifies that the documented design of each project defines and justifies its intended target groups. • Participant profile data collected by MEL team demonstrates that projects are meeting their intended target groups (including percentage of enrolled beneficiaries who are women and under the poverty line). • Percentage of enrolled beneficiaries who are women and under the poverty line. • Systematic assessment of evidence provided by the MEL team demonstrates extent to which projects achieve changes for their intended target groups.

Table 10: MUVA VFM standards, Mk 2

Excellent VFM	Performance targets generally met or exceeded Excellent performance on all aspects; there may be room for incremental improvements
Good VFM	Performance targets generally all met Reasonably good performance overall; might have a few areas for improvement
Adequate VFM	Performance targets generally or nearly met Fulfilling minimum requirements; showing acceptable progress overall; though significant improvements may be needed
Poor VFM	Performance targets generally not met Clear evidence of unsatisfactory functioning; immediate and major improvements are needed.

To support judgements of effectiveness as an urban female economic empowerment program, an additional rubric (Table 11) was used to support consistent judgements from findings set out in the evaluation reports of individual MUVA projects. Effectiveness was judged relative to expectations set out in project theories of change and associated documents, while scalability was assessed on the basis of three factors: cost at scale, beneficiary and partner buy-in, and relevance.

Table 11: Rubric for assessing effectiveness and scalability

	Effectiveness	**Cost at scale (overall, and per beneficiary)**	**Beneficiary and partner buy-in**	**Relevance**
Excellent	Evidence to date indicates project is exceeding its intended contributions to FEE as defined in project ToC.	Low overall cost, affordable for partners to scale up; substantially lower cost per beneficiary than usual or past interventions of a similar nature.	Strong buy-in; beneficiaries and partners consider the project/ pathway/ method to be clearly worthwhile to adopt	Project/ pathway/ method is clearly well aligned with community needs and with priorities of DFID and Mozambique governments
Good	Evidence to date indicates project is making its intended contributions to FEE as defined in project ToC.	Moderate overall cost, reasonably affordable for partners to scale up; slightly lower cost per beneficiary than usual or past interventions of a similar nature.	Moderate buy-in; beneficiaries and partners generally consider the project/ pathway/ method to be sufficiently worthwhile to adopt	Project/ pathway/ method is reasonably well aligned with community needs and/or DFID/ Mozambique government priorities
Adequate	Evidence to date indicates project is partially making its intended contributions to FEE as defined in project ToC.	Relatively high overall cost for partners to scale up but not impossible; acceptable cost per beneficiary compared to usual or past interventions of a similar nature.	Just-adequate beneficiary and partner support to be viable	Project/ pathway/ method is tangentially but defensibly aligned with community needs and/or DFID/ Mozambique government priorities
Poor	Evidence to date indicates project is not making its intended contributions to FEE as defined in project ToC.	Unaffordable for partners to take up.	Beneficiaries and/or partners do not consider the project/ pathway/ method to be worthwhile	Project/ pathway/ method is not well aligned with community needs or DFID/ Mozambique government priorities

The processes of gathering evidence, analysing evidence, synthesising evidence, and making judgements, proceeded in the manner already described. The third VFM assessment found that MUVA was delivering good VFM overall as defined by the criteria and standards, based on judgements made for all criteria except cost-effectiveness.

Thematic assessment against the propositions

In this section, the case described above is systematically analysed to investigate the extent to which it corroborates the model's theoretical propositions. Three core propositions are tested in the case studies – namely, the fourth, fifth, and sixth propositions: that CBA can enhance an evaluation of VFM, that CBA is usually insufficient to fully answer an evaluative question about VFM, and that a stronger approach would involve explicit evaluative reasoning, supported by judicious use of economic and other methods –a 'value for investment' (VFI) approach.

The MUVA case study illustrates several ways in which **CBA can enhance an evaluation of the merit, worth and significance of resource use, by yielding insights that would otherwise be difficult to gain**.

Firstly, the CBA used in the MUVA VFM assessment followed the prescribed structure and rules which promoted systematic and rational analysis of costs and consequences. Early application of CBA in MUVA was intended to illustrate the use of break-even analysis, for one of MUVA's projects. The CBA assessed the break-even position (the level of effectiveness at which benefits equal costs) as well as the potential costs and consequences of scaling up the intervention.

However, the CBA did have some limitations in regard to its ability to demonstrate the merits of the method. CBA was only conducted at individual project level and not the whole program. The perspective taken was limited to that of the target group of program participants, and not Mozambique society overall. Nonetheless, the case corroborates the sub-proposition that CBA provides a structure and set of rules that promote systematic and rational analysis – identifying, measuring, valuing, and comparing the costs and consequences of alternative courses of action.

Secondly, MUVA illustrates another advantage of CBA: that valuing and reconciling costs and consequences in commensurable units enables them to be reconciled in the final synthesis, as a single indicator. The MUVA VFM assessment used this feature of CBA to assess the break-even point of a project. The analysis identified the level of effectiveness that would be required to achieve a net present value of zero. Examining both costs and consequences in commensurable units provided insights for program and project managers and evaluators that would have been difficult to intuit. For example, it was estimated that the project would need to produce positive employment outcomes for 15% of participants to break even.

There were some limitations in regard to the comprehensiveness of the CBA, however: it did not include all costs and consequences. Costs were limited to financial costs incurred by the donor agency, and consequences focused on employment outcomes for graduates, valued in terms of modelled wage increases based on Mozambique wage data in relevant sectors. This decision to limit the scope of analysis provided a robust appraisal of these tangible factors without the levels of uncertainty that would have been present if intangible costs and benefits had also been estimated.

The limited study scope is not an inherent limitation of the CBA method but is often seen in practice (Sinden et al., 2009). The CBA could, in principle, have taken a broader perspective by estimating and including monetary values for wider costs and benefits. For example, intangible costs such as unpaid participant time, and intangible benefits such as measured increases in agency and self-efficacy of participants, could have been valued monetarily. Notwithstanding these limitations, the MUVA case corroborates the sub-proposition that valuing and reconciling costs in commensurable units can provide unique insights that enhance an evaluation of VFM.

Thirdly, the MUVA case study provides a demonstration of discounting as a strength of CBA. The MUVA VFM assessment used this feature of CBA to assess the break-even point of one of MUVA's projects. A social discount rate of 8.5% was used, representing the social opportunity cost of capital.

The case also illustrates one of the common challenges in selecting an appropriate discount rate for social investments. The social discount rate is not directly measurable. While some governments specify a discount rate to be used in appraising public investments, Mozambique does not. The 8.5% discount rate was a proxy based on the Mozambique deposit rate, sourced from World Bank data. Still, the case corroborates the sub-proposition that discounting is a strength of CBA that can enhance an evaluation of VFM by explicitly and systematically taking the time value of money into account.

Fourthly, this case study demonstrates the use of sensitivity and scenario analysis, to facilitate transparency and robust thinking about uncertainty and risk. Sensitivity analysis was conducted to understand the extent to which the results of the analysis were sensitive to changes in the input values of certain variables, in particular the value per outcome (i.e., the assumed increase in earnings for a person who benefits from the project, based on published wage data), and the discount rate. The MUVA VFM assessment will, in future, include scenario analysis to assess the break-even position of projects under different sets of conditions – in particular, different scale-up scenarios.

The sensitivity analysis undertaken in this instance was fairly rudimentary, focused on a small set of variables which were manipulated individually. No attempt was made to explore statistical relationships

between variables. Nonetheless, the case demonstrates the value of sensitivity and scenario analysis in CBA.

The fifth sub-proposition was that CBA can accurately measure values, empirically determining the monetary 'weightings' that should be applied to different costs and benefits. The MUVA case neither corroborates, nor falsifies this sub-proposition. The CBA was conducted prospectively. Financial costs of pilots were determined from budget data (not actual costs), with other valuations being informed by published economic data (e.g., average wages for different occupations, and discount rate). This is typical of many CBAs; valuations were determined by assumptions, informed by literature, with scenario and sensitivity analysis to understand the nature and extent of uncertainty.

The MUVA case study corroborates the proposition that, when it comes to evaluating VFM in social policies or programs, **CBA is usually insufficient to fully answer an evaluative question about the merit, worth and significance of resource use.** Eight sub-propositions were identified, seven of which are corroborated here.

The first sub-proposition was that CBA may not capture all criteria: CBA provides an estimate of efficiency (Adler & Posner, 2006) whereas the merit, worth and significance of resource use may be multi-criterial. The VFM evaluation included assessment of equity and the effectiveness of MUVA as a learning and an influencing program, in addition to economy and efficiency. By design, and in line with typical application, the CBA used in the MUVA assessment estimated the economic efficiency of a project and did not attempt to examine other criteria such as equity, relevance or sustainability.

Furthermore, the CBA provided an imperfect estimate of economic efficiency, because of the limited scope of costs and benefits included, and because the perspective taken was that of the beneficiary group and not Mozambique society (therefore displacement effects, where a member of the target group gains employment by displacing another Mozambican citizen, were ignored). Moreover, the exclusion of intangible outcomes (e.g., acquisition of 'soft skills' that enhanced agency at work and at home) may under-value projects where such benefits are strongly evident and highly valued by beneficiaries and employers. A more comprehensive CBA would not have been feasible with the time, resources and data available. For all of these reasons, the case corroborates the sub-proposition that CBA may not capture all criteria in an evaluation of VFM.

The second sub-proposition was that CBA, by virtue of the Kaldor-Hicks criteria, reflects a particular normative position on equity that subordinates distributive goals to overall efficiency. The CBA in the MUVA assessment examined aggregate value for all beneficiaries and not the distribution of costs and benefits (e.g., differences between the most and least disadvantaged beneficiaries). Therefore the CBA implicitly reflected the Kaldor-Hicks criterion that a net gain in overall value would be

worthwhile for the project regardless of the impact on individual beneficiaries.

This is not the only analytical option with CBA, but it is the default one. In principle it is possible to conduct CBA from multiple perspectives (for example, to compare and contrast net value from the perspective of the donor and beneficiaries). Moreover, DFID's own guidance recognises that reaching the most disadvantaged may involve additional costs. MUVA projects were working with disadvantaged groups. Working effectively with these target groups may be more intensive and more costly than working with middle income groups. If so, this would set a higher bar for these projects to break even. In a context like this, explicit consideration of equity is critical. This consideration does not preclude the use of CBA but gives weight to the argument that CBA should not be the only method used.

The third sub-proposition was that CBA reflects a normative position that all goods are fungible and should be valued in commensurable units. The CBA valued costs and benefits monetarily in order to derive a net present value. While this is a strength of CBA as already noted, it would at the same time have been a limitation if it were the only method used in the VFM assessment. For example, each of MUVA's three objectives (testing approaches, learning, and influencing) needed to be considered individually before making an overall determination of VFM – and would have been challenging to value monetarily. Overall, the case corroborates the sub-proposition that CBA treats all values as fungible.

The fourth sub-proposition was that commensuration may obscure diversity in people's values. Aggregation of values using a common metric may obscure qualitative differences in the perspectives of different groups (e.g., power disparities, diverging values or interests). Addressing power imbalances and diverging worldviews are at the heart of the MUVA program's long-term intent of improving economic empowerment of young women in urban Mozambique. Achieving such change requires sustained, multi-level and multi-faceted intervention aimed at changing social and institutional norms as well as the individual agency of women in the target groups, and attitudes of men in their communities. The CBA, in contrast, took a reductionist perspective that valued outcomes in monetary terms, from the perspectives of beneficiaries (young women) in aggregate. Perspectives of other groups were not included, nor did the analysis consider within-group differences.

The analysis could have been broadened to consider costs and benefits to other groups. For example, some men might have viewed increased agency for women as a dis-benefit. The utilitarian model of CBA aggregates the self-interested preferences of individuals (Adler & Posner, 2006). In a male-dominated society, including such perspectives would diminish and possibly negate the value of the projects. If this analysis had been undertaken it could have provided a more powerful demonstration of the proposition that aggregating values obscures differences in perspectives, and that doing so can perpetuate existing disparities in

power and privilege. Overall, the case corroborates the sub-proposition that commensuration obscures diversity in people's values.

The fifth sub-proposition was that commensuration may obscure qualitative differences between things of value, such as economic, social, cultural and environmental value. The CBA focused on financial costs and benefits. The market wage data used in the analysis was a limited proxy that did not represent the full value of outcomes. For example, the wage might obscure differences in employment quality (e.g., whether the work is stable and satisfying, or insecure and menial).

Moreover, increased income, and increased agency (e.g., self-efficacy and ability to advocate for oneself) are qualitatively very different. This means, for example, that two projects could be of equivalent net value but it would not be evident in the net present value that one project mainly conferred financial benefits while the other brought intangible gains in agency for participants.

The analysis could, in principle, have been broadened to consider further categories of intangible costs and benefits. For example, community leaders invested and risked their reputations in facilitating MUVA teams' access to communities. Indirect benefits to the families and communities of direct participants could also have been explored. This would have provided a more powerful demonstration of the obscuring effect of aggregating values using a common metric. In summary, the case corroborates the sub-proposition that valuing benefits and costs in commensurable units can obscure qualitative differences in what is valuable.

The sixth sub-proposition was that CBA reflects a consequentialist perspective and has limited ability to evaluate the merit, worth or significance of resource use during implementation, in terms of processes or before outcomes can be measured. This CBA examined only the costs and consequences of a MUVA pilot. As it was too soon to measure costs and consequences of the pilot, break-even analysis was conducted prospectively, to appraise potential value. Project processes were not taken into account beyond the use of basic statistics like throughputs to estimate future costs and outcomes. The case corroborates the sub-proposition that CBA focuses on costs and consequences and does not evaluate process value.

The seventh sub-proposition was that the scope of a CBA may be constrained by what is measurable. In practice, important values may be excluded from a CBA because they are too hard to estimate. The scope of the CBA was deliberately limited to financial costs and consequences. Valuation of intangible benefits was outside the reach of available evaluation resources. Intangible outcomes were considered in the wider VFM evaluation but were not monetised. If CBA had been the only method used in this evaluation, its scope could, in principle, have been widened to more comprehensively encompass intangible benefits. Nonetheless it would have remained difficult to value benefits such as self-efficacy, self-

confidence and agency. The case corroborates the sub-proposition that in a social investment, key intangible benefits may be too hard to monetise.

The eighth sub-proposition was that CBA is not explicitly required to fully adhere to program evaluation standards. In this instance, CBA was one component of a wider evaluation and was conducted on a desktop basis, using available cost and outcome data, some engagement with project managers to develop realistic assumptions, but no engagement with beneficiaries. The CBA was designed and conducted without reference to evaluation standards. If the evaluators had consulted evaluation standards when designing and conducting the CBA, the principles of contextual viability and negotiated purposes may, conceivably, have resulted in a decision not to conduct CBA if, for example, it was found that monetary valuation of women's economic empowerment was unacceptable to stakeholders. However, the evidence from this case study does not provide a test of the sub-proposition.

Given the strengths and limitations of CBA in the aforementioned propositions, the model proposes **that a stronger approach would involve explicit evaluative reasoning, supported by judicious use of economic and other methods**. There are ten sub-propositions, nine of which are corroborated in this case study.

First, the case corroborates the sub-proposition that the approach can accommodate multiple and diverse criteria. This evaluation included explicit criteria and standards for economy, efficiency (technical, allocative, and dynamic), equity, effectiveness (as a female economic empowerment program, as a learning program, and as an influencing program) and for cost-effectiveness (outcome efficiency). In particular, the inclusion of equity underscores the trade-offs involved in development investments and that VFM of development aid depends not only on efficiency, but on reaching women and girls, young people, poor people and other marginalised groups. There is a tension between efficiency and equity; working to improve the lives of the most disadvantaged may be more costly than improving the lives of moderately disadvantaged people. Criterial evaluation of efficiency and equity makes the trade-offs explicit.

Second, the case corroborates the sub-proposition that the approach allows equity to be considered separately from efficiency. It can take the Kaldor-Hicks efficiency criterion (with its inbuilt position on equity) into account without being limited to this position exclusively. Equity criteria considered the performance of MUVA in targeting, reaching, and achieving outcomes for specific target groups. At the same time, CBA was undertaken at project level to consider the aggregate break-even point of an intervention.

Third, the case corroborates the sub-proposition that the approach does not require commensuration; it has the ability to contrast and support deliberation on qualitative differences between groups of people or between things of value. This evaluation clearly demonstrates this point. Value was defined at multiple levels of the theory of change, for multiple

criteria, using a mix of quantitative and qualitative standards and evidence. This VFM assessment did not include deliberation on qualitative differences in the values held by different groups of people. Nonetheless, the case provides sufficient evidence to corroborate the sub-proposition.

Fourth, the case corroborates the sub-proposition that the approach does not take an exclusively consequentialist perspective; a practical upshot of this is that it can be used to evaluate the merit, worth and significance of resource use during implementation (e.g., for diagnostic or formative purposes). Using DFID's VFM criteria of economy, efficiency and equity, early VFM assessment focused on measures of VFM performance during implementation, namely: good management of resources (funding) to procure inputs (e.g., staff, consultants, equipment, and transport); good management of inputs to produce outputs (including sub-criteria for allocative, technical and dynamic efficiency); and 'design equity' (stakeholder engagement and needs assessment to inform intervention design and identification of target groups). Later, as outcome evidence became available, measures of effectiveness and return on investment were added. In this way, a consequentialist perspective was one perspective included in the evaluation, but process value was also examined.

Fifth, the case corroborates the sub-proposition that the approach can incorporate a broader range of analytical options about value than CBA. This evaluation took an analytic approach, separately analysing and evaluating economy, efficiency, effectiveness, cost-effectiveness, and equity, and then using these judgements collectively to make a determination of VFM overall. Within each criterion, judgements were made holistically against multiple criteria using a rating rubric. A mix of qualitative and quantitative synthesis methods were used (with quantitative synthesis being an inherent part of CBA modelling, and qualitative synthesis used as the overarching approach to integrating multiple streams of evidence). As a result, it was possible to work with diverse values in ways that would have been challenging in CBA – for example, to describe how and why soft skills were important to employers.

Sixth, the case corroborates the sub-proposition that the approach does not prescribe the methods to be used to gather evidence; rather it is flexible to enable matching of methods to context, recognising that no single method can address all VFM criteria and that no single method should be regarded as gold standard. In this VFM assessment, decisions about what forms of credible evidence would be necessary and sufficient to support judgements about performance and VFM were made after criteria and standards had been agreed. This sequence of evaluation design ensured the methods of data collection addressed the aspects of performance and VFM that were agreed to be important among stakeholders.

Seventh, the case corroborates the sub-proposition that the approach has the ability to accommodate mixed methods evidence, including the results

of quantitative and qualitative analysis. This VFM assessment was a mixed methods evaluation, triangulating multiple streams of evidence to make judgements against the rubrics. Methods included analysis of financial and administrative data, findings from rigorous outcome evaluations of projects, qualitative stakeholder feedback, and CBA.

Eighth, the case corroborates the sub-proposition that the approach is able to incorporate economic evidence, including the results of economic analysis. The VFM assessment included, but was not limited to, economic analysis of project costs and consequences. The results of the CBA supported determination at the 'cost-effectiveness' (outcome efficiency) level of the VFM assessment, alongside other evidence and reasoning used to address the other four criteria. As already acknowledged, the CBA had a limited scope and estimated efficiency from a narrow perspective. To date, CBA has only been undertaken for one MUVA project but this is sufficient to corroborate the sub-proposition.

Ninth, the case corroborates the sub-proposition that evaluative reasoning using qualitative weighting and synthesis can be combined with CBA to strengthen VFM assessment. This case demonstrated that the two approaches are not mutually exclusive. The overarching rubric-based evaluation incorporated CBA as a method of analysis. This enabled the evaluation to gain from the strengths of CBA (as validated in the first set of propositions above) while mitigating weaknesses (as outlined in the second set of propositions).

The tenth sub-proposition was that the approach can be conducted in full adherence with program evaluation standards. This could not be tested from documentary evidence as program evaluation standards were not explicitly referenced in this case study. Retrospective assessment against the standards suggested a general level of adherence, with some areas where the standards could have been followed more comprehensively. For example, beneficiaries and delivery partners were involved in program design but were not involved directly in the development of rubrics. No standards were identified that the assessment would have been categorically unable to adhere to.

The findings above are summarised later in this chapter, in Table 14, which also examines the replication of findings across the two cases. First, however, the second case study is presented: the Pakistan Sub-National Governance Program.

Case study 2: The Pakistan Sub-National Governance Program

This case study presents the Pakistan Sub-National Governance (SNG) Program, a DFID-funded and OPM-delivered program supporting reforms in public financial management, planning and innovative service improvement pilots in two provinces of Pakistan – Punjab and Khyber Pakhtunkhwa (KP). The program operated from 2013 to March 2018 in 12 districts (six in each province). The VFM framework was designed during August-September 2016 and approved by both OPM and DFID prior to VFM assessments proceeding. VFM assessments were carried out in February-March 2017 and February-March 2018.

Program and context

The business case for the SNG program (DFID, 2012) noted that Pakistan had long faced significant social, economic and security challenges. These challenges included increasing poverty and socioeconomic inequalities. Provision of basic services fell short of international norms and many groups, especially women, were excluded from access to services. As a result, over one-third of Pakistanis remained poor and uneducated, with some vulnerable to involvement with violent groups.

Challenges in governance were closely related to Pakistani provinces' economic and social problems. Dominance of military, bureaucratic and political elites had distorted governance and resource allocation. Weak social indicators reflected public sector mismanagement of resources. Weak capability and accountability had been compounded by corruption. These problems undermined public confidence in the state and hampered progress toward Sustainable Development Goals (DFID, 2012).

Further challenges had compounded these problems, including rapid population growth which outstripped the existing capacities of basic services such as health, education and water, sanitation and hygiene. There had been significant structural changes in the assignments of authority and fiscal responsibility between different levels of government – including abolition of locally elected government bodies in 2010, devolution from federal level in 2011, and decentralisation of responsibility for the management of education and health services from provincial to district level (implemented in KP in 2015 and Punjab in 2017).

In this context, the SNG program sought to work with provincial and district governments, to improve their capability and address governance problems that hampered service delivery, such as weak planning capacity, a lack of evidence on which to base policies and budgets, low levels of staff capability, corruption, and inadequate accounting to citizens and legislatures.

The intended outcome of the SNG Program was 'Poor people in Punjab and Khyber Pakhtunkhwa report that services are better meeting their needs'. The program supported reforms in public financial management, governance and planning, and operated a challenge fund to finance innovative service improvement pilots. Collectively, these workstreams aimed to deliver results cross three main areas: 'Decisions by government based on evidence'; 'Public services that are more responsive to people's needs' and 'Strengthened government capability to deliver basic services'.

The evidence base on how to effectively improve district governance was limited, especially in conflict-affected states like Pakistan. Innovation was therefore integral to the program's design. The SNG program sought to develop and pilot new initiatives and approaches planning, budgeting and financial management, to scale up some promising initiatives, and to build an evidence base for this program and for other, similar endeavours in the future.

The reform processes in the SNG program were iterative and evolved with the changing context (e.g., the aforementioned devolution of responsibilities to district governments). The interventions and activities of the program were tailored to the specific policy context, needs and challenges of each province.

In the first three years of implementation, prior to the adoption of an evaluation-specific approach to VFM assessment, VFM reporting had been limited to selected indicators of economy and efficiency. When the SNG program reached its fourth year, the program team and DFID required VFM assessments to start covering the full results chain including effectiveness in achieving early outcomes and assessing progress toward longer-term outcomes including equity.

A VFM framework was needed that could take into account the complexity of the program and its evolving context, to assess the program's achievements relative to what could reasonably be expected in this environment, and to take into account the adaptive and learning-focused dimensions of the program.

Developing the VFM framework

The VFM framework was designed and developed with the SNG teams of Punjab and Khyber Pakhtunkhwa, in a one-week workshop in Lahore in August 2016. Following the VFI process model, the VFM design followed the stages of developing an understanding of the program through its theory of change, developing criteria and standards, and identifying sources of evidence needed to address the criteria and standards. Following evaluation design, VFM assessment was carried out by gathering the necessary evidence, analysing each stream of evidence, and synthesising findings to make judgements about performance and VFM against the agreed criteria and standards. These steps are detailed in the following paragraphs.

Understand the program

The evaluation design process commenced with a review of core program documentation, including its business case, inception report, previous VFM reports, and briefings from program leaders. Terms of reference for the VFM framework were developed, with clear objectives and information needs stemming from previous DFID annual reviews.

The theory of change for the SNG program was used as a key point of reference for developing criteria and standards aligned with the intended design, delivery and outcomes of the program. The SNG program had an existing theory of change. However, it was found to be insufficiently detailed to support the development of a VFM framework. In particular, the intended outcome had been defined at a high level ('Services better meet the needs of poor people in Punjab and Khyber Pakhtunkhwa'). This left a relatively long conceptual journey from outputs to the intended outcome.

To address this gap in the theory of change, the author and colleagues reviewed the theory of change, in consultation with the SNG teams of the two provinces – including their monitoring and evaluation (M&E) advisors, technical advisors, and management – to define a set of intermediate outcomes between outputs and final outcomes. These intermediate outcomes provided the basis for identifying sub-criteria and standards for evaluating performance at the 'effectiveness', 'cost-effectiveness' (outcome efficiency), and 'equity' levels of the VFM framework.

The theory of change (Figure 6) connects the four SNG work streams of public financial management, governance, geographic information systems, and challenge fund (at the bottom of the diagram), to the program's intended impact and outcomes (at the top of the diagram). The middle sections of the theory of change were added by the author and colleagues. Two sets of intermediate outcomes were distinguished: changes occurring within governments (at provincial and district levels), and resultant changes that would affect the lives of citizens in Punjab and KP.

In reviewing the SNG program's logframe, it was observed that many of the changes labelled 'outputs' were in fact changes occurring within governments, and therefore more accurately characterised as intermediate outcomes. A distinction was made between outputs delivered directly by the SNG teams (for example, draft legislation, rules and guidelines, and training provided to governmental staff), and the first wave of outcomes which involved some action on the part of other actors in the system (for example, improved public financial management and planning systems). The practical upshot was that the intermediate outcomes at the level of government systems were to a large extent captured by the SNG program's existing logframe indicators. There were a few notable exceptions to this, however, and these additional outcomes not represented within the logframe were specified (in red text) in the theory of change. For example, these additional outcomes captured:

SNG's important role in facilitating coordination between donors and government; the program's work around strengthening citizen access to information and channels to participate; and output-based budgeting. Additionally, cross-cutting issues of sustainability, gender and equity were identified.

Of particular note, one key population-level intermediate outcome was identified, connecting changes within governments to the final outcome and impact indicators. This outcome, labelled 'increased funding and/or efficiency improvements for service delivery, targeted to needs' became a critical component of the VFM framework at cost-effectiveness (outcome efficiency) level, as explained later. It was reasoned that the changes within the provincial and district governments should result in additional resources for service delivery, targeted to needs, ultimately resulting in increased resources being utilised by districts and their 'service delivery units' (e.g., health and education services), thus impacting on citizens.

Additionally, it was noted that the value of the SNG program related not only to achieving specified outcomes, but critically to the program's capacity to generate learning that could influence other relevant programs, and its capacity to be adaptive, responding to lessons learnt as well as emergent opportunities and challenges. These features were also included in the theory of change and, subsequently, in the VFM framework.

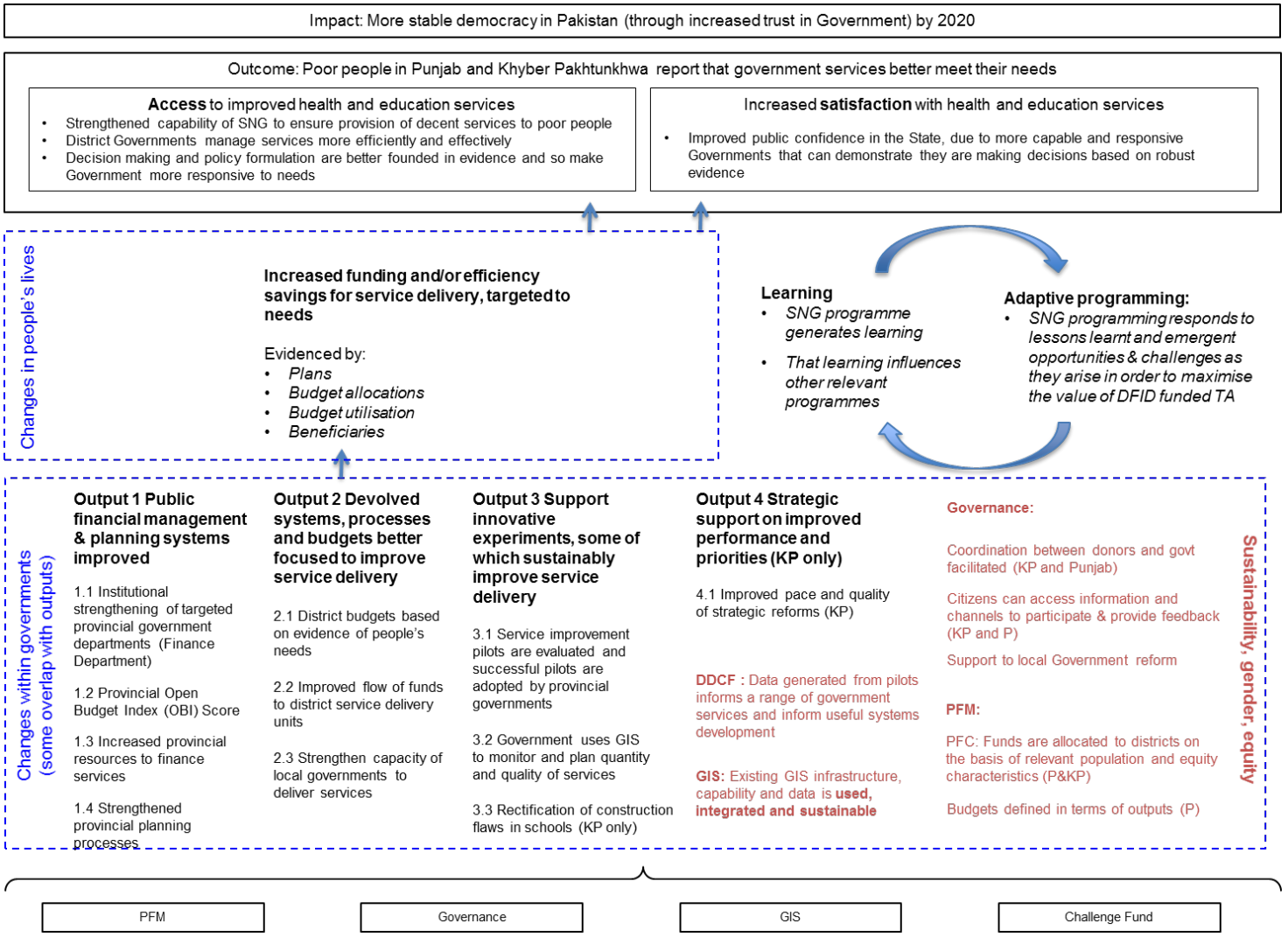

Figure 6: Theory of change for Pakistan sub-national governance program. Reprinted from King & Allan (2016). Reprinted with permission.

Criteria and standards

Drawing on the theory of change, the author and colleagues facilitated a series of workshops with SNG M&E advisors, technical advisors and management, to develop program-specific definitions of economy, efficiency, effectiveness, cost-effectiveness (outcome efficiency) and equity, together with more detailed sub-criteria. This intensive, participatory process was important because it helped to ensure that the VFM framework was based on a sound understanding of the context and real-world functioning of the program. The process of developing criteria and standards surfaced a range of values and perspectives from participants, and provided a forum to reconcile and integrate these with the program performance requirements set out in documentation, in particular the program's logframe.

Criteria and standards were summarised in rubrics. There are multiple ways of structuring rubrics, and research into choosing between types of rubric is in its infancy (Dickinson & Adams, 2017; Martens, 2018b; Wehipeihana et al., 2018). For the purposes of this evaluation, a separate rubric was developed for each of DFID's five criteria. This type of approach has been described as analytic (Martens, 2018b), because it facilitates judgement-making against each criterion individually before synthesising these judgements into an overall evaluative judgement about VFM. The maxim was followed that rubrics should include "rich descriptive language that differentiates quality and/or success and is user friendly to stakeholders" (Dickinson & Adams, 2017, p. 114).

Table 12 summarises DFID's generic question for each criterion, alongside the SNG-specific criteria and sub-criteria that were developed. Performance standards were also developed (Table 13), using a visual 'traffic light' system to define five levels of performance. The criteria and standards were reviewed and approved by DFID prior to the use of the VFM framework to assess program VFM.

A consideration that influenced the design of the performance standards was the decision to incorporate logframe indicators as sub-criteria at the effectiveness level of the VFM framework. This decision was taken primarily on the basis that the indicators (when triangulated with additional evidence) were relevant and valid indicators of program performance. In part this validity reflected the processes that had been used in selecting and agreeing the logframe indicators, in which the SNG teams had been centrally involved. Secondarily, from a pragmatic standpoint, it was agreed that conceptual alignment of the VFM assessment and logframe would make for a more coherent and efficient VFM assessment, making use of existing evidence already being collected by the program's monitoring and evaluation teams.

Flowing from this decision to incorporate logframe indicators, the performance standards were also defined in a way that corresponded to the five levels of the DFID scoring system for Annual Reviews. This concordance between the VFM standards and DFID's scoring system is

shown at the economy and effectiveness levels of Table 13, where the definitions of the levels (from "substantially exceeded expectation" down to "substantially did not meet expectation") are drawn directly from that scoring system.

Table 12: VFM criteria used in the SNG Program

VFM criteria	DFID's VFM questions	SNG-specific definition	Sub-criteria (overview)
Economy	"Are we or our agents buying inputs of the appropriate quality at the right price? (Inputs are things such as staff, consultants, raw materials and capital that are used to produce outputs)" (DFID, 2011, p.4)	The SNG team manages program resources economically, buying inputs of the appropriate quality at the right price.	Average unit costs of consultants (national/ international) compared to contract; trends in average costs of significant items compared to contract; supplier negotiation and contract management ensuring inputs remain cost-competitive; one-off cost savings secured through negotiation.
Efficiency	"How well do we or our agents convert inputs into outputs? (outputs are results delivered by us or our agents to an external party. We or our agents exercise strong control over the quality and quantity of outputs)" (DFID, 2011, p.4)	The SNG program produces the intended quality and quantity of deliverables, within the available resources.	Comparison of actual delivery against work plans. "Deliverables" refers to completion of planned activities and delivery of 'products' such as draft legislation, rules and guidelines, and training. Recognition of program activities that varied from the work plan to pursue new opportunities as they arose and drop activities without traction. In this fixed budget environment, and bearing in mind the bespoke, heterogeneous nature of deliverables and consequent lack of benchmarks, cost-output ratios were not used.
Effectiveness	"How well are the outputs from an intervention achieving the desired outcome on poverty reduction? (Note that in contrast to outputs, we or our agents do not exercise direct control over outcomes)" (DFID, 2011, p.4)	The SNG program achieves its intended changes in public financial management, governance and planning systems, and service improvement pilots, in Punjab and Khyber Pakhtunkhwa.	"Intended changes" were further specified in sub-criteria corresponding to each outcome area specified in the logframe (e.g., "district budgets based on evidence of people's needs"). These were individualised to each province, given the different issues and strategies at play.
Cost-effectiveness (outcome efficiency)	"How much impact on poverty reduction does an intervention achieve relative to the inputs that we or our agents invest in it?" (DFID, 2011, p.4)	The SNG program contributes to increased funding and/or efficiency improvements for services to meet identified needs in Punjab and Khyber Pakhtunkhwa.	Ratio of fiscal value created (increased funding allocations and utilisation to education and health services, plus efficiency gains) to fiscal value consumed (resources invested in the SNG program).
Equity	"When we make judgements on the effectiveness of an intervention, we need to consider issues of equity. This includes making sure our development results are targeted at the poorest and include sufficient targeting of women and girls" (DFID, 2011, p.3).	Changes in needs-based planning and resource allocation contribute to reducing inequities by targeting resources to poor people, women and girls.	Proportion of increased funding allocated to, and utilised by services for poor people, women and girls.

Table 13: VFM Standards used in the SNG Program

					
Economy	Substantially exceeded expectation	Moderately exceeded expectation	Met expectation	Moderately did not meet expectation	Substantially did not meet expectation
Efficiency	SNG deliverables for the year substantially exceeded work plan and in line with allocated budget	SNG deliverables for the year moderately exceeded work plan and in line with allocated budget	SNG deliverables for the year completed according to work plan and in line with allocated budget	SNG deliverables for the year moderately did not meet work plan and/or moderately exceeded budget	SNG deliverables for the year substantially did not meet work plan and/or substantially exceeded budget
Effectiveness	Substantially exceeded expectation	Moderately exceeded expectation	Met expectation	Moderately did not meet expectation	Substantially did not meet expectation
Cost-effectiveness (outcome efficiency)	Increased funding utilisation + efficiency gains exceed combined DFID resourcing for SNG and consequential provincial and district government investments in governance, planning and public financial management reform	Increased funding utilisation + efficiency gains exceed DFID resourcing for SNG program	Increased funding allocation + efficiency gains exceed DFID resourcing for SNG program	Funding allocation for services moderately below DFID resourcing for SNG program	Funding allocation for services substantially below DFID resourcing for SNG program
Equity	A substantial proportion of increased funding is utilised by services for poor people, women and girls.	A moderate proportion of increased funding is utilised by services for people, women and girls.	Needs-based planning and resource allocation includes explicit targeting of services for poor people, women and girls.	Needs-based planning and resource allocation includes implicit targeting of services for poor people, women and girls.	Needs-based planning and resource allocation does not target services for poor people, women and girls.

Evidence sources and methods

After the criteria and standards were agreed, the author and colleagues worked with SNG staff to determine what evidence was necessary to support evaluative judgements. Key streams of evidence included: program cost data, disaggregated across work streams; quarterly contract monitoring reports summarising planned against executed delivery; logframe ratings, produced each year as part of DFID's annual review process; together with output briefs, case examples, and economic analysis detailed as follows.

Output briefs provided narrative evidence, prepared specifically for the VFM evaluation, to support accurate and well-evidenced ratings. An output brief was written for each logframe 'output' (which, as noted earlier, corresponded to the intermediate outcomes in the revised theory of change). Each output brief detailed the problem to be addressed, the interventions, institutional context, the most significant changes in government, and an assessment of the sustainability of those changes. In doing so, the output briefs collectively told the story of how changes in government systems contributed to increased service delivery which was responsive to needs (or could reasonably be expected to, if too early to actually observe such changes). Additionally, a small number of cross-cutting case examples were prepared, illustrating how the SNG program's interventions interacted to contribute to improvements in access to needed services.

Economic analysis

Economic analysis was designed to address DFID's cost-effectiveness criterion (which refers to the general notion of outcome efficiency and not the economic method of cost-effectiveness analysis). The VFM framework argued that the outcome efficiency of the program was ultimately tied to improvements in democratic governance, resulting from better access and satisfaction, which should be achieved in the two provinces over the longer term. However, it was acknowledged that these improvements would result from multiple factors, including the efforts of provincial governments themselves, as well as other international development programs with overlapping objectives. Moreover, it was beyond the reach of the VFM framework to assess outcome efficiency according to these long-term outcomes (which were to be assessed by an independent impact evaluation after the program had ceased operating). Nonetheless, the VFM framework proposed that it would be possible to track the program's direction of travel by measuring intermediary outcomes against the inputs of the program.

Based on the population-level intermediate outcome identified in the theory of change, the following outcome efficiency criterion was defined for the SNG program: 'The SNG program contributes to increased funding and/or efficiency improvements for services to meet identified needs in Punjab and Khyber Pakhtunkhwa'. The standards developed for this criterion (shown in Table 13 above) reflected the concept that the SNG

program should create more value than it consumes – i.e., provide a positive return on investment. It was reasoned that a measurable and relevant near-term proxy for 'value created' would be the increased flow of funds to districts, together with possible efficiency improvements (e.g., savings to governments stemming from improved debt management and other initiatives). According to the theory of change, the increased funding and/or efficiency improvements should result from the intended governmental changes (e.g., increased provincial resources available to fund services; needs-based planning and budgeting). The increased utilisation of those funds by health, education and other services would be necessary (though not sufficient) to achieve longer-term benefits for citizens. It was acknowledged that these long-term outcomes would also depend on those services being of adequate quality.

'Value consumed' was represented by the collective resources invested in governance, planning and public financial management reforms by DFID and by the provincial governments of Punjab and KP. To meet the standard for the highest level of outcome efficiency, program stakeholders agreed that the indicator for 'value created' should exceed the total resources invested by DFID and the provincial governments combined. However, to meet the minimum standard for an acceptable level of outcome efficiency, it was agreed that the cost side of the equation should take the narrower perspective of DFID as donor.

It was further argued that increased funding *allocation* to districts exceeding DFID's resourcing would represent a reasonable minimum benchmark for outcome efficiency in that SNG's influence over funding allocations was greater than its increases in funding *utilisation* by health and education services in those districts. However, increased funding utilisation was agreed to be ultimately more important than increases in allocations, because it is a necessary precursor to any real impact on access to services for citizens. Therefore, increases in utilisation were used at the top two levels of the outcome efficiency standards.

The time horizon for the assessment against the outcome efficiency standards was cumulative across the entire period of SNG implementation. Furthermore, if the SNG program was successful, the increased flow of funds to districts should be sustained beyond the period of the program. Therefore, determining the value of increased allocations and utilisation should incorporate not only retrospective assessment of results to date, but also should forecast future funding flows based on a transparent and reasonable set of assumptions, supported by scenario and sensitivity analysis.

A time horizon of ten years was agreed to be appropriate for forecasting purposes, in keeping with the expected minimum life of the new public financial management legislation and guidance. A base discount rate of 8% was used, in line with the discount rate understood to be commonly in use in CBA of public projects in Pakistan. An approach was devised for estimating what share of increased value should be attributed to SNG, as detailed later.

Learning and adaptive programming

In addition to the qualitative, quantitative, and economic streams of evidence described above, one further stream of evidence was needed: evidence of learning and adaptive programming. The SNG program operated in a context that necessitated sensitivity to the political economy of public finance and governance reform. This required an approach that went beyond technical solutions and included attention to stakeholders, institutions and processes. This context demanded that the SNG program be flexible, adaptive, and take calculated risks (not all of which would pay off). The SNG program was also expected to generate learning that could inform similar reforms in the future.

Accordingly, the VFM assessment acknowledged the importance of learning and adaptive programming as defined in the theory of change: 'Learning: The SNG program generates learning, which influences other relevant programs'; and 'Adaptive programming: SNG programming response to lessons learnt and emergent opportunities and challenges as they arise, in order to maximise the value of DFID-funded technical assistance'.

These two factors were included within the VFM framework. Rather than assess the effectiveness of these features against predetermined targets, however, it was considered more important to describe the learning and adaptation that had occurred. Accordingly it was agreed that the VFM report would include narrative documenting the cumulative learning and adaptation.

First VFM assessment

The first VFM assessment using this framework was carried out during February-March 2017. Evidence was gathered by the SNG teams' monitoring and evaluation advisors, and by program management staff at OPM's head office in Oxford, England.

Analysis, synthesis and judgements were facilitated by the author and colleagues, in a series of workshops with SNG staff during a week-long visit to program's Islamabad offices. Analysis involved examining each stream of evidence individually. Synthesis involved combining the multiple streams of evidence to reach a holistic understanding of SNG performance – for example, through triangulation (Davidson, 2005). Based on the synthesised evidence, judgements were made to determine an overall performance rating for each criterion of VFM (economy, efficiency, effectiveness, cost-effectiveness, and equity) using the standards developed for the SNG program.

Subsequently, an overall judgement of VFM was made, in a workshop with the SNG teams. This included deliberation on the relative importance of the five criteria and thus how much weight each should receive in the overall judgement. As this VFM assessment was being conducted in the fourth year of the program, it was agreed that cost-effectiveness

(outcome efficiency) and equity were the most important criteria to contribute to the overall judgement of VFM.

The approach to understanding the contribution of the SNG program to the observed intermediate outcomes evolved between the first and second VFM assessments. In the first VFM assessment, a judgement-based approach was adopted. It was reasoned that the results of the SNG program should be conceptualised as the synergistic effect of interventions (e.g., ideas, training, resources, documents) produced by SNG teams, and the adoption and spread of such interventions by the provincial governments. As other DFID programs may also contribute to these outcomes, the SNG program may be able to claim a contribution to the outcomes. The challenge lies in determining the extent of such contribution amongst other influencing factors.

It was argued that a notional split between SNG and other relevant programs would not be meaningful, as it would fail to recognise the synergies and interdependences between the various influences. Nevertheless, such a percentage estimate was understood to be one of DFID's expectations of the VFM assessment. A framework of four considerations was used to provide a systematic way of assessing the SNG program's contribution to outcomes. These considerations were labelled 'deadweight', 'displacement', 'attribution', and 'drop-off' (Nicholls et al., 2012).

'Deadweight' referred to changes that would have occurred without any intervention – for example, due to environmental factors or an internal motivation to change on the part of provincial governments. For example, tax revenues may increase not only because of improvements in public financial management, but also due to economic growth. 'Displacement' prompted the SNG teams to consider whether changes due to SNG had displaced any other positive effects or had unintended consequences. 'Attribution' referred to the notion that a portion of the observed outcomes might be attributable to another program external to SNG. 'Drop-off' prompted consideration of whether the results of an intervention may diminish over time.

The contribution of SNG to each intermediate outcome was considered individually, and included at the end of each output brief, with percentage estimates for each factor being deliberated upon in a workshop with SNG M&E advisors, technical experts and management from each province. Consensus was reached that SNG could conservatively claim ten percent of the total changes in these provincial-level fiscal indicators.

The findings from the outcome efficiency results in the first VFM assessments indicated that the value of additional funding for health and education services well exceeded the resources invested in the program. In this context it was recognised that the credibility of the assessment could be strengthened by taking a more robust approach to contribution analysis. This was added in the second VFM assessment, as described below.

Second VFM assessment

The second (and final) VFM assessment was carried out in the final months of the program, during February-March 2018. The approach largely followed that of 2017, as described above. However, in response to feedback from DFID, a new approach was taken to strengthen analysis of the SNG program's contribution to outcomes.

Contribution tracing is a relatively new approach to evaluating causal claims (Befani & Mayne, 2014). It combines elements of contribution analysis (Mayne, 2008) and process tracing (Van Evera, 1997), with the insight that evidence testing can apply the logic of Bayesian inference (Befani & Stedman-Bryce, 2016; Bayes, 1763). Contribution tracing aims to explore causality in situations where traditional counterfactual approaches are difficult, either because the intervention is extremely complex or aimed at a whole population, or for ethical or practical reasons, or because evaluation sponsors are concerned about why an intervention worked, not just whether it worked. This approach was undertaken for selected intermediate outcomes (each corresponding to an individual output brief), in a series of workshops with SNG M&E advisors and technical experts.

The approach to contribution tracing is not further detailed here, as this dissertation does not aim to contribute to the application of this methodology. It is worth noting, however, that the addition of contribution tracing strengthened the credibility of the VFM assessment. It was found that it was possible to make a strong case that SNG activities had produced significant improvements in planning and budgeting processes, but not that these improved planning and budgeting processes had yet improved service delivery.

Additionally, the framework from 2017 was reapplied to consider, at outcome efficiency level, the proportion of increased funding allocations to districts and efficiency gains that may be attributed to SNG. This analysis was informed by the findings from contribution tracing and reaffirmed that SNG should claim ten percent of the total changes in the provincial-level fiscal indicators.

Thematic assessment against the propositions

In this section, the SNG case is systematically assessed to investigate the extent to which it corroborates the model's theoretical propositions.

The first of these propositions is that **CBA can enhance an evaluation of the merit, worth and significance of resource use, by yielding insights that would otherwise be difficult to gain**. This proposition, and all of its sub-propositions, are corroborated in the SNG case study.

Firstly, the CBA used in the SNG VFM assessment followed the prescribed structure and rules which promoted systematic and rational analysis of costs and consequences. CBA methods were used as one component of the SNG VFM assessment. Not all costs and consequences were included;

costs were limited to financial inputs of DFID and the provincial governments. Ideally the CBA should have included the value of impacts of improved health and education on the lives of citizens. In this instance, however, benefits were represented by the value of increased funding flows to districts – a relevant intermediate measure aligned with the program theory of change. In summary, the case corroborates the sub-proposition that CBA promotes systematic and rational analysis of costs and consequences.

Secondly, the case demonstrates the sub-proposition that valuing and reconciling costs and consequences in commensurable units enables commensuration of both sides. This CBA valued costs and consequences in pounds sterling (£), and calculated a benefit:cost ratio from the discounted cashflows. The case corroborates the sub-proposition.

Thirdly, the case illustrates the use of discounting to take the time value of money into account. The CBA used an 8% discount rate to represent the social opportunity cost of capital. The process of discounting adjusted for the differential timing of costs and benefits, converting them to present values. The discount rate was subjected to sensitivity analysis to understand the implication of using a higher or lower discount rate for the results and conclusions of the analysis. The case corroborates the sub-proposition that discounting is a strength of CBA.

Fourthly, the case demonstrates a way in which sensitivity and scenario analysis are strengths of CBA. The CBA explored how the results varied in response to changes in key assumptions such as future rates of utilisation of funding flows to districts and the proportion of total increases in funding attributed to SNG. The analysis concluded that the benefit:cost ratio remained positive under any plausible combination of assumptions, strengthening the conclusion that the program was cost-effective. The case corroborates the sub-proposition that sensitivity and scenario analysis are strengths of CBA.

Fifthly, the case provides an illustration that CBA is capable of meeting a key condition for numerical weighting and synthesis to provide valid results (Scriven, 1991) – namely, that weights can be determined empirically. The CBA compared costs, quantified retrospectively, with benefits, measured retrospectively up to the date of assessment together with forecasting of future benefits over a ten-year time horizon. Values of cost and benefits were predominantly drawn from fiscal data, including the value of DFID and provincial governments' investments in the reforms, and increases in funding to health and education by provincial and district governments. Forecasting of future benefits relied on assumptions (as already described) together with sensitivity and scenario analysis. Overall, the case partially corroborates the sub-proposition (for retrospective valuation only).

The second overarching proposition tested in the case study is that, when it comes to evaluating VFM in social policies or programs, **CBA is usually insufficient to fully answer an evaluation question about VFM.**

The first sub-proposition is that CBA may not capture all criteria. CBA principally provides an estimate of outcome efficiency, whereas VFM in DFID programs also encompasses additional criteria such as good stewardship of resources (economy), delivery of outputs through adaptive programming (dynamic efficiency) and reaching the most disadvantaged groups in the two provinces (equity). The CBA conducted in the VFM assessment estimated the economic efficiency of the program and did not attempt to examine the other criteria. The case corroborates the sub-proposition.

Related to the first sub-proposition, the second sub-proposition is that CBA (in its standard form) does not consider equity separately from efficiency. CBA reflects a normative position on equity (aligned with Kaldor-Hicks efficiency) that any net gain in overall value is worthwhile regardless of distributive impacts. There is a form of CBA that includes distributional weights to adjust the analysis for equity considerations, but this requires a sound basis for setting and justifying numerical weights, which was not considered feasible in this evaluation. The CBA only looked at aggregate value for all citizens and not the distribution of costs and benefits (e.g., differences between the most and least disadvantaged citizens). The case corroborates the sub-proposition.

The third sub-proposition was that CBA reflects a normative position that all goods are fungible and should be valued in commensurable units. The CBA in this case took a limited scope that considered fiscal costs and benefits. This provided a clear estimate of the relative value of program costs and intermediary outcomes. It meant that the costs and benefits in this case truly were fungible as they were all measured in fiscal flows. If a broader scope had been adopted – for example, to include monetary valuations of intangible benefits such as increased trust in government among the citizens of the participating districts, the case would have better illustrated the potential controversies in treating diverse values being fungible. Nevertheless, the case corroborates the sub-proposition.

The fourth sub-proposition is that commensuration may obscure diversity in people's values; aggregation of values using a common metric may obscure differences in the perspectives of different groups, such as differences in social position that can reinforce inequities when values are aggregated. This CBA considered the value of the program in aggregate and made no attempt to consider differences in the perspectives of different groups. The case corroborates the sub-proposition.

The fifth sub-proposition is that commensuration may obscure qualitative differences between things of value. The CBA used funding flows as a proxy for outcomes. The funding flows were quantified, but the quality of their targeting and use, in a diverse range of health and education services, was not. This was an acknowledged limitation of the method and the use of funding flows as a proxy. Nevertheless the case corroborates the sub-proposition that commensuration can mask qualitative diversity.

The sixth sub-proposition is that CBA takes a consequentialist perspective, focusing on costs and consequences without considering processes. This is true of the CBA conducted in this case, and corroborates the sub-proposition.

The seventh sub-proposition is that the scope of a CBA may be constrained by what is measurable. In practice, this can result in important values being excluded from a CBA because they are too hard to estimate. The scope of this CBA was deliberately limited to fiscal costs and benefits. From a pragmatic standpoint this was sufficient to demonstrate a positive return on investment. It would, however, have been challenging to estimate impacts of improved health and education services on the lives of citizens, or improved democracy in Pakistan. Such estimates would have been subject to significant uncertainty in both attribution and monetary valuation of benefits. The case corroborates the sub-proposition.

The eighth sub-proposition is that CBA is not explicitly required to fully adhere to program evaluation standards. In this evaluation, program evaluation standards were not explicitly referenced, so the sub-proposition cannot be tested. In this instance, CBA was one component of a wider evaluation and was conducted on a desktop basis. If CBA had been the only method used it could not have fully met evaluation standards, as explained in the gap analysis.

The propositions and sub-propositions above should be relatively non-controversial as they reflect features of CBA that are well covered in the extant literature. It is in the final proposition that the novel contribution of this research is put to the test. Bearing in mind the strengths and limitations of CBA as outlined above, the model proposes **that a stronger approach would involve explicit evaluative reasoning, supported by judicious use of economic and other methods**. There are ten sub-propositions, nine of which are corroborated in this case study.

First, the case corroborates the sub-proposition that the model can accommodate multiple and diverse criteria. This evaluation included explicit criteria and standards for equity as well as efficiency – demonstrating that the approach can accommodate multiple criteria.

Second, the case corroborates the sub-proposition that the model allows equity to be considered separately from efficiency. For example, it can take Kaldor-Hicks efficiency into account without being limited to this position exclusively. The VFM assessment of SNG used equity criteria tailored to the program and aligned with its theory of change, assessing the extent to which changes in needs-based planning and resource allocation contributed to reducing inequities by targeting resources to poor people, women and girls.

Third, the case corroborates the sub-proposition that the model does not require commensuration; it has the ability to contrast and support deliberation on qualitative differences between values as well as the

ability to aggregate them. In this evaluation, value was defined at multiple levels, for multiple criteria, using a mix of quantitative and qualitative standards and evidence. The VFM assessment did not, however, involve deliberation on qualitative differences in the values held by different groups of people. This would have been possible but was not deemed necessary in this case. Nonetheless, the case demonstrates the central point that commensuration is not required in the VFI model.

Fourth, the case corroborates the sub-proposition that the model does not take an exclusively consequentialist perspective. A practical upshot of this is that it can be used to evaluate VFM of implementation and delivery, independently of whether outcomes are achieved or evaluated. Using DFID's VFM criteria, the evaluation separately evaluated economy (good management of resources to procure inputs) and efficiency (good management of inputs to produce outputs), as well as consequentialist criteria of effectiveness (achievement of objectives), cost-effectiveness (value of outcomes relative to costs) and equity (reaching the intended groups). The overall judgement of VFM took both the consequentialist and the process-based evaluations into account.

Fifth, the case corroborates the sub-proposition that the model can incorporate a broader range of analytical options about value than CBA. This evaluation took an analytic approach by first evaluating economy, efficiency, effectiveness, cost-effectiveness and equity separately, and then using the five evaluative judgements collectively to reach a judgement about VFM overall. Within each criterion, judgements were made holistically against multiple sub-criteria using rating rubrics. A mix of qualitative and quantitative synthesis methods were used, with quantitative synthesis being an inherent part of CBA and qualitative synthesis being used as the overarching approach to integrating multiple streams of evidence. As a result, it was possible to work with diverse values in ways that would have been challenging in CBA.

Sixth, the case corroborates the sub-proposition that the model does not prescribe the methods to be used to gather evidence; rather it is flexible to enable a mix of methods to be matched to context, recognising that no single method can address all VFM criteria and that no single method should be regarded as gold standard. In this evaluation, in fidelity with the model, decisions about what forms of credible evidence were necessary and sufficient to support judgements about performance and VFM were made after the criteria and standards had been agreed. This sequence of evaluation design ensured the methods of data collection addressed the aspects of VFM that were considered important to stakeholders.

Seventh, the case corroborates the sub-proposition that the model has the ability to accommodate mixed methods evidence. Methods included analysis of program financial data, monitoring reports, narrative evidence from output briefs, provincial budget tracking data, and CBA. These multiple streams of evidence contributed to judgements against the

criteria and standards and enabled triangulation of findings from different sources.

Eighth, the case corroborates the sub-proposition that the model has the ability to incorporate economic evidence. The VFM assessment included, but was not limited to, economic analysis of program costs and consequences. The results of economic analysis supported determination at the outcome efficiency level of the VFM assessment, alongside other evidence and reasoning used to address the other four criteria.

Ninth, and building on the observation above, the case corroborates the sub-proposition that the VFI model and CBA are compatible and can be integrated. The overarching model of explicit evaluative reasoning can incorporate CBA as one of its supporting methods, and doing so can strengthen evaluation compared to using either approach without the other. This evaluation gained from the use of CBA (as validated in the first set of propositions above) while mitigating some of its weaknesses (as validated in the second set of propositions).

The tenth sub-proposition was that the model can be applied in full adherence with program evaluation standards. This sub-proposition was not tested through documentary evidence in the case study, as program evaluation standards were not explicitly referenced in the VFM evaluation of the SNG program. Retrospective assessment against the standards suggested the evaluation had generally been conducted in adherence, but some areas were identified where the standards could have been followed more comprehensively. For example, citizens of the participating districts were not consulted on the evaluation design and methods. No standards were identified that the assessment would have been unable to adhere to.

The findings above are summarised in Table 14 in the following section, which systematically examines the replication of findings across the two case studies.

Replication

Together, the two case studies follow a replication design (Yin, 2009). Under this design, the conclusions from each individual case study can be used to expand and generalise a theory (analytic generalisation). The conclusions from each case can also be compared with other cases, and may be considered more potent if two or more cases support the same theory (Yin, 2009). In this study, replication is investigated with regard to the theoretical propositions. The main proposition is that, given the strengths and limitations of CBA, a stronger approach should usually involve explicit evaluative reasoning and mixed methods, incorporating economic evaluation without limiting the evaluation to this method alone.

The 'rival' argument is that CBA is the gold standard for evaluating VFM. The cases do not directly provide evidence for or against the rival explanation, because stand-alone CBAs were not conducted on the two programs. Although CBAs were conducted as part of each evaluation, they were intended to address one criterion only (outcome efficiency) and their scope was kept narrow to include only costs and benefits that were fiscal or readily monetisable. If CBA had been the only method used, a decision might have been made to include a wider scope of costs and benefits. Nonetheless, the relative merits of CBA have been explored through thought experiment, in reference to the propositions.

Table 14 summarises the theoretical propositions and the conclusions reached in regard to each case. Replication is then assessed.

Findings

Overall, the two cases corroborate the proposition that **CBA can enhance an evaluation of the merit, worth and significance of resource use** by providing insights that would otherwise be difficult to gain. Four of the five sub-propositions are replicated, reinforcing the stated strengths of CBA – namely, that CBA: promotes systematic and rational analysis of costs and consequences; enables costs and consequences to be valued and evaluated in commensurable units; uses discounting as a way to rationally consider the opportunity cost of the investment; and enables transparent and robust thinking about uncertainty and risk through the use of sensitivity and scenario analysis.

In both cases, CBA contributed analysis of benefits and costs valued in pounds sterling (£) which were commensurated into a single indicator (net present value or benefit:cost ratio), providing results that would have been difficult to reliably intuit. In both cases, a discount rate was selected that represented the social opportunity cost of the investment. This means that a positive net present value (if achieved) would indicate that the program compared favourably to alternative uses to which the resources could have been applied. In both cases, sensitivity analysis provided transparency about the extent to which key assumptions impacted on results. In both cases, scenario analysis was undertaken to

better understand the sets of conditions and assumptions required for benefits to exceed costs. These findings are consistent with the literature on CBA (e.g., Drummond et al., 2005; Levin & McEwan, 2001) and represent expected strengths of CBA conducted in fidelity with the standard methodology.

The fifth proposition – that CBA can accurately measure values – was partially demonstrated in the SNG program through the retrospective valuation of costs and benefits. Overall, however, both cases highlight that it is not always possible for the values used in CBA to be determined empirically in real-world evaluations. In particular, both cases involved modelling of future costs and benefits, which necessitated the use of assumptions.

The two cases also corroborate the proposition that in social policies and programs, **CBA is usually insufficient to fully answer an evaluative question about the merit, worth and significance of resource use.** Seven of the eight sub-propositions were replicated. Specifically, the two cases provide clear demonstrations of some limitations of CBA: not capturing all criteria; not considering equity separately from efficiency; treating all values as fungible; obscuring diversity in people's values; obscuring diversity in things of value; ignoring processes; and excluding values that are too hard to monetise.

In both cases, the mandatory application of DFID's VFM criteria required consideration not only of outcome efficiency (or 'cost-effectiveness') but also of program resource management ('economy'), the quality, quantity and productivity of processes and delivery ('efficiency'), achievement of objectives ('effectiveness') and the extent to which program processes and outcomes reach the intended priority groups ('equity'). These aspects of performance could not have been fully or adequately explored, had the VFM assessment been restricted to CBA alone.

DFID's allocation of resources to international development programs recognises that "reaching marginalised groups may entail additional effort and cost" (ICAI, 2018, p. i). In both cases, CBA, if conducted as a stand-alone analysis, would have assessed the aggregate value of the programs without explicit reference to their performance in reaching and benefiting their intended target groups. This is consistent with the expectations of standard CBA methodology (Drummond et al., 2005; Levin & McEwan, 2001) – though, as already discussed, it is theoretically possible, albeit technically challenging, to add distributive weights to CBA (Adler & Posner, 2006).

In both cases, CBAs valued costs and benefits monetarily, aggregating all values into a single indicator of net value. This inherently treated the values as being fungible, consistent with standard CBA methodology (Drummond et al., 2005). In both cases, CBAs examined net value at a level of abstraction that did not consider diversity in individual values or qualitative differences between things of value. This is an inherent feature of commensuration. It does not mean that economic analysis cannot

consider such diversity, only that it is not built into the standard prescription for the CBA method and must be analysed using different methods.

A worthy adjunct to the standard method may be the conduct of CBAs from the perspectives of different groups, to better understand the incidence of costs, benefits and net value (Gargani, 2017). Such analysis could contribute evidence in support of judgements about non-efficiency values related to equity and social justice. This possibility serves to further highlight the multi-criterial nature of VFM and the need to make an overall determination from multiple streams of evidence. It provides a further illustration of a way in which CBA could provide evidence to support an overarching process of evaluative reasoning.

The eighth sub-proposition, that CBA is guided by technical but not ethical considerations, and is not able to fully adhere to program evaluation standards, was not directly tested in the case studies but was adequately demonstrated at a theoretical level in the gap analysis already presented.

Together, the two case studies corroborate the main proposition of the theoretical model: that **given the strengths and limitations of CBA, a stronger approach would involve explicit evaluative reasoning, supported by mixed methods including economic methods where feasible and appropriate**. Of the ten sub-propositions, nine were replicated and one (full adherence with program evaluation standards) was not tested.

In both cases, the VFM assessments accommodated multiple and diverse criteria, and made the trade-offs between them explicit when reaching an overall judgement about VFM. The inclusion of explicit equity criteria in each case required the evaluations to examine program performance in reaching and benefiting their intended target groups, in addition to their efficiency. Both cases demonstrated that the model did not require commensuration of values; in both cases, value was defined at multiple levels, for multiple criteria, using a mix of quantitative and qualitative standards and evidence.

On one hand, combining CBA with broader criteria and evidence weakens one of the principal benefits of CBA in that it departs from measuring all value consumed and created by the program in monetary units, instead opting for a system of qualitative weighting and deliberative balancing. This is not a fatal flaw, however. The move to deliberative balancing prompted clarity in other ways. In particular, it allowed for an inclusive and transparent appraisal of VFM, explicitly balancing efficiency and equity, qualitative evidence, and features of program performance such as responsiveness to evolving context which would not have been evaluated using CBA alone.

In both cases, the inclusion of DFID's five VFM criteria expanded the evaluation beyond a consequentialist perspective alone, and meant that the VFM assessments could be (and were) used for formative purposes,

supporting learning and improvement as well as summative and accountability purposes.

In both cases, a mix of valuing approaches were used. Judgements were made analytically against DFID's criteria – defined through multiple sub-criteria set out in rating rubrics. VFM was defined as the collective performance across the five criteria together. A mix of qualitative (using rubrics) and quantitative (using CBA) synthesis methods were used. As a result, it was possible to work with diverse values in ways that would have been challenging within a textbook CBA. Presenting explicit evaluative judgements made both the evidence and the reasoning transparent, opening the evaluation to scrutiny. As a result, findings were more openly traceable and challengeable – and this improved their credibility to stakeholders. This was important in the DFID context where both programs' VFM assessments formed part of the information turned over to external annual review teams.

In both cases, decisions about what forms of credible evidence were necessary and sufficient to support judgements about VFM were made: on the basis of systematic analysis of agreed criteria and standards; and in consultation with stakeholders. The sequential approach to evaluation design (first defining aspects and levels of performance, and subsequently determining the metrics and methods to be applied) ensured methods of data collection addressed the priority aspects of performance and VFM that were agreed, by stakeholders, to be important.

In both cases, addressing the criteria and standards involved mixed methods, collecting and analysing quantitative and qualitative sources of data, including financial and administrative data, monitoring reports, narrative evidence, results of evaluations already conducted, and economic analysis of program costs and benefits. Both cases incorporated CBA and demonstrated possible ways in which the findings from CBA can be integrated within a larger evaluation. As shown through the corroboration of the propositions above, doing so can strengthen evaluation compared to using either approach without the other.

Table 14: Replication analysis

Core proposition	Sub-propositions	Rating (MUVA)	Rating (SNG)	Replication
CBA can enhance an evaluation of the merit, worth and significance of resource use by yielding insights that would otherwise be difficult to gain.	CBA promotes systematic and rational analysis of costs and consequences	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	Valuing and reconciling costs and consequences in commensurable units provides unique insights	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	Discounting is a strength of CBA	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	Sensitivity and scenario analysis are strengths of CBA	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	CBA can accurately measure values	Case does not corroborate sub-proposition	Case partially corroborates sub-proposition	Not replicated
When it comes to evaluating VFM of social policies and programs, CBA is usually insufficient to fully answer an evaluative question about the merit, worth and significance of resource use.	CBA may not capture all criteria	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	CBA (in its standard form) does not consider equity separately from efficiency	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	CBA treats all values as fungible	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	Commensuration may obscure diversity in people's values	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	Commensuration may obscure diversity in things of value	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	CBA is concerned with costs and consequences, not processes	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	The scope of a CBA may be constrained by what is measurable	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	CBA is not explicitly required to fully adhere to program evaluation standards	Sub-proposition not tested in the case study	Sub-proposition not tested in the case study	Not replicated

Given the aforementioned strengths and limitations of CBA, a stronger approach would involve explicit evaluative reasoning, supported by judicious use of economic and other methods (VFI). The minimum requirements for such an approach are that it: pose, and answer, an evaluative question about the merit, worth and significance of resource use; use explicit evaluative reasoning; match methods to context; and be conducted in keeping with program evaluation standards.	VFI can accommodate multiple and diverse criteria	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI allows equity to be considered separately from efficiency	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI does not require commensuration	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI does not take an exclusively consequentialist perspective	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI can incorporate a broader range of valuing options than CBA	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI does not prescribe the methods to be used to gather evidence	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI has the ability to accommodate mixed methods evidence	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI has the ability to incorporate economic evidence	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI and CBA can be combined	Case corroborates sub-proposition	Case corroborates sub-proposition	Replicated
	VFI can be conducted in full adherence with program evaluation standards	Sub-proposition not tested in the case study	Sub-proposition not tested in the case study	Not replicated

Overview of findings

Aside from being VFM evaluations of DFID-funded international development programs, the two case studies represented very different contexts in which to apply the model of VFM assessment. The SNG program aimed to improve democracy in Pakistan by strengthening systems of governance and public financial management. Its locus of influence was principally top-down, providing technical assistance to government agencies to improve the allocation of resources to meet needs. In contrast, MUVA's locus of influence was bottom-up, working in urban Mozambican communities with local organisations, young women and girls, to test new approaches to female economic empowerment, to learn from them, and to influence the adoption of successful approaches locally and internationally. The replication of findings in these two different contexts strengthens the generalisability of the theoretical propositions. Findings are summarised as follows.

Applicability of the theoretical propositions

By applying the model of VFM assessment in the two contexts, and systematically analysing the applicability of the models' theoretical propositions, it has been found that each case study corroborates nearly all sub-propositions, and that the findings are replicated across nearly all sub-propositions.

The case studies demonstrate that CBA can enhance an evaluation of VFM by promoting systematic and rational analysis, by using commensurable units for costs and consequences, by incorporating opportunity cost into the analysis through discounting, and through the use of scenario and sensitivity analysis. In both cases, CBA was used to address DFID's criterion of 'cost-effectiveness', by providing an estimate of the net present benefit or benefit:cost ratio of the program. In both cases the analysis clarified the parameters under which the programs would create more value than they consumed.

Neither case study corroborated the sub-proposition that CBA is capable of being conducted entirely on the basis of empirically-measured values. The two cases did not 'disprove' this – it is a point of fact that methods exist for measuring and valuing outcomes in retrospective program evaluations (Gargani, 2017). Rather, they demonstrated that it is not always feasible for weights to be determined empirically within real-world constraints of available data, time, resources, and/or expertise. In both cases, some values were observed or measured (for example, market wages) while others were not (for example, future rates of employment). In both cases, CBA entailed modelling future costs and benefits, which meant it was necessary to incorporate assumptions into the models. Another well-documented feature of CBA – the ability to explore the implications of uncertainty and risk through scenario and sensitivity analysis – helped to compensate for this. Overall, it can be concluded that CBA enhanced VFM assessment in both cases.

The case studies provided two examples of situations in which CBA would have been insufficient to fully answer an evaluative question about VFM – demonstrating that CBA, in such cases, falls short of being a 'gold standard' for evaluating VFM. In particular, circumstances exist in which CBA: may not capture all criteria; does not consider equity separately from efficiency; treats all values as fungible; obscures diversity in people's values or diversity between things of value; focuses on costs and consequences divorced from processes; and can be constrained by what is measurable in a particular context.

In both cases, CBA was able to contribute an estimate of outcome efficiency (addressing DFID's 'cost-effectiveness' criterion) but would have struggled to support a multi-criterial assessment of frugal management of resources, the value of adaptive management and learning, and the extent to which program processes and outcomes supported equity-related objectives. Furthermore CBAs in both cases would have struggled to incorporate some of the intangible benefits of the programs (such as improved agency of young Mozambican women, and improved democracy in Pakistan).

While the strengths and limitations of CBA were derived from literature, the final proposition offered a novel theoretical framework for evaluation of VFM. It was this proposition, in particular, that warranted empirical investigation.

It was proposed that a stronger approach than the use of CBA alone would involve explicit evaluative reasoning, supported by judicious use of economic and other methods. It was suggested that the minimum requirements for such an approach would be that it: pose, and answer, an evaluative question about the merit, worth and significance of resource use; use explicit evaluative reasoning; match methods to context; and be conducted in keeping with program evaluation standards. A practical model was developed, comprising eight steps for implementing the theoretical model, using rubrics and mixed methods.

The two case studies were conducted with fidelity to the practical model. Both cases corroborated the sub-propositions that such an approach: can accommodate multiple criteria; can accommodate any normative position on equity; does not require commensuration; does not take an exclusively consequentialist perspective; can incorporate a broader range of valuing options than CBA; does not prescribe the methods that should be used to gather evidence; has the ability to accommodate mixed methods evidence; and can incorporate results of economic analysis. Both cases demonstrate that qualitative weighting and synthesis and CBA are not mutually exclusive – they can be combined, and doing so can strengthen evaluation compared to using either approach without the other. All of these sub-propositions were replicated in both cases. One further sub-proposition – that the model can be implemented in fidelity with program evaluation standards – was not empirically investigated but is well covered from a theoretical perspective in the gap analysis.

The analysis overall provides analytic generalisation (Yin, 2009) from the cases to the theory, through the corroboration and replication of theoretical propositions. Experience from the case studies also reveals some ways in which the theoretical and practical models could be refined. These are summarised below.

Refinements to the model

The process followed the principle of parsimony, providing a level of guidance considered to be necessary and sufficient to guide a seasoned evaluator in conducting a VFM assessment that meets the theoretical requirements for a good VFM assessment. Experience from the two case studies indicated that while the model was generally fit for purpose, there were some aspects of the model that might be investigated and developed further.

The first area of possible refinement to the model relates to the value of taking a **participatory approach to the evaluative reasoning process**. Both case studies involved participatory approaches, reflecting contemporary evaluation-specific models of validity and credibility (Donaldson et al., 2015; Griffith & Montrosse-Moorhead, 2014; Wehipeihana & McKegg, 2018). Accordingly, the evaluations treated stakeholders as a critical source of values, promoting dialogue and collective valuing as well as individual values in evaluation (House & Howe, 1999; Julnes, 2012b). It may be posited that if the function of a rubric (or some other presentation of criteria and standards) is to articulate an agreed set of inter-subjective values, and if rubrics are vulnerable to shared bias in their development and use, and if engaging an appropriately broad mix of stakeholder perspectives in dialogue reduces bias, then a well-managed participatory approach should improve the validity of evaluation.

Throughout the phases of an evaluation, primary intended users and other key stakeholders were engaged in defining the programs and their theories of change, identifying and agreeing appropriate criteria and standards, identifying credible evidence sources, determining what methods were acceptable and feasible to provide credible evidence, and interpreting evidence against the criteria and standards to reach evaluative judgements about VFM (Martens, 2018b). Through these processes, shared understandings were reached. As a result, evaluation credibility and use may be enhanced (King et al., 2013).

The second area for further investigation is the opportunity to **more explicitly align VFM assessment with other monitoring and evaluation activity**. Conceptually, VFM criteria such as efficiency, effectiveness and equity overlap with the focal points of process and outcome evaluations, and evaluations into specific aspects of program design and delivery such as approaches to gender equity. Given the potential for overlap and duplication, deliberate steps should be taken to align and coordinate VFM assessment with monitoring and evaluation, in several ways.

Firstly, decisions need to be made about the scope of a VFM assessment relative to the scope of other evaluation activity, and to identify where areas of overlap exist. This is both a conceptual and a pragmatic decision; conceptually, all evaluation could be subsumed within the ambit of VFM – for example, the merit, worth and significance of resource use arguably spans all of the Development Assistance Committee (DAC) evaluation criteria (OECD DAC, 2012). Pragmatically, however, in the case studies, the timing of VFM assessments was more frequent (at least annual) than that of process and outcome evaluations (which may only be done once over the life of a multi-year program), and the five overarching VFM criteria were stipulated which provided a boundary for the VFM assessment.

Secondly, the content of a program's VFM framework should cohere with the content of monitoring and evaluation frameworks in the same program. For example, VFM and evaluation frameworks should be aligned with a unifying theory of change and should use criteria, standards and other definitions that are conceptually consistent, or at least non-contradictory. This conceptual alignment should support the selection of methods by considering how each method will be implemented and how they will interact with each other through the evaluation process (McKegg et al., 2018). This alignment should also facilitate consistency, clarity and credibility of evaluation findings, thereby supporting evaluation use (Patton, 2008).

Thirdly, for the sake of efficient evaluation processes, VFM assessment should be coordinated with other monitoring and evaluation activity to avoid overlap and duplication of effort, and to ensure the right evidence is gathered at the right time to serve both purposes. For example, in the SNG program, collation of evidence in the first VFM assessment took considerable time and effort. In the second VFM assessment, VFM reporting was better harmonised with other routine reporting and was undertaken more efficiently (King & Allan, 2018). In the MUVA program, evaluations of trialled approaches to female economic empowerment formed part of the evidence used in the VFM assessments.

Ideally, the team responsible for VFM evaluation, and other teams responsible for monitoring and evaluation, should collaborate from the outset to ensure alignment and coordination of their respective evaluation designs, methods and processes. For example, all of these teams should be involved in developing a theory of change to ensure it meets the needs of all evaluations in addition to providing a valid representation of the program.

The third area for further development in the prototype model was the opportunity to strengthen VFM assessment by seeking to understand and **evaluate VFM from the perspectives of intended beneficiaries**. DFID's demand for VFM assessment stems principally from political accountability drivers, though the value of VFM assessment for learning and improvement is also recognised (ICAI, 2018). Accordingly, the two VFM assessments featured in the case studies principally took a donor-

centric perspective of VFM. Neither VFM assessment involved beneficiaries directly in the evaluation design or synthesis of findings. However, the MUVA program was strongly participatory, involving stakeholders – including young urban Mozambican women and girls – in the design of its theory of change, the design of approaches to female economic empowerment, and evaluations of those approaches. The VFM assessment built on these elements, so indirectly reflected beneficiary input. However, the VFM assessment could be further strengthened by incorporating analysis of costs and benefits from the perspectives of beneficiaries.

The fourth opportunity for further development of the model was an acknowledgement of the role of **systems thinking and complexity-informed approaches** in evaluation of adaptive programs. The model was designed with the intent of being sufficiently flexible to accommodate a diversity of evaluation approaches. In both case studies, the design and conduct of program-specific VFM assessments drew on insights from the literature on complexity in program management (Andrews, 2010; Andrews, 2013; Andrews, Pritchett & Woolcock, 2017; Olson & Eoyang, 2001). For example, both cases recognised that there may be multiple actors and the behavior of the system may be unpredictable; that direction-setting when attempting to bring about system change requires participatory approaches with multiple stakeholders (rather than a top-down approach); and that causality may be non-linear, with multiple variables and inter-relationships (Olson & Eoyang, 2001). In such contexts, the performance of programs needs to take into account not only what was delivered (or not) and how well, but also *why*, recognising that when a program is responsive to an evolving context, some parts of its strategy may not be implemented, and some new strategies may be added as time goes on (Mintzberg & Waters, 1985). In such contexts, responsiveness to context may be more relevant to VFM than delivery to work plans.

These considerations were reflected in revisions to the model, which was subsequently published by King and OPM (2018) (see *Appendix*).

Summary

Two case studies were carried out, applying the process model to evaluations of VFM in two international development programs. The case studies: illustrated the use of the model in practice; provided real-world settings to investigate the applicability of the theoretical propositions; and provided experiential evidence to inform future refinements to the model of VFM evaluation.

The case studies corroborated the propositions of the theoretical model. Findings were replicated across the two case studies to a sufficient extent that analytic generalisation from the case studies to the theoretical model may be claimed. The model was tested on two international development programs but, in theory, is applicable to any context where a social investment is being evaluated for VFM. It is sufficiently flexible to

incorporate any mix of criteria, standards and methods, and evaluator orientations, within the general framework.

When evaluating VFM in international development programs, the design and conduct of VFM assessments should be carried out in the manner described by the eight-steps of the process model. VFM assessments should use explicit evaluative reasoning, select methods according to context, seek to make use of economic evaluation to the extent feasible and appropriate, and balance this with other methods that contribute other forms of evidence and address criteria that complement CBA. Experience from the case studies suggests that there is value in such evaluations being conducted in a participatory manner and with a view to ensuring evaluation use.

The practical model for VFM evaluation is fit for purpose. Experience from the case studies identified ways in which the model can be improved and refined. As is appropriate in theory development, the model is not 'finished' and remains open to further research, development and revision.

Chapter 8: Discussion

This thesis aimed to contribute new knowledge about what VFM means and how it should be evaluated. It has done so through the development of conceptual and process models, which were investigated through case studies. This chapter addresses these aims and critically evaluates the extent to which they were met.

The chapter first provides a summary of the results from the research, and what they cumulatively contribute to the field of evaluation. The results are evaluated against commonly-held virtues of a 'good' theory. Following this, limitations and research opportunities are identified. The chapter concludes with remarks about the nature of VFM and the requirement for sound evaluative reasoning, with methods being selected and used in the service of robust evidence and logical argument.

Summary of findings in this thesis

This study was a new area of research, investigating the integration of evaluation-specific methodology with economic methods of evaluation to address VFM. The contribution of this research is evidenced in its end products. A conceptual model has been developed, proposing a definition of VFM and a set of requirements for good evaluation of VFM in social investments. The research suggests that CBA, in theory and in practice, falls short of being a gold standard for meeting these requirements, but can contribute in important ways to an evaluation of VFM. The research contributes a process model, laying out a series of steps and principles to guide an evaluation of VFM that meets the requirements of the conceptual model. Testing of the conceptual and process models through case studies provides proof of concept. The significant and novel contributions of this thesis are summarised as follows.

This research has found that **VFM is an important construct** that is broader than efficiency, return on investment, or aggregate wellbeing. **VFM is an evaluative question** – a question demanding a judgement, based on logical argument and evidence (Davidson, 2005; Fournier, 1995; Patton, 2018; Schwandt, 2015). Resource allocation decisions need to be evaluated against multiple criteria, which might entail trade-offs, ambiguities, deliberative dialogue, and tension between different cultural values and ways of knowing (Goodwin et al., 2015; House & Howe, 1999; Schwandt, 2018). VFM cannot be determined by an algorithm or subbed out to a formula, though algorithms and formulae can help provide insights from evidence (King, 2015). Evaluative reasoning (Scriven, 1991; Yarbrough et al., 2011), to answer a VFM question, is mandatory.

VFM is a shared domain of two disciplines: evaluation, through its connection to the evaluative concepts of merit, worth and significance (Scriven, 1991); and economics, through its connection to the fundamental economic problem of resource allocation (Drummond et al., 2005). This research has grappled with the validity and feasibility of

integrating paradigms and procedures from both disciplines, and found this to be viable through a model of evaluative reasoning. Evaluation and economics can and should be combined to address questions about VFM.

CBA is not a rival to evaluation – it *is* evaluation, in that it implements the general logic of evaluation through a sophisticated system of valuing and synthesis (King, 2017). Moreover, the system of valuation in CBA, which converts gains and losses in utility into money, and aggregates them (Adler & Posner, 2006) gives CBA distinctive strengths that can bring unique insights to evaluation. **CBA estimates an important dimension of VFM** – net benefit or aggregate welfare (Sunstein, 2018). It does so imperfectly (Adler & Posner, 2000; 2006; Julnes, 2012b) – but it is the most fit-for-purpose method available for addressing this dimension (Adler & Posner, 2006). "Whether or not an analysis of costs and benefits tells us everything we need to know, at least it tells us a great deal that we need to know. We cannot safely proceed without that knowledge" (Sunstein, 2018, p. xi). Evaluators should use CBA more (Gargani, 2017; Julnes, 2012c; Yates, 1996).

CBA, however, is not the whole evaluation. Efficiency is "morally relevant, but not necessarily morally decisive" (Adler & Posner, 2006, p. 154). There are definable circumstances in which CBA may have capacity to enhance evaluation of VFM, but cannot comprehensively address the evaluative question (Drummond et al., 2005; Gargani, 2017; Levin & McEwan, 2001; Sunstein, 2018). Issues such as relevance, sustainability, equity or distributive justice, cultural and historical significance, and deontological ethics, are all relevant to VFM in social investments (Adler & Posner, 2006; Boston, 2017; Pinkerton et al., 2002). Intangible values, including social, cultural and collective values, matter (Goodwin et al., 2015; Julnes, 2012b). Multiple forms of evidence (qualitative and quantitative) and ways of creating knowledge should contribute to evaluative judgements about complex social issues (Deane & Harré, 2016; Greene, 2005; Mertens & Hesse-Biber, 2013; Wehipeihana & McKegg, 2018). Properly balancing diverse values and power differences requires deliberation rather than aggregation (House & Howe, 1999; Julnes, 2012b).

Given the centrality of these issues in social policies and programs, CBA will usually be insufficient on its own. CBA should be used more widely in evaluation, but not as a stand-alone method. **CBA should be used in a supporting role to a wider process of evaluative reasoning and in conjunction with other methods**.

This appraisal of CBA provides a strong argument for the use of **evaluative thinking** (Patton, 2018; Vo & Archibald, 2018). Understanding the strengths and limitations of methods, interpreting findings in light of those strengths and limitations, and bringing wider considerations into evaluative judgements, should always feature in safe and effective use of evaluation methods.

Explicit evaluative reasoning is essential to making valid judgements about VFM. The conceptual model, process model and case studies highlight the worth of evaluation as a discipline, and evaluative reasoning as the backbone of evaluation. Evaluative reasoning is the "superprocedure" that eluded Adler & Posner (2006, p. 158), an overarching approach capable of combining the insights from CBA with additional considerations, which holds the key to using economic methods evaluatively (King, 2017). As much as evaluators need economic analysis, economists need evaluative reasoning.

In social policies and programs, where VFM may have multiple and diverse criteria, **qualitative weighting and synthesis** is fit for purpose as an overarching process of reasoning to evaluate VFM, combining criteria and metrics from economic evaluation with wider criteria and evidence. While the theory and practice of evaluation includes diverse approaches to evaluative reasoning, the approach developed and tested in this research built on a technocratic (Schwandt, 2015) foundation of qualitative valuing and synthesis (Scriven, 1994), using rubrics (Davidson, 2005) as a practical tool. Evaluation theory, corroborated in the case studies, suggests that qualitative valuing and synthesis serves a wider range of evaluation purposes and contexts than numerical approaches (including CBA). This does not preclude the use of other approaches to evaluative reasoning where circumstances allow.

In addition to satisfying the requirements of a technocratic approach to evaluative reasoning, the processes of developing and using rubrics in the case studies was intentionally designed to **foster stakeholder engagement and participation in evaluation**. Reflecting wider literature on rubrics, this approach facilitated situational responsiveness, validity, and evaluation use (Davidson, 2005; 2014; Dickinson & Adams, 2017; King et al., 2013; McKegg et al., 2018; Martens, 2018b; Wehipeihana et al., 2018). The use of rubrics provided a focal point for engagement with evaluation users and other program stakeholders, facilitating negotiation and explicit agreement of the basis upon which judgements should be made and the types of evidence that should be gathered in the service of the evaluation, prior to data collection commencing (Davidson, 2005; 2014; Dickinson & Adams, 2017; King et al., 2013; McKegg et al., 2018; Oakden & King, 2018; Wehipeihana et al., 2018). In this way, the use of rubrics fostered communication with stakeholders about the intended uses of VFM assessments and helped to focus the evaluation design.

During VFM assessment, criteria and standards similarly provided a central point of reference for stakeholder engagement in reviewing the evidence and making evaluative judgements. In reporting findings, criteria and standards provided a transparent link between the evidence and judgements, opening findings to scrutiny and making judgements challengeable. In these ways, the use of rubrics helped to position evaluative reasoning as "a collaborative, social practice" (Schwandt, 2018, p. 125) – providing opportunities to contextualise and validate

judgements by accessing stakeholders' in-depth knowledge, providing an inclusive and rigorous judgement-making process, and reducing potential bias in rubric design and use (King et al., 2013; Wehipeihana et al., 2018). Tacit and deliberative approaches to evaluative reasoning (Schwandt, 2015) and experiential judgements of quality (Stake & Schwandt, 2006) were called into action with program stakeholders, as strategies for guiding, checking, challenging and validating judgements.

Program evaluation standards should guide economic evaluation. Perhaps reflecting core values of detached rationality that are tied to economic thinking and positivism (Sunstein, 2018), standards for the conduct of economic evaluations focus on technical aspects of applying economic methods to ensure precision, accuracy and reliability. Ethical principles, not explicit in economic evaluation methodology, are also required. For example, as recently proposed by one economist:

> *First, act in service to human prosperity in a flourishing web of life, recognising all that it depends upon. Second, respect autonomy in the communities that you serve by ensuring their engagement and consent, while ever aware of the inequalities and differences that may lie within them. Third, be prudential in policymaking, seeking to minimise the risk of harm – especially to the most vulnerable – in the face of uncertainty. Lastly, work with humility, by making transparent the assumptions and shortcomings of your models, and by recognising alternative economic perspectives and tools* (Raworth, 2017, p. 138).

Using program evaluation standards to guide economic evaluation requires that the evaluator (or economist) remain open to the possibility of *not* conducting an economic evaluation. Selection of methods should be contextually determined. No method should be preordained or imposed as the 'best approach' without due concern for consequences and influence, attention to stakeholders, negotiated purposes, and contextual viability.

The case studies are themselves a contribution to the evaluation field. The case study analysis models the use of probative inference – "inference to a conclusion that has been established, so the utterer claims and is prepared to support, as beyond reasonable doubt" (Scriven, 2012, p. 23). Theoretical propositions were identified that differentiated the strengths and limitations of CBA, and the features of a proposed 'Value for Investment' (VFI) model of evaluation that might harness the strengths of CBA while compensating for its limitations. These propositions, together with evidence from the case studies, underpinned a process of reasoning to assess the validity of the conceptual model. The theoretical propositions can be used again in future research and meta-evaluation of the VFI model. Similarly, the general approach of developing such propositions about the nature of evaluation, and testing them empirically through case studies, can be used in other research on evaluation.

Is it a good theory?

Theory-building research aims to contribute to an ongoing process of knowledge accumulation (Lakatos, 1970). Neither CBA, nor the VFI model developed in this research, should be described using terms like 'best practice' or 'gold standard'. The findings of this research should more appropriately be seen as contributions to the 'interim struggle' of theorising (Shepherd & Suddaby, 2017).

The approach to theory building was, at its core, a deductive approach, moving from literature to theory development to empirical testing (Wacker, 1995). Within this overarching model, however, the research demonstrated pragmatic empirical theorising and engaged scholarship as described by Shepherd & Suddaby (2017). Throughout the phases of research, theory development was informed by exploring the interface between theory-based and practice-based knowledge, as described by Coryn & Stufflebeam (2014).

Model development, as is often the case (Shepherd & Suddaby, 2017) started with a conjecture, born of real-world experience. The notion of reaching across disciplinary boundaries to combine evaluative and economic thinking was explored through literature. Literature search and review evolved and expanded the model as unanticipated concepts came to light. The model was presented to evaluation theorists and practitioners at conferences, and subjected to peer review, which raised new conundrums to be explored through further literature search. The process model was operationalised by combining literature with experiential knowledge. The steps in the process started as a concept sketch. Literature clarified the sequence of steps, the rationale and processes within each step, and the overarching principles that should guide the whole process. When the model was tested through case studies, formal and systematic analysis against theoretical propositions supported analytic generalisation and replication (Yin, 2009), while tacit learning and reflection provided additional depth, identifying wider issues and possible refinements to the model.

Features of 'good' theory are replete in the literature (Coryn & Stufflebeam, 2014; Miller, 2010; Patterson, 1986; Popper, 1957; Quine & Ullian, 1980; Shepherd & Suddaby, 2017; Wacker, 1995) and offer a checklist for considering whether the theory developed in this research should be considered a good theory.

According to the criterion of uniqueness, good theory can be differentiated from other theories while striking a balance between novelty and conformity (Shepherd & Suddaby, 2017; Wacker, 1998). This theory makes a number of unique contributions, as noted above. It defines VFM as a distinct construct that is broader than efficiency; it connects VFM to the two disciplines of evaluation and economics; it identifies CBA as an approach to evaluative reasoning; it establishes that CBA is valuable but insufficient to address VFM questions; it sets out a model for integrating CBA with evaluative reasoning and mixed methods; and it argues for

economic evaluation to be guided by program evaluation standards. These contributions link existing evaluation and economic theory through a novel definition of VFM. According to the criterion of uniqueness, this is a good theory.

A good theory is also conservative – it has few rivals and only replaces those rivals if shown to be superior (Shepherd & Suddaby, 2017; Wacker, 1998). The 'rival theory' is that CBA is the gold standard for evaluating VFM in social investments. This research has demonstrated that it is not. The new theory is additive; it doesn't discard CBA but combines it with evaluative reasoning and mixed methods. Conclusions from the theoretical and empirical phases of the research suggest the new theory represents an advance on the use of CBA alone. But the theory is conservative in recognising and building on the strengths of CBA. The circumstances in which the new theory is superior to the rival theory have been clarified and remain open to further research. On these grounds the theory meets the conservation criterion.

The criterion of generalisability is most relevant to causal theories, where it is desirable that such theories have wide application (Shepherd & Suddaby, 2017) and are clear about the circumstances and evaluation questions to which they apply (Miller, 2010). It is worth noting, however, that the conceptual model does have wide application – specifically, to the evaluation of VFM in social policies and programs. The process model and case studies centred on international development programs. Nonetheless, the eight steps of the process model could be applied to any social policy or program, with contextually determined criteria. Through this research, the model has taken its first steps toward generalisability and has been found to be viable.

The criterion of fecundity highlights the value of a theory that generates new models that can be tested, leading to new knowledge (Patterson, 1986). The literature referring to this criterion has explanatory theories in mind. However, it is worth noting that the conceptual model developed in this research, contributing a definition of VFM and a set of requirements and processes for its evaluation, brings together established theories and practices to create a new model, which has been tested through case studies. Possibilities abound for further research on the model, as will be outlined later. The conceptual model meets the criterion of fecundity.

Good theory should meet the criterion of parsimony – it only contains necessary and sufficient complexity and assumptions (Patterson, 1986; Shepherd & Suddaby, 2017; Wacker, 1998). This theory incorporates CBA within a wider framework, adding further requirements and procedures, so is more complex than CBA. However, the findings suggest that the trade-off is necessary and helps to make evaluation and CBA more fit-for-purpose than either method alone when it comes to evaluating VFM in social policies and programs. The model comprises processes and principles that will be familiar to many evaluators and which provide a simple guiding structure for a complex and contextual practice. Therefore, the theory meets the criterion of parsimony.

Theories should be internally consistent – they should be logical and free from contradictions. This criterion is more relevant to theories that use mathematics or symbolic logic (Wacker, 1998) than to the conceptual model developed in this research. Nevertheless, the theory-building approach contained processes aimed at exposing and addressing any logical flaws, including exposing the theory to iterative peer review processes and case study analysis. Assessment of the theoretical propositions corroborated their conceptual quality and coherence.

It has been argued that theories should be empirically risky – they should do more than state the obvious, and have a credible risk of being wrong (Wacker, 1998). The critical risk in this research was that it had the temerity to question the well-established and widely-used method of CBA, and proposed augmentations to the method that had not been tried before. On the basis of case study analysis, it is concluded that the theory is not wrong. It has a long hill to climb, however, if it is to establish credentials as a strong alternative to CBA. This prospect is beyond the reach of this current research, and relies on others adopting, testing and refining the model in years to come.

The criterion of abstraction states that good theory should be independent from time and place, and capable of being abstracted to a higher-order set of propositions that have broad applicability. The theoretical component of the research is a middle-level abstraction theory, proposing requirements and a process for evaluating VFM in social investments. The empirical component of the research, based on two case studies, provides abstraction from the cases to the theory through analytic generalisation and replication (Yin, 2009). Further empirical evidence will only accumulate through continued application and adaptation of the model in diverse settings and circumstances.

Additional features of good theory are that it should be operational, that is, capable of being turned turned into a procedure, and practical, having utility to practitioners (Patterson, 1986). Both of these criteria have been met, with the two case studies providing concrete illustrations of its use in two settings. Finally, good theory is important – it is significant and relevant for evaluators in the real world (Coryn & Stufflebeam, 2014). The contributions of this research to theory and practice, as described in the beginning of this chapter, meet the thresholds for significance and relevance.

On the basis of these considerations, it is concluded that the model of VFM evaluation is a good theory, and represents a novel and significant contribution to the interim struggle of finding valid and fit-for-purpose ways to evaluate VFM in social policies and programs, to assist good resource allocation decisions that serve the public interest.

Research opportunities

While the two case studies corroborate the model's theoretical propositions, and provide replication of findings at a minimal level, much remains to be done to further research and refine the approach.

One such opportunity is to test the applicability of the model in a broader range of social policies and programs. Learning from a more diverse set of cases would serve further abstraction and refinement of the model. Aside from applying the model in different contexts, another opportunity for diversification would involve adapting the model to apply to a broader set of evaluative questions about VFM. The process model and case studies within this research addressed a principally summative and accountability-focused question about the overall VFM of the programs. Opportunities abound to apply the conceptual model to a diversity of evaluation purposes and questions, such as: informing incremental, stepwise resource allocation decisions to support innovation (e.g., in developmental evaluation); empowering communities to understand their own VFM (empowerment evaluation); and extending theory-based models of causality (such as realist evaluation) to incorporate theories of value.

A potential criticism of the research is that the case study analysis, and the evaluations examined within the case studies, were conducted by the same researcher who developed the model, introducing potential for real or perceived bias. To mitigate these risks, care was taken to make the analysis transparent and replicable. The systematic assessment against theoretical propositions, and the independent review of case study reports supported these objectives. Nevertheless, future research on the model by independent researchers would provide additional assurance.

Further research is also needed to more fully explore potential ways to integrate economic and evaluative thinking. This research focused principally on the use of CBA within the model. The inclusion of other economic methods, such as CEA and CUA, warrants further investigation. The economic literature provides guidance on choosing between these methods, but not on their specific application within the overarching evaluation model proposed here. There is also potential to consider the applicability of economic thinking more broadly, recognising that the field of economics, like that of evaluation, is diverse and expanding. What new insights might behavioural economics, for example, and post-GDP measures of wellbeing contribute to evaluations of programs and policies?

CBA itself continues to develop, and future enhancements to the method may affect the VFI model. Future developments in measuring wellbeing might improve or replace CBA, in time (Sunstein, 2018); but a higher-order system of explicit evaluative reasoning and complementary methods will continue to be needed to handle criteria and evidence that fall outside the schema of the method. Further research on VFI is needed to keep abreast of advances in evaluative and economic methods.

There are some challenges or limitations to the model itself that could be addressed through future research. One such challenge is the issue of consistency and comparability of VFM assessments. Criteria that are contextually determined for each evaluation may limit the extent to which different programs can be compared or ranked according to their VFM. Where such comparison is important, these challenges might be addressed through the specification of more general criteria and standards intended for comparative purposes, and/or through the inclusion of standardised economic and other metrics where these offer valid points of comparison.

Possible stakeholder bias toward indicators of efficiency, such as net present value, is a potential challenge to the adoption and use of the VFI model. The ability to synthesise program outcomes and costs, compare them to an alternative and summarise the result in a single indicator has beguiling simplicity. In contrast, using criteria and standards to make evaluative judgements from diverse evidence requires a more nuanced and discursive presentation of findings – though these findings can still be summarised within a simple statement such as 'this program meets the threshold for good VFM as defined in the criteria and standards'.

The output of a CBA may also have an appearance of values-neutrality or 'objectivity' that some stakeholders may perceive to be more robust than an evaluative judgement based on rubrics and mixed methods evidence. This perception is demonstrably incorrect. CBA, as we have seen, is as much a judgement-oriented practice as any other evaluation method, requiring multiple decisions about scope, perspective, time horizon, discount rate, and methods of monetary valuation that affect its results. Explicit evaluative reasoning guards against personal subjectivity in judgements, because criteria and standards, agreed at the outset of an evaluation, provide a transparent basis for gathering, analysing and synthesising relevant evidence and presenting judgements. Criteria and standards are inter-subjective – that is, an agreed social construct used by a group of people for an agreed purpose. As with other inter-subjective constructs, such as bank accounts, employment contracts and systems of democracy, evaluative rubrics are real, verifiable, and serve an important purpose. Further research is needed on mainstreaming the general logic of evaluation as a fundamental underpinning of sound policymaking and economic analysis.

Conclusions

At the commencement of any evaluation, an evaluator should invest time in understanding the program, its context, the primary evaluation users and stakeholders, and their information needs (Scriven, 2013; Patton, 2008). Evaluation questions will stem from these information needs (Davidson, 2005). If there is an evaluation question about VFM, the evaluator should proceed as follows (Figure 7).

Develop a clear and shared understanding of what needs the program is intended to meet and how it is intended to function. For example, a theory of change can facilitate clarity about these matters, providing a critical point of reference for structuring causal and evaluative claims.

Define context-specific criteria of VFM – the features or aspects of 'good resource use' that should be examined in the evaluation. A comprehensive evaluation of VFM should consider features of the resource use itself ('what did we put in?'), consequences of the resource use ('what did we get out?') and the basis upon which these two factors should be reconciled ('was it worth it?').

Develop standards – defined levels of performance – for each criterion. The standards should specify what the evidence should look like at different levels of performance (Davidson, 2014) and provide a clear and agreed basis for making judgements.

Determine what approach should be used for weighting and synthesis of the criteria, standards and evidence. Options include qualitative or numerical weighting and synthesis. In general, qualitative approaches (for example, rubrics) are feasible and support valid judgements in a wider range of circumstances than numerical weighting and synthesis (Scriven, 1991; Davidson, 2005). A numerical approach would be a viable option where there is an empirical basis to justify weights and sound mathematical logic is possible – for example, sufficiently few criteria to avoid swamping and mutually exclusive criteria to avoid interactions between criteria. CBA should only be used as the overarching approach to evaluative reasoning if: maximising aggregate welfare is the sole criterion of VFM; all relevant and material values can be fairly and accurately represented monetarily; qualitative distinctions not reflected in monetary valuations are unimportant; aggregation of values is appropriate; and only costs and consequences (not processes) matter.

Identify what forms of evidence are needed and will be credible to address the criteria and standards, and what design and and mix of methods should be used to collect and analyse the evidence (e.g., suitable approaches to causal inference). Method selection should be contextual and negotiated (Montrosse-Moorhead, Griffith, & Pokorny, 2014; Patton, 2018; Yarbrough et al., 2011).

In an evaluation of VFM, the inclusion of economic methods of evaluation should be given due consideration. Economic methods can supply part of the evidence toward an overall determination of VFM. Economic methods are candidates for inclusion where efficiency is a relevant criterion. CBA should be considered where aggregate welfare or net benefit is relevant, summation of monetisable costs and consequences is likely to give a sufficiently robust estimate of net benefit, and impacts can meaningfully be monetised and aggregated. CEA or CUA are candidates for inclusion where two or more alternatives are being compared, comparative outcome efficiency is relevant, and it is feasible to derive sufficiently robust estimates of impacts and/or utility. Disaggregated economic

analysis should be considered to investigate differences in the incidence of costs and consequences between subgroups.

Gather the needed evidence using the selected methods, following accepted good practices and ethical standards associated with each method (Coryn & Stufflebeam, 2014; Donaldson et al., 2015; Drummond et al., 2005; Schwandt, 2015; Tolich & Davidson, 2018).

Analyse each stream of evidence individually to identify findings that are relevant to the evaluation questions, criteria and standards.

Synthesise the evidence: bring the individual streams of evidence together to make judgements against the criteria and standards (Davidson, 2005; Fournier, 1995). Although this model is built on a technocratic foundation of criterial inference, the use of tacit, all-things-considered and/or deliberative approaches (Schwandt, 2015) can be used as supporting strategies for checking, challenging, validating and contextualising evaluative judgements.

Communicate the findings: present a clear and accurate account of evaluative judgements, with supporting evidence and reasoning.

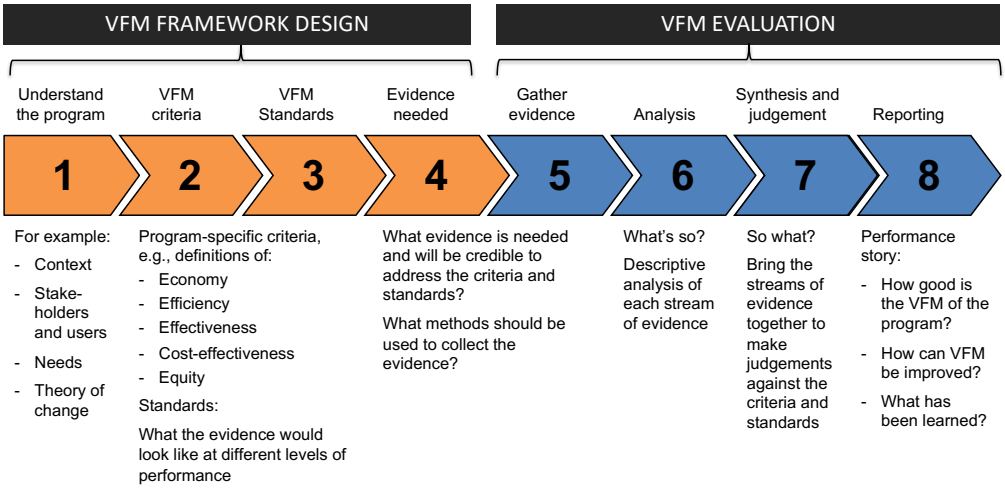

Figure 7: Overview of value for investment process model. Adapted from King & OPM (2018).

Throughout this process, stakeholders should be engaged in the evaluation to facilitate validity, understanding and use. VFM criteria, standards, evidence, and judgements should be considered from a range of perspectives such as beneficiaries, society and investor/donor.

These steps should be aligned with any other monitoring and evaluation activity being conducted in the same context, both for conceptual alignment and for practical coordination. Evidence used in the evaluation does not have to be new evidence; it might come from existing sources (McKegg et al., 2018).

This model provides an overarching logic to guide the evaluation of VFM, using a shared set of values agreed in advance with primary users and key stakeholders. Fidelity to this model should strive to ensure the evaluation: poses, and answers, an evaluative question about VFM; uses explicit evaluative reasoning; matches methods to context; and is guided by a suitable set of program evaluation standards.

Concluding remarks

Value for money, in essence, is good resource use. When the resources in question are invested in social change, a good resource allocation is one that serves the public interest. This is a matter of context and perspective, requiring not only robust evidence, but sound reasoning and engagement with values. The public interest is not well served by slavish adherence to any one method. There are no gold standards in evaluation, and CBA is no more a gold standard for evaluating VFM than the randomised controlled trial is for evaluating causality.

Irrespective of the approaches and tools used to guide evaluative reasoning, evaluation is "not first and foremost about methods, but is about making sense of evidence and creating a coherent, logical, and, ultimately, if successful, persuasive argument about what the evidence shows" (Patton, 2018, p. 18). Accordingly, this research also urges the use of evaluative thinking (Vo & Archibald, 2018) as a mainstay of VFM evaluation, recognising the importance of combining critical, creative, inferential and practical thinking (Patton, 2018) to determine and defend an appropriate combination of methods, tools, techniques, and theories (Vo, Schreiber & Martin, 2018) to evaluate VFM.

It has been said that "there is nothing as practical as a good theory".[3] This adage aptly describes the ambition of this research. This thesis contributes a practical theory for the evaluation of VFM. The theory combines good evaluation practice with economic methods of evaluation to support well-evidenced, well-reasoned judgements about VFM. It does this by joining the forces of explicit evaluative reasoning and mixed methods, including economic methods of evaluation where feasible and appropriate.

The VFI model builds on the strengths of CBA while compensating for its limitations. CBA enhances evaluation of VFM by providing a structure for systematic and rational analysis of costs and consequences, by valuing costs and consequences in commensurable units and reconciling them a single indicator, by taking opportunity cost into account through discounting, by facilitating transparency and robust thinking about uncertainty and risk, and by measuring an important value construct. However, while CBA addresses a single criterion, the VFI model can incorporate multiple and diverse criteria. While CBA bundles efficiency and

[3] This aphorism is commonly attributed to Kurt Lewin, though Bedeian (2016) traced it back further to Friedrich W. Dörpfeld's 1873 book, *Grundlinien einer Theorie des Lehrplans, zunächst der Volks- und Mittelschule.*

equity together, VFI unpacks them and makes trade-offs explicit. While CBA concentrates on the sameness of values, VFI facilitates deliberation on differences. While CBA is consequentialist, VFI can take account of process value. While CBA imposes a monetary valuing schema and excludes values that are too hard to monetise, VFI can incorporate a range of analytical approaches to value. For these reasons VFI is more versatile and more inclusive (of diverse stakeholders, values, and evidence) than CBA.

The significance of this research cuts to the core of both evaluation and economics. When it comes to investments in social change, good resource allocation is as much about social justice, equity and fairness, environmental and cultural sustainability, as it is about efficiency. Much as students of economics might question the adequacy of homo economicus (the rational man) as a model for human economic behaviour, and whether GDP is an adequate proxy for national wellbeing, CBA may be called into question as a defensible arbiter of resource allocation decisions. The assumptions and values embedded in each of these models serve to reinforce an individualistic, materialistic view of what is socially desirable. As growing inequalities in many nations fracture neoliberal political models, the time is ripe for a model of evaluation that recognises that good resource allocation decisions cannot be determined on the basis of economic efficiency alone.

Good policy isn't just about having evidence about 'what works', though such evidence is important. Good policy is equally about 'what matters' – and this means the merit, worth and significance of resource allocation decisions must be evaluated on the basis of values that are relevant to people who have a stake in the policy. Evaluating resource allocation requires a mix of technical and social, individual, collective and systemic thinking. A narrow focus on return on investment runs the risk of privileging some types of evidence and values over others: quantitative over qualitative; efficiency over equity; ends over means; aggregation over deliberation; consensus over difference; majority rule over balancing voices at the policy table; tangible over intangible values. CBA (and by extension, related methods such as CEA, CUA and SROI) has great capacity to inform good decisions, provided it is used to serve sound reasoning, not supplant it. A more generalised model of evaluative reasoning, capable of accommodating the paradigms and constructs of economic evaluation, as well as a broader set of axiologies from the diverse field of program evaluation, offers flexibility to support nuanced and context-sensitive judgements about VFM that respond to stakeholder needs and values. Ultimately, this thesis provides proof of concept for a practical theory to guide evaluation of VFM in social policies and programs.

References

Abend, G. (2008). The meaning of 'theory'. *Sociological theory, 26*(2), 173-199

Adler, M.D., & Posner, E.A. (2000). Implementing cost-benefit analysis when preferences are distorted. In M.D. Adler & E.A. Posner (Eds.), *Cost-Benefit Analysis: Legal, Economic, and Philosophical Perspectives* (pp. 269-312). Chicago, IL: University of Chicago Press.

Adler, M.D., & Posner, E.A. (2006). *New Foundations of Cost-Benefit Analysis.* Cambridge, Mass: Harvard University Press.

Ahmed, J.U. (2010). Documentary research method: New dimensions. *Indus Journal of Management & Social Sciences, 4*(1), 1-14.

Alkin, M. C. (Ed.). (2004). *Evaluation Roots: Tracing Theorists' Views and Influences*. Thousand Oaks, CA: Sage.

Alkin, M., & King, J. (2016). The historical development of evaluation use. *American Journal of Evaluation, 37*(4), 568-579.

Alvesson, M., & Karreman, D. (2007). Constructing mystery: Empirical matters in theory development. *Academy of Management Review, 32*(4), 1265-1281.

American Economic Association. (2013). What is economics? [web page]. Retrieved from: http://www.aeaweb.org/students/

American Economic Association. (2018). AEA Code of Professional Conduct, adopted April 20, 2018. [web page]. Retrieved from: https://www.aeaweb.org/about-aea/code-of-conduct

Andrews, M. (2010). *How Far Have Public Financial Management Reforms Come in Africa?* Faculty Research Working Paper. Cambridge, MA: Harvard Kennedy School.

Andrews, M. (2013). *The Limits of Institutional Reform in Development: Changing Rules for Realistic Solutions.* New York, NY: Cambridge University Press.

Andrews, M., Pritchett, L., & Woolcock, M. (2017). *Building State Capability.* Oxford, England: Oxford University Press.

ANZEA & Superu. (2015). *Evaluation standards for Aotearoa New Zealand.* Wellington, NZ: Aotearoa New Zealand Evaluation Association and Social Policy Evaluation and Research Unit.

Archibald, T., Sharrock, G., Buckley, J., & Young, S. (2018). Every practitioner a "knowledge worker": Promoting evaluative thinking to enhance learning and adaptive management in international development. In A.T. Vo & T. Archibald (Eds.). *Evaluative Thinking. New Directions for Evaluation, 158,* 73-91.

Argyrous, G. (2013). *A review of government cost-benefit analysis guidelines.* SSC/ANZSOG Occasional Paper. Canberra, Australia: Australia and New Zealand School of Government.

Arvidson, M., Lyon, F., McKay, S., & Moro, D. (2010). *The Ambitions and Challenges of SROI.* Working Paper 49. Birmingham, England: Third Sector Research Centre.

Australian Government Department of Finance. (2018). Value for money. [web page]. Retrieved from: https://www.finance.gov.au/procurement/procurement-policy-and-guidance/commonwealth-procurement-rules/march/value-for-money/

Australian Government Department of Foreign Affairs and Trade. (n.d.). Value for money principles. [web page]. Retrieved from: https://dfat.gov.au/aid/who-we-work-with/value-for-money-principles/Pages/value-for-money-principles.aspx

Backhouse, R.E. (2016, August). *The origins of New Welfare Economics. Preliminary draft written to provoke discussion at a workshop in welfare economics*, Hitotsubashi University, Japan.

Bamberger, J.M., Rugh, J., & Mabry, L.S. (2011). *RealWorld evaluation: Working under budget, time, data, and political constraints* (2nd ed). Newbury Park, CA: Sage.

Bamberger, J.M., (2012). *Introduction to Mixed Methods in Impact Evaluation.* Impact Evaluation Notes, No.3, August 2012. New York, NY: The Rockefeller Foundation.

Baron, J. (2004). Normative models of judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*, pp. 19-36. London: Blackwell.

Barr, J., & Christie, A. (2014). *Better value for money: an organising framework for management and measurement of VFM indicators.* Hove, England: Itad.

Baskarada, S. (2013). *Qualitative Case Study Guidelines.* Victoria: Australian Government Department of Defence. Joint and Operations Analysis Division.

Bayes, T., Price, Mr. (1763). "An Essay towards solving a Problem in the Doctrine of Chances". *Philosophical Transactions of the Royal Society of London*. 53 (0), 370–418.

Bedeian, A.G. (2016). A note on the aphorism "there is nothing as practical as a good theory". *Journal of Management History, 22*(2), 236-242.

Befani, B., & Mayne, J. (2014). Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation. *IDS Bulletin, 45*(6), 17-36.

Befani, B., & Stedman-Bryce, G. (2016). Process Tracing and Bayesian updating for impact evaluation: *Evaluation*, *23*(1), 42-60.

Bills, K. (2013). We're all economists now: Scarcity lessons for high school students. *Forbes*. [web page]. Retrieved from: https://www.forbes.com/sites/kurtbills/2013/04/18/were-all-economists-now-scarcity-lessons-for-high-school-students/#1ff440527cc0

Brent, R.J. (2006). *Applied Cost-Benefit Analysis.* Second Edition. Cheltenham, UK: Edward Elgar Publishing Ltd.

Boston, J., & Gill, D. (Eds.). (2017). *Social investment: A New Zealand policy experiment.* Wellington, New Zealand: Bridget Williams Books.

Boston, J. (2017). The intertemporal dimension. In J. Boston & D. Gill (Eds.), *Social investment: A New Zealand policy experiment* (pp. 91-120). Wellington, New Zealand: Bridget Williams Books.

Boxenbaum, E, & Rouleau, L. (2011). New knowledge products as bricolage: Metaphors and scripts in organizational theory. *Academy of Management Review, 36*(2), 272-296.

Carlton, D.W., & Perloff, J.M. (1994). *Modern industrial organization (2nd Ed.).* New York, NY: Harper Collins.

Chapple, S. (2013). Forward liability and welfare reform in New Zealand. *Policy Quarterly, 9*(2), 56-62.

Chapple, S. (2017). Corked wine in a cracked bottle. In J. Boston & D. Gill (Eds.), *Social investment: A New Zealand policy experiment* (pp. 355-379). Wellington, New Zealand: Bridget Williams Books.

Christie, C.A., & Alkin, M.C. (2012). An evaluation theory tree. In M.C. Alkin (Ed.), *Evaluation Roots* (2nd ed; pp. 12-66). Thousand Oaks, CA: Sage.

Coryn, C.L.S., & Hobson, K.A. (2011, Fall) *Eval 6000: Foundations of Evaluation.* Lecture presented at Western Michigan University, Kalamazoo, MI.

Coryn, C.L.S. & Stufflebeam, D.L. (2014). *Evaluation Theory, Models, & Applications.* San Francisco, CA: Jossey-Bass.

Creedy, J., & Passi, H. (2017). *Public sector discount rates: a comparison of alternative approaches*. Working Paper 17/02. Wellington, New Zealand: NZ Treasury.

Crotty, M. (1998). *The foundations of social research: meaning and perspectives in the research process*. London, England: Sage.

Damart, S., Roy, B. (2009). The uses of cost-benefit analysis in public transportation decision-making in France. *Transport Policy, 16*, 200-212.

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

Davidson, E.J. (2006). The RCTs-Only Doctrine: Brakes on the Acquisition of Knowledge? *Journal of Multidisciplinary Evaluation (6)*, ii-v.

Davidson, E.J. (2014). Evaluative Reasoning. *Methodological Briefs: Impact Evaluation 4*. Florence: UNICEF Office of Research.

Davis, K.E., Frank, R.G. (1992). Integrating costs and outcomes. *New Directions for Evaluation (54),* 69-84.

Deane, K.L., & Harré, N. (2016). Developing a thoughtful approach to evaluation: Values-driven guidelines for novice evaluators. *Evaluation Matters—He Take Tō Te Aromatawai*, *2,* 53-78.

Department for International Development. (2011). *DFID's Approach to Value for Money (VfM).* London, England: DFID, UK Government.

Destremau, K., & Wilson, P. (2017). Defining social investment, Kiwi-style. In J. Boston & D. Gill (Eds.), *Social investment: A New Zealand policy experiment* (pp. 32-79). Wellington, New Zealand: Bridget Williams Books.

Dickinson, P., Adams, J. (2017). Values in evaluation – the use of rubrics. *Evaluation and Program Planning, 65,* 113-116.

Dodgson, J.S., Spackman, M., Pearman, A., & Phillips, L.D. (2009). *Multi-criteria decision analysis: a manual.* London, United Kingdom: Department for Communities and Local Government.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddard, G. L. (2005). *Methods for the economic evaluation of health care programs.* Oxford, England: Oxford University Press.

Dumaine, F. (2012). When one must go: The Canadian experience with strategic review and judging program value. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New Directions for Evaluation, 133*, 65–75.

Eberle, W.D., & Hayden, F.G., (1991). *Critique of Contingent Valuation and Travel Cost Methods for Valuing Natural Resources and Ecosystems.* Economics Department Faculty Publications. Paper 13. Lincoln: University of Nebraska.

Ellerman, D. (2014). On a fallacy in the Kaldor-Hicks efficiency-equity analysis. *Constitutional Political Economy, 25*(2), 125-136.

Emmi, A., Ozlem, E., Maja, K., Ilan, R., Florian, S., (2011). *Value for Money: Current Approaches and Evolving Debates.* London, United Kingdom: London School of Economics. Retrieved from: http://bigpushforward.net/wp-content/uploads/2011/09/vfm-current-approaches-and-evolving-debates.pdf

Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism*. Princeton, NJ: Princeton University Press.

Fetterman, D. M., Kaftarian, S. J., & Wandersman, A. (Eds.). (2015). *Empowerment Evaluation: Knowledge and Tools for Self-Assessment, Evaluation Capacity Building, and Accountability (2nd ed.).* Thousand Oaks, CA: Sage.

Fleming, F. (2013). *Evaluation methods for assessing Value for Money*. Melbourne, Australia: Better Evaluation.

Flyvbjerg, B. (2011). Case study. In N. K. Denzin & Y. S. Lincoln (Eds). *The SAGE Handbook of Qualitative Research (4th Ed).* Thousand Oaks, CA: SAGE

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. In D. M. Fournier (Ed.), *Reasoning in Evaluation: Inferential Links and Leaps. New Directions for Evaluation, (58)*, 15-32.

Frank, R. (2000). Why is cost-benefit analysis so controversial? In M.D. Adler & E.A. Posner (Eds.), *Cost-Benefit Analysis: Legal, Economic, and Philosophical Perspectives* (pp. 77-94). Chicago, IL: University of Chicago Press.

Funnell, S.C., & Rogers, P.J. (2011). *Purposeful Program Theory: effective use of theories of change and logic models.* San Francisco, CA: Jossey-Bass.

Gargani, J. (2014, July). *New European Standard for Social Impact Measurement*. [web page]. Retrieved from: https://evalblog.com/2014/07/16/new-european-standard-for-social-impact-measurement/

Gargani, J. (2016, October). *Presidential keynote address* to the American Evaluation Association Conference, Atlanta, GA.

Gargani, J. (2017). The leap from ROI to SROI: farther than expected? *Evaluation and Program Planning, 64*, 116-126.

Gargani, J. (2018, October). *Impact measurement in the private sector*. Paper presented at the European Evaluation Society Conference, Thessaloniki, Greece.

Gehman, J., Glaser, V.L., Eisenhardt, K.M., Gioia, D., Langley, A., & Corley, K.G. (2018). Finding theory-method fit: a comparison of three

qualitative approaches to theory building. *Journal of Management Inquiry, 27*(3), 284-300

Giddens, A. (1999). *The third way: The renewal of social democracy*. Malden, MA: Blackwell.

Greene, J.C. (2002). With a splash of soda, please: Towards active engagement with difference. *Evaluation* 8(2), 249-258.

Greene, J.C. (2005). The generative potential of mixed methods inquiry. *Westminster Studies in Education, 28*(2), 207-211.

Greene, J.C. (2007). *Mixed methods in social inquiry.* San Francisco, CA: Jossey-Bass.

Greene, J.C. (2013). Reflections and ruminations. In D.M. Mertens & S. Hesse-Biber (Eds). *Mixed methods and credibility of evidence in evaluation. New Directions for Evaluation, 138,* 109-119.

Greenhalgh, T. (2018). *How to implement evidence-based healthcare.* Oxford, England: Wiley.

Griffith, J.C., & Montrosse-Moorhead, B. (2014). The value in validity. In J.C. Griffith & B. Montrosse-Moorhead (Eds.). *Revisiting truth, beauty, and justice: Evaluating with validity in the 21st century. New Directions for Evaluation, 142,* 17-30.

Goodwin, D., Sauni, P., Were, L. (2015). Cultural fit: an important criterion for effective interventions and evaluation work. *Evaluation Matters—He Take Tō Te Aromatawai*, *1,* 25-46.

Guba, E., & Lincoln, Y.S. (1981). *Effective evaluation: improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco, CA: Jossey-Bass.

Henry, G.T., Mark, M.M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation*, *24*, 293-314

Herman, P.M., Avery, D.J., Schemp, C.S., Walsh, M.E. (2009). Are cost-inclusive evaluations worth the effort? *Evaluation and Program Planning, 32*(1), 55-61

Hicks, J.R., (1939). The Foundations of Welfare Economics. *Economic Journal, 706*.

HM Treasury. (2006). *Value for Money Assessment Guidance.* Norwich, England: Her Majesty's Stationery Office.

HM Treasury. (2018). *The Green Book: Central government guidance on appraisal and evaluation.* Norwich, England: Her Majesty's Stationery Office.

Hogan, R.L. (2007). The historical development of program evaluation: Exploring the past and present. *Online Journal of Workforce Education and Development, II*(4), 1-14.

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

Independent Commission for Aid Impact (ICAI). (2011). *ICAI's Approach to Effectiveness and Value for Money.* *http://www.bond.org.uk/data/files/ICAIs-Approach-to-Effectiveness-and-VFM.pdf*

Independent Commission for Aid Impact. (2018). *DFID's approach to value for money in program and portfolio management: A performance review*. London, United Kingdom: ICAI.

Jackson, P. (2012). *Value for money and international development: deconstructing myths to promote a more constructive discussion*. Paris, France: OECD Development Co-operation Directorate.

Julnes, G. (2012a). Editor's Notes. In G. Julnes (Ed). *Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation*. *New Directions for Evaluation, 133*, pp 1-2.

Julnes, G. (2012b). Managing valuation. In G. Julnes (Ed). *Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation*. *New Directions for Evaluation, 133*, pp 3-15.

Julnes, G. (2012c). Promoting valuation in the public interest. In G. Julnes (Ed). *Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation*. *New Directions for Evaluation, 133*, pp 109-129.

Kaldor, N. (1939). Welfare Propositions in Economics. *Economic Journal, 549*.

Kaplan, A. (1964), The conduct of inquiry: methodology for behavioural science. San Francisco, CA: Chandler.

Keeney, R.L, and Raiffa, H. (1976) *Decisions with Multiple Objectives: Performances and Value Trade-Offs*. Wiley, New York.

King, J. (2015). Use of cost-benefit analysis in evaluation. Letter to the editor. *Evaluation Journal of Australasia, 15*(3), 37-41.

King, J. (2017). Using Economic Methods Evaluatively. *American Journal of Evaluation, 38*(1), 101-113.

King, J., Allan, S. (2018). Applying Evaluative Thinking to Value for Money: The Pakistan Sub-National Governance Programme. *Evaluation Matters—He Take Tō Te Aromatawai, 4,* pp. 207-235. Retrieved from: https://www.nzcer.org.nz/system/files/journals/evaluation-maters/downloads/Online_Articles_txt_King_FA_0.pdf

King, J., & OPM VfM Working Group. (2018). OPM's approach to assessing VfM: A guide. Oxford, England: Oxford Policy Management Ltd. Retrieved from: http://www.opml.co.uk/publications/opm's-approach-assessing-value-money

King, J., Guimaraes, L. (2016). Evaluating Value for Money in International Development: The Ligada Female Economic

Empowerment Program. *eVALUation Matters,* Third Quarter, 2016, pp. 58-69. Africa Development Bank. Retrieved from: http://idev.afdb.org/sites/default/files/documents/files/Evaluating%20value%20for%20money%20in%20international%20development-.pdf

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

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

Korzybski, A. (1933). *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. The International Non-Aristotelian Library Publication Company.

Kuhn, T.S., & Hacking, I. (2012). *The structure of scientific revolutions* (4th ed.). Chicago, IL: University of Chicago Press.

Lakatos, I. (1970) Falsification and the methodology of scientific research programmes, in Lakatos, I. and Musgrave, A. (Eds.) *Criticism and the Growth of Knowledge*. Cambridge, United Kingdom: Cambridge University Press.

Levin, H. (1987). Cost-Benefit and Cost-Effectiveness Analyses. In D.S. Cordray, H.S. Bloom, and R.J. Light (Eds). *Evaluation Practice in Review. New Directions for Program Evaluation, 34*.

Levin, H.M., McEwan, P.J. (2001). *Cost-Effectiveness Analysis*. 2nd Ed. Thousand Oaks: Sage.

Levy, H., & Sarnat, M. (1994). *Capital Investment & Financial Decisions (5th Ed).* Hertfordshire, United Kingdom: Prentice Hall.

Liddle, J., Wright, M., Koop, B. (2015). Cost-benefit analysis explained. *Evaluation Journal of Australasia*, *15*(2), 33-38.

Løkke, A., Sørensen, P.D. (2014). Theory testing using case studies. *The Electronic Journal of Business Research Methods, 12*(1), 66-74.

Mabry, L. (2010). Critical social theory evaluation: slaying the dragon. *The evaluation marketplace: exploring the evaluation industry. New Directions for Evaluation, 127,* 83-98.

Mankiw, G. (1999). *Macroeconomics* (4th ed.). New York: Worth.

Mark, M.M., Henry, G.T., Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey-Bass.

Martens, K.S.R. (2018a). How program evaluators use and learn to use rubrics to make evaluative reasoning explicit. *Evaluation and Program Planning, 69*, 25-32.

Martens, K.S.R. (2018b). Rubrics in program evaluation. *Evaluation Journal of Australasia, 18*(1), 21-44.

Mayne, J. (2008). Contribution Analysis: An approach to exploring cause and effect, *ILAC methodological brief.* Retrieved from: http://www.cgiar-ilac.org/files/ILAC_Brief16_Contribution_Analysis_0.pdf

McKegg, K., Oakden, J., Wehipeihana, N., King, J. (2018). *Evaluation Building Blocks: A Guide.* Wellington, New Zealand: Kinnect Group. Retrieved from: http://kinnect.co.nz/to/wp-content/uploads/EvaluationBuildingBlocks_A-Guide_FINAL_V1.pdf

Mertens, D.M., Hessie-Biber, S. (2013). Mixed methods and credibility of evidence in evaluation. In D.M. Mertens & S. Hessie-Biber (Eds.), *Mixed methods and credibility of evidence in evaluation. New Directions For Evaluation, 138*, 5-13.

Miles, M., Huberman, A., Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd Ed). Thousand Oaks, CA: SAGE.

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

Ministry of Foreign Affairs and Trade, New Zealand (2011, July). *Value for Money Guideline*. Retrieved from: https://www.mfat.govt.nz/assets/Aid-Prog-docs/Tools-and-guides/Value_for_Money_Guideline.pdf

Mintrom, M. (2017). Broader perspectives. In J. Boston & D. Gill (Eds.), *Social investment: A New Zealand policy experiment* (pp. 80-98). Wellington, New Zealand: Bridget Williams Books.

Montrosse-Moorhead, B., Griffith, J.C., & Pokorny, P. (2014). House with a view: Validity and evaluative argument. In J.C. Griffith & B. Montrosse-Moorhead (Eds.). *Revisiting truth, beauty, and justice: Evaluating with validity in the 21st century: New Directions for Evaluation, 142,* pp. 95-105.

Moon, K., & Blackman, D. (2014). A guide to understanding social science research for natural scientists. *Conservation Biology, 28*(5), 1167-1177.

Morel, N., Palier, B., Palme, J. (Eds.) (2012). *Towards a social investment welfare state?: Ideas, policies and challenges.* Bristol, England: Bristol University Press.

National Audit Office. (2013). *Value for money*. London: England: NAO. [web page]. Retrieved from: https://www.nao.org.uk/successful-commissioning/general-principles/value-for-money/

Newbold, P., Bos, T., (1990). *Introductory Business and Economic Forecasting* (2nd ed.). Cincinnati, Ohio: South-Western Publishing Co.

Nicholls, J., Lawlor, E., Neitzert, E., Goodspeed, T. (2012). *A Guide to Social Return on Investment*. January 2012. Haddington, England: The SROI Network.

Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation* Science, 10(53), 1-13.

Nunns, H., Peace, R., Witten, K. (2015). Evaluative reasoning in public-sector evaluation in Aotearoa New Zealand: How are we doing? *Evaluation Matters—He Take Tō Te Aromatawai*, *1,* 137-163

Nussbaum, M.C. (2000). The costs of tragedy: some moral limits of cost-benefit analysis. In M.D. Adler & E.A. Posner (Eds.), *Cost-Benefit Analysis: Legal, Economic, and Philosophical Perspectives* (pp. 169-200). Chicago, IL: University of Chicago Press.

Oakden, J. & King, J. (2018). Evaluation, in M. Tolich & C. Davidson (Eds.). *Social science research in New Zealand: An introduction* (3rd ed; pp. 180-193). Auckland, New Zealand: Auckland University Press.

OECD DAC. (2012). *DAC Guidelines and Reference Series. Quality Standards for Development Evaluation.* Evaluation Network of the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD). Retrieved from http://www.afdb.org/fileadmin/uploads/afdb/Documents/Evaluation-Reports/Quality%20Standards%20for%20Development%20Evaluation.pdf

Olson, E.E. & Eoyang, G.H. (2001) *Facilitating Organization Change: Lessons from Complexity*. Chichester, England: Wiley.

Patel, M. (2013). African Evaluation Guidelines. *African Evaluation Journal, 1*(1), 1-5.

Patterson, C.H. (1986). *Theories of counselling and psychotherapy* (4th ed.). Philadelphia: Harper & Row.

Patton, M.Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M.Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York, NY: Guildford.

Patton, M.Q. (2017a). *Evaluation Flash Cards: Embedding evaluative thinking in organizational culture.* St Paul, MN: Otto Bremer Trust.

Patton, M. Q. (2017b). *Principles-focused evaluation: The guide*. New York: Guilford Publications.

Patton, M.Q. (2018). A historical perspective on the evolution of evaluative thinking. In A.T. Vo & T. Archibald (Eds.). *Evaluative Thinking. New Directions for Evaluation, 158,* 11-28.

Pawson, R. (2013). *The Science of Evaluation: A Realist Manifesto.* Thousand Oaks, CA: Sage.

Persaud, N. (2007). Is cost analysis underutilized in decision making? *Journal of Multidisciplinary Evaluation (JMDE:2).*

Pinkerton, S.D., Johnson-Masotti, A.P., Derse, A., Layde, P.M., (2002). Ethical issues in cost-effectiveness analysis. *Evaluation and Program Planning 25*, 71-83.

Popper, S.K. (1957). Philosophy of science: a personal report. In C.A. Mace (Ed.), *British Philosophy in Mid-Century.* (pp. 182-183). London, England: George Allen & Unwin.

Quine, W.V., & Ullian, J.S. (1980). Hypothesis. In E.D. Klempe, R. Hollinger, & A.D. Kline (Eds.), *Introductory Readings in the Philosophy of Science.* Buffalo, NY: Prometheus.

Rawls, John (1971). *A Theory of Justice*. Cambridge, Massachusetts: Belknap Press.

Renard, R., Lister, S. (2015). *Measuring and Reporting on Value for Money: A conceptual framework for MDBs.* Oxford, England: Mokoro Ltd.

Rogers, P. (2014). Theory of Change. *Methodological Briefs: Impact Evaluation 2*. Florence, Italy: UNICEF Office of Research.

Roorda, M. (2019). *Developing defensible criteria for Australian and New Zealand public sector evaluations*. (Draft Doctoral dissertation). Melbourne, Australia: University of Melbourne.

Samuelson, P.A., (1948). *Economics*. New York, NY: McGraw-Hill.

Sen, A. (2000). The discipline of cost-benefit analysis. In M.D. Adler & E.A. Posner (Eds.), *Cost-Benefit Analysis: Legal, Economic, and Philosophical Perspectives* (pp. 95-116). Chicago, IL: University of Chicago Press.

Shapiro, S.A., & Schroeder, C. (2008). Beyond cost-benefit analysis: a pragmatic reorientation. *Harvard Environmental Law Review, 32,* 433-502.

Schiere, R. (2016). What is new in value for money? *eVALUation Matters,* Third Quarter, 2016, pp. 22-33. Africa Development Bank.

Schwandt, T. (2015). *Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice.* Redwood City: Stanford University Press.

Schwandt, T. (2018). Evaluative thinking as a collaborative social practice: The case of boundary judgement making. In A.T. Vo & T. Archibald (Eds.), *Evaluative Thinking. New Directions for Evaluation, 158,* 125-137.

Scott, G. (2017). Governance, public policy and public management. In J. Boston & D. Gill (Eds.), *Social investment: A New Zealand policy experiment* (pp. 477-498). Wellington, New Zealand: Bridget Williams Books.

Scriven, M. (1980). *The logic of evaluation.* Thousand Oaks, CA: Sage.

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

Scriven, M. (1993). The Nature of Evaluation. *New Directions for Program Evaluation, 58,* 5-48.

Scriven, M. (1994). The Final Synthesis. *Evaluation Practice, 15*(3), 367-382.

Scriven, M. (1995). The Logic of Evaluation and Evaluation Practice. *New Directions For Evaluation, 68*, 49-70.

Scriven, M. (2007, June). *The Logic of Evaluation.* Paper presented at the Ontario Society for the Study of Argumentation Conference. Windsor, Canada.

Scriven, M. (2008). A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research. *Journal of Multidisciplinary Evaluation, 5*.

Scriven, M. (2012). The logic of valuing. In G. Julnes (Ed.), *Promoting valuation in the public interest: Informing policies for judging value in evaluation. New directions for evaluation, 133*, 17-28.

Scriven, M. (2013, March 22). *Key evaluation checklist (KEC).* Retrieved from: http://www.michaelscriven.info/images/KEC_3.22.2013.pdf

Scriven, M. (2013). Reconstructing the foundations of evaluation: Practical philosophy of science vs. Positivist philosophy of science. Paper presented to the Australasian Evaluation Society International Conference. Brisbane, Australia.

Shadish, W.R., Cook, T.D., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York, NY: Houghton Mifflin.

Shepherd, D.A., & Suddaby, R. (2017). Theory building: A review and integration. *Journal of Management, 43*(1), 59-86.

Stake, R.E. (1978). *The case study method in social inquiry. Educational Researcher,* 5-8.

Stake, R., Migotsky, R.D., Cisneros, E.J., Depaul, G., Dunbar, C., Farmer, R., Feldovich, J., Johnson, E., Williams, B., Zurita, M., Chaves, I. (1997). The evolving syntheses of program value. *Evaluation Practice, 18* (2), 89-103.

Stake, R.E, & Schwandt, T.A. (2006). On discerning quality in evaluation. In I. Shaw, J.C. Greene, & M.M. Mark (Eds). *The SAGE Handbook of Evaluation: Policies, programs and practices.* London: SAGE.

Stufflebeam, D.L. (2001a). Evaluation Models. *New Directions for Evaluation, 89,* 7-98.

Stufflebeam, D.L. (2001b, March). *Evaluation Values and Criteria Checklist.* Retrieved from: www.wmich.edu/evalctr/checklists

Sunstein, C.R. (2000). Cognition and cost-benefit analysis. In M.D. Adler & E.A. Posner (Eds.), *Cost-Benefit Analysis: Legal, Economic, and Philosophical Perspectives* (pp. 223-268). Chicago, IL: University of Chicago Press.

Sunstein, C. R. (2018). *The Cost-Benefit Revolution.* Cambridge, MA: MIT Press.

Talvitie, A. (2018). Jules Dupuit and benefit-cost analysis: making past to be the present. *Transport Policy, 70,* 14-21.

Tolich, M., Davidson, C. (Eds.). (2018). *Social science research in New Zealand: An introduction* (3rd ed.). Auckland, New Zealand: Auckland University Press.

Tolich, M., Davidson, C. (2018). Science and social science. in M. Tolich & C. Davidson (Eds.). *Social science research in New Zealand: An introduction* (3rd Ed.). (pp. 34-47). Auckland, New Zealand: Auckland University Press.

Toulmin, S.E. (1964). *The Uses of Argument.* New York, NY: Cambridge University Press.

Toulmin, S.E. (1972). *Human Understanding.* Princeton, NJ: Princeton University Press.

United Nations Evaluation Group. (2016). *Norms and Standards for Evaluation.* New York, NY: UNEG.

United States Government Accountability Office. (2007). *Government Auditing Standards, July 2007 Revision.* By the Comptroller General of the United States. Washington DC: USGAO.

United States Government Accountability Office. (1990). *Case study evaluations*. By the Comptroller General of the United States. Washington, DC: USGAO.

Van De Ven, A.H., & Johnson, P.E. (2006). Knowledge for theory and practice. *Academy of Management Review, 31,* 802:821.

Van Evera, S. (1997). *Guide to Methods for Students of Political Science*. Ithaca NY: Cornell University Press

Vo, A.T., & Archibald, T. (2018). New directions for evaluative thinking. In A.T. Vo & T. Archibald (Eds.). *Evaluative Thinking. New Directions for Evaluation, 158*, 139-147.

Vo, A.T., Schreiber, J.S., & Martin, A. (2018). Toward a conceptual understanding of evaluative thinking. In A.T. Vo & T. Archibald (Eds.). *Evaluative Thinking. New Directions for Evaluation, 158*, 29-47.

Von Neumann, J., and Morgenstern, O. (1947) *Theory of Games and Economic Behaviour*. *Second Edition*. Princeton, NJ: Princeton University Press.

Wacker, J.G. (1998). A definition of theory: research questions for different theory-building research methods in operations management. *Journal of Operations Management, 16*, 361-385.

Walsh, B., & King, J. (2017). Realist economic evaluation. Paper presented at the International Conference for Realist Research, Evaluation and Synthesis, Brisbane, Australia.

Wehipeihana, N., & McKegg, K. (2018). Values and culture in evaluative thinking: Insights from Aotearoa New Zealand. In A.T. Vo & T. Archibald (Eds.). *Evaluative Thinking. New Directions for Evaluation, 158,* 93-107.

Wehipeihana, N., Oakden, J., King, J., McKegg, K. (2018, May). Rubrics – A GPS for Evaluation: Our learning from 10 years' experience. Paper presented at the Canadian Evaluation Society Conference, Vancouver, Canada.

World Bank. (2016). *Value for money: Achieving VfM in investment projects financed by the World Bank.* Washington DC: Author.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., and Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage

Yates, B.T. (1996). *Analyzing costs, procedures, processes, and outcomes in human services.* Applied Social Research Methods Series Volume 42. Thousand Oaks, CA: Sage.

Yates, B.T., (2009). Cost-inclusive evaluation: A banquet of approaches for including costs, benefits, and cost-effectiveness and cost-benefit analyses in your next evaluation. *Evaluation and Program Planning, 32*(1), 52-54.

Yates, B.T. (2012). Step arounds for common pitfalls when valuing resources used versus resources produced. *Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation. New Directions for Evaluation*, *133*, 3-52

Yin, R. K. (2009). *Case study research: Design and methods* (4th Ed.). Thousand Oaks, CA: Sage.

Appendix: VFM publications subsidiary to thesis

King, J., Allan, S. (2018). Applying Evaluative Thinking to Value for Money: The Pakistan Sub-National Governance Programme. *Evaluation Matters—He Take Tō Te Aromatawai, 4,* pp. 207-235. Retrieved from: https://www.nzcer.org.nz/system/files/journals/evaluation-maters/downloads/Online_Articles_txt_King_FA_0.pdf

King, J., Guimaraes, L. (2016). Evaluating Value for Money in International Development: The Ligada Female Economic Empowerment Program. *eVALUation Matters,* Third Quarter, 2016, pp. 58-69. Africa Development Bank. Retrieved from: http://idev.afdb.org/sites/default/files/documents/files/Evaluating%20value%20for%20money%20in%20international%20development-.pdf

King, J., & OPM VfM Working Group. (2018). *OPM's approach to assessing VfM: A guide*. Oxford, England: Oxford Policy Management Ltd. Retrieved from: http://www.opml.co.uk/publications/opm's-approach-assessing-value-money