

**Mediating Social Change:
Building Adaptive Learning Systems through Developmental Evaluation**

Barbara Szijarto

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the
Ph.D. in Education

May 2019

Supervisor: Dr. J. Bradley Cousins

Committee Members: Dr. Peter Milley, Dr. John Sylvestre, Dr. David Trumpower

© **Barbara Szijarto, Ottawa, Canada, 2019**

Abstract

Complex social problems are at the forefront of our awareness. We are witnessing intensifying political, social and environmental challenges and waning confidence in our ability to engineer solutions. We are also seeing a proliferation of large scale, multi-agency interventions that seek change at the level of systems, and through which actors pursue adaptive learning as a means to develop effective solutions. Proponents assert that the prediction and control on which conventional program design and evaluation are based are not available under complex conditions. They propose instead that learning through experience in a program's own context can create more responsive, impactful and sustainable interventions. These ideas offer a potentially transformative opportunity. However, they need to be complemented with a better understanding of implementation - the 'ways of doing things' that bring them to life.

This study focused on developmental evaluation as an example of an adaptive learning (AL) approach for the development of innovative social interventions. The study was informed by 'sensemaking' theories and research in organizational learning, knowledge mobilization and program evaluation. Through an exploratory lens and a mixed methods design, this study sheds light on the role of specialized intermediaries in an AL process; how the role is performed in practice; and what this implies for adaptive learning in the domain of social interventions. The study documents how an intermediary can help actors navigate recognized challenges of developing interventions under complex and dynamic conditions. The findings have implications for how an AL process is understood and implemented. They provide an empirical contribution to an emerging field of study on the design of AL systems, to support future research and real-world practice as AL approaches become mainstream.

Dedication

To my husband, Rick, and my daughters, Sydney and Gillian, for your enthusiasm and countless willing sacrifices. To my parents and my sisters, for your resilience and strength which inspire me. And to Brad Cousins. You share your humour, confidence, and compassion with everyone around you. You challenge us to do better and are a truly generous human being. Thank you.

Acknowledgements

I would like to acknowledge my thesis committee members, Peter Milley, John Sylvestre and David Trumpower, who contributed their time and insight, and asked very good questions.

I am also grateful to the participants of this study, who shared their experiences and knowledge with me.

Funding for this research was provided by the Social Sciences and Humanities Research Council, the University of Ottawa, and the O'Meara Family Scholarship Fund.

Contents

Abstract	ii
Dedication	iii
Acknowledgements	iii
Contents	iv
List of Figures	vii
List of Tables	viii
Part I. Conceptual Foundation	1
Chapter 1. Introduction	1
Statement of the Problem.....	2
Study Purpose, Focus and Research Questions	3
Thesis Overview	4
Chapter 2 Complexity Thinking and the Potential of Intermediaries	6
Scope of the Literature Review.....	6
The Spread of Complexity Thinking	7
Adaptive Learning	15
Intermediaries in a Learning Process	28
Conclusion	35
Chapter 3 Conceptual Framework	37
Working Definitions	37
Conceptual Framework	40
Guiding Assumptions.....	42
Part II. Methodology	45
Chapter 4 Study Design, Methods and Procedures.....	45

Mediating Social Change

Study Design	45
Strand 1: Multi-Case Study.....	47
Strand 2: Group Concept Mapping	52
Integration of the Strands.....	63
Quality and Validity.....	64
Part III. Strand Level Findings.....	66
Chapter 5 Within-case Findings.....	66
Common Features of the Cases.....	66
Case 1. A Better System for At-Risk Youth	67
Case 2. The ‘Game Changer’: Higher Education for Low-Income Children	76
Case 3. Transforming Cities through the Arts	88
Case 4. A System for Agricultural Innovation.....	96
Closing Note	108
Chapter 6 Cross Case Analysis	109
Making Space for Adaptive Inquiry	109
Learning Forward.....	123
Performing Roles	130
Chapter 7 Group Concept Mapping.....	149
Overview of Strand 2	149
Results.....	149
Interpretation of the Results.....	154
Summary	163
Part IV. Integration	165
Chapter 8 Integration of the Strands: Roles and Activities.....	165
Step 1. Roles	165

Step 2. Domains of Activities	173
Chapter 9 Implications for Adaptive Learning	186
Implications for an Adaptive Learning Process.....	186
Integrating Three Conditions for Adaptive Learning: A Proposed Model.....	195
Implications for Practice	200
Chapter 10 Conclusion.....	208
A Summary of the Study’s Contributions.....	209
Limitations of the Study.....	212
An Agenda for Future Research	214
Closing	216
References.....	217
Appendix A. Content Analysis Template, Strand 1.....	245
Appendix B. Interview Guide Template.....	246
Appendix C. Letter of Informed Consent: Interviews	249
Appendix D. Letter of Informed Consent: GCM Sorting and Rating.....	251
Appendix E. Letter of Informed Consent: GCM Expert Panel.....	253
Appendix F. Case Trajectory Diagram, Case 2	255
Appendix G. GCM Statement List by Cluster	256
Appendix H. Focus Group Guide	260
Appendix I. Full Case Study Reports	262

List of Figures

Figure 2.1. Focus of the foundational literature review.....	6
Figure 3.1. Adaptive learning in a social learning system.....	38
Figure 3.2. Mediation as an intervention in sensemaking within a social learning system.....	39
Figure 3.3. Conceptual framework for collaborative sensemaking in a social learning system...	41
Figure 4.1. Study design	46
Figure 4.2. Steps in the multi-case study data collection and analysis.	50
Figure 4.3. Concept mapping steps.....	54
Figure 4.4. Implementation of the two strands over time	62
Figure 6.1. Primary roles.	132
Figure 6.2. Continuum of mediation strategies.....	133
Figure 6.3. Three relational targets.	141
Figure 7.1. Point map output of the MDS analysis.....	150
Figure 7.2. Concept map of DE activities.....	151
Figure 7.3. Bridging statement locations.	152
Figure 7.4. Figure 7.4. Spanning diagram illustrating relationships of Statement #26 “Be a good strategy coach”.....	152
Figure 7.5. Cluster-level ratings.....	154
Figure 7.6. Clusters in conversation: 3 ‘regions’ of activity.....	157
Figure 7.7. Clusters in conversation, metaphor of ‘space’.	158
Figure 7.8. A proposed process for statement #32.....	161
Figure 8.1. Size and average rating of cross-case role categories based on GCM statement coding.....	166
Figure 8.2. Distribution of statements relating to subthemes of the technical role.	167
Figure 8.3. Distribution of statements by subtheme in the guide role.	168
Figure 8.4. Distribution of statements in the learner role.	170
Figure 8.5. Location of four roles in terms of directiveness and domain focus.	171
Figure 8.6. Overview of NVivo coding against the GCM framework.	174
Figure 9.1. A model of proposed pathways.	198
Figure 9.2. Three conditions to create space for AL.....	200

List of Tables

Table 5.1. Description of the cases	67
Table 6.1. Factors affecting engagement	111
Table 6.2. Strategies for making space.	115
Table 6.3. Efforts to gain traction.	124
Table 6.4. Reported outcomes.	129
Table 6.5. Evaluator relationship to program actors	142
Table 6.6. Relationships among program actors.	146
Table 7.1. Names and description of the clusters.	155
Table 8.1. Primary role categories	171
Table 8.2. Domains of activity.	173
Table 8.3. Components of alignment.	175
Table 8.4. Examples of alignment of GCM items with CAE principles	181
Table 8.5. GCM activities and cross-case extensions for Domain 3	182
Table 9.1. Examples of divergent and convergent thinking.	197

Part I. Conceptual Foundation

In the first part of this thesis, I situate the study and describe its foundations. In Chapter 1, I introduce the background context, the problem, and my research questions. In Chapter 2, I discuss ideas from existing literature that have informed the study. In the third chapter, I describe the study's conceptual framework, including definitions of key constructs and epistemological assumptions guiding my approach.

Chapter 1. Introduction

Social programs are meant to improve lives. Their purpose is to improve social conditions and alleviate suffering, referred to generally as social betterment (Mark, Henry, & Julnes, 2000). To achieve social betterment, actors involved in the delivery of social programs are encouraged to commit to evidence-based approaches, often by adopting a program model that has been developed and tested elsewhere. If this is not available, actors are urged to integrate established best practice or formal theory into operational plans and then to adhere to these plans throughout implementation. This approach to program design and implementation is driven by a legitimate desire of stakeholders for accountability and effective use of public funds, and is sometimes reinforced by program evaluation (Gerrits & Verweij, 2015).

Alternative approaches have recently become more common, based on an understanding of social interventions as complex. Proponents assert that the prediction and control on which conventional design and evaluation are based are not available to complex interventions. Rather, social programs are subject to emergent influences, interactions and effects. These make their trajectories challenging to predict in advance and their outcomes more difficult to interpret. Moreover, change and surprise are not only inevitable but desirable; ongoing adaptation of a program within its own context is essential. Although the threads of this argument run back at least 40 years (for example Weiss & Rein, 1969 with response from Campbell, 1970), momentum has developed rapidly in the last decade around a call for complexity-informed approaches.

In particular, adaptive learning (AL) approaches have gained traction in the past decade. This is reflected, for example, in “place-based” interventions (e.g., Bellefontaine & Wisener, 2011), and in developmental evaluation (e.g., Patton, McKegg, & Wehipeihana, 2016). Developmental evaluation (DE) aims to support interventions through an AL cycle. It focuses on enabling data-informed reflection on results and attending to unexpected events as learning opportunities. Other iterative program development approaches include interaction design (Engeström, 2006) and various “facilitated sensemaking” approaches (Lanham et al., 2013). These are said to be more consistent with complexity, able to cultivate greater sensitivity to local contexts and more synergistic and sustainable interventions. They are believed to offer a richer understanding of *how* programs are working (or not working) to support ongoing development and adaptation to change.

Statement of the Problem

These approaches represent important innovations in program design and evaluation, and they warrant attention. Rationale for their use draws heavily on complexity theories and the implications of these theories for social interventions. These arguments will be reviewed in the next chapter, along with what the literature tells us about challenges that can impair learning under complex conditions.

This last presents something of a paradox. The same conditions that make adaptive learning so important also challenge the learning process itself. Moreover, methods embedded in some approaches, such as timely feedback based on ‘safe-fail’ experiments (Snowden, 2011), remain conceptually and instrumentally ambiguous and challenging to operationalize. The involvement of a specialized intermediary is often proposed to help overcome such challenges. They are typically a researcher or evaluator who provides facilitative and technical support. For example, intermediaries are proposed to help make a learning process more intentional, enable balanced participation of diverse perspectives, and improve quality of feedback mechanisms. Although a potentially vital support to the process, the intermediary’s role, functions and the nature of their influence are not well documented. There is little empirical research about how an intermediary supports learning within a group and the translation of that learning to social programs.

Study Purpose, Focus and Research Questions

This study's purpose is to contribute to a better understanding of the role of intermediaries in adaptive learning approaches. The study was designed to be exploratory, with an aim to shed light on the intermediary role, how it is performed in practice, and what this implies for adaptive learning in the domain of social interventions.

I review relevant literature on how such roles are conceptualized in the next chapter, however in short, intermediaries are usually described as third parties and sometimes called facilitators, brokers or mediators. The term “facilitator” is common and implies an actor who enables others to learn or generate knowledge. This term implies neutrality (Kemmis & McTaggart, 2008a). Like “brokering”, facilitation is a passive term that may obscure a more influential function. I chose to adopt the term “mediation” in this study, because it helps to maintain awareness of the potential for the function to be located inside the process and to have direct influence.

The study centred on developmental evaluation as an approach in which an intermediary (the evaluator) actively seeks to foster adaptive learning through a complexity-informed, situated approach. Introduced in the mid 1990's, DE has become a prominent evaluation approach particularly in the last 10 years (Milley, Szijarto, Svensson, & Cousins, 2018). Patton (1994) first described DE and set it apart from formative and summative evaluation based on its purpose. In essence, the core purpose is to support development, not refinement or implementation fidelity (formative evaluation), and not a summative judgement of whole-program value, merit or worth (summative evaluation). In DE, “the evaluator is part of a team whose members collaborate to conceptualize, design, and test new approaches” in a process of ongoing adaptation and intentional development (p. 317). Proposed applications for DE include support to the development of innovative approaches and systems change interventions, supporting the adaptation of practices developed elsewhere to new contexts, and rapidly exploring solutions in situations of crisis (Patton, 2011). It is more prospective in orientation than conventional forms of evaluation. DE evaluators explicitly seek to enable the use of empirical data and critical thinking in frequent cycles, working in collaboration with program actors in a process of adaptive learning (Patton, 2016a). A complexity perspective is meant to inform all aspects of the evaluation. This includes informing how actors make sense of the problems they are targeting. A

complexity perspective carries assumptions of limited control and predictability, a need to change and adapt, and a need to attend to effects that may be unexpected in degree and in kind (Patton, 2016b, see a review of complexity ideas in Chapter 2). These features make DE a useful and rich exemplar of mediated approaches to adaptive learning. The study of DE seemed likely to produce rich insights on the process as applied in social interventions.

Research questions. I focused this study on two questions, with the first extended via targeted sub-questions:

1. In the context of developmental evaluation, what role does mediation play in place-based learning in social program design and implementation?
 - What are the most important activities or points of intervention for a mediator? Why?
 - How is mediation enacted (e.g., activities, tools, timing)? How does it unfold? Why?
2. What are the implications for an adaptive learning process?

These questions relate to the proposed importance of intermediaries to an adaptive learning process, and particularly their importance in the design and development of complex interventions, as discussed in the next chapter. As noted above, applications of DE include early-stage support for innovative and systems change approaches and for programs seeking to make major changes in response to changing conditions. These are interventions that lack a predefined model or blueprint matching the context or proposed program theory, and for which stakeholders elect to adopt an adaptive approach. The bounding of the study through its focus on DE is discussed further in Chapter 4, where I describe the study's methodology and related methods in more detail, including decisions related to casing and data sources.

Thesis Overview

This thesis is organized in ten chapters. In the next chapter, I review literature relevant to adaptive learning and intermediary roles across three fields: program evaluation, organizational learning and knowledge mobilization. Chapter 3 presents the study's conceptual framework along with working definitions of key constructs such as *adaptive learning* and *mediation*. I describe the methodology in Chapter 4. I used a mixed methods design for the study with two

parallel components: a multiple case study and group concept mapping. These components were later integrated to produce overall findings. In Chapters 5 and 6, I report on the case-based component of the study: results of four individual case studies followed by a cross-case analysis. I then report on the concept mapping component in Chapter 7. Findings from the integration of the two components are presented in Chapters 8 and 9; in Chapter 8, findings respond to my research question and sub-questions on the role of mediation. Implications for adaptive learning are addressed and discussed in Chapter 9. The final chapter closes the thesis with an outline of the study's contributions, limitations, and important directions for future research.

Innovative approaches to social intervention design and development based on adaptive learning are coming into more widespread use. Advocates propose that learning through experience in a program's own context can create more responsive, impactful and sustainable interventions. Their work offers a potentially transformative opportunity to serve better our diverse, dynamic communities. However, these need to be complemented with a better understanding of implementation - the 'ways of doing things' that bring these ideas to life (Howard, Flores, & Hambleton, 2015). As observed in the next chapter, empirical research is still scarce in this area. Using DE as an exemplar approach, this thesis documents how a specialized intermediary can help actors navigate challenges of developing interventions under complex and dynamic conditions. It identifies an array of strategies, functions and roles for the intermediary, ways in which these are performed and why, and some risks to which actors might attend. The findings have implications for how an adaptive learning process is understood and provide an important empirical contribution to an emerging field of study on the design of adaptive learning systems.

Chapter 2 Complexity Thinking and the Potential of Intermediaries

In this chapter, I review existing research on ideas central to the study. The review's purpose was to lay a foundation for the conceptual framework and provide sensitizing concepts for use in collecting and analysing data (Blumer, 1954). I also sought to assess the extent of empirical inquiry in relevant areas of research and to identify gaps in understanding.

Scope of the Literature Review

Adaptive learning mediation is a new area of study, particularly with respect to developing social interventions. Relevant lines of research are available but located in different academic specialties and are not well integrated. At the outset of this study, I reviewed research in knowledge mobilization (KMb), program evaluation and organizational learning. These fields are relevant because they offer different perspectives on how organized groups 'learn'. Broadly, the organizational learning literature concentrates at the level of the organization or organizational subunit: how the entity or its teams learn and adapt to change. Evaluation's focus is on the social intervention. This leads scholars in this field to look both inside and across organizational boundaries, and at macro influences of policy and power. KMb looks across sectors (e.g., academia, policy, service delivery) with concern for how evidence-based innovations are spread across contexts. I focus this chapter on themes shared by these domains that are especially pertinent to my study, depicted in the overlapping area of Figure 2.1. These are: the implications of complexity for social interventions, characteristics of adaptive learning, and proposed roles for intermediaries.

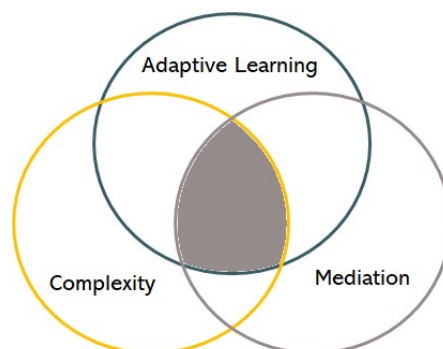


Figure 2.1. Focus of the foundational literature review

Consistent with an exploratory study, more focused themes emerged as the study progressed. I sought out and reviewed other literature to help me develop these themes and to situate findings. For example, preliminary findings from cross-case analysis led me to locate and discuss a framework from the conflict mediation literature, and to read more deeply on the concept of role in evaluation. This literature is discussed primarily with the study's findings in Chapters 6-9 as it was not part of the foundational work for the study.

The Spread of Complexity Thinking

Ideas from theories of complexity feature in abundance now in the literature on evaluation (e.g., Patton et al., 2016), knowledge mobilization (e.g. Ferlie, Crilly, Jashapara, & Peckham, 2012) and organizational learning (e.g., Snowden & Boone, 2007). Much of this work is conceptual, although empirical studies are emerging, primarily single case (e.g., Edwards, Rowan, Marck, & Grinspun, 2011) and multiple case studies (e.g., Berta, Virani, Bajnok, Edwards, & Rowan, 2014). They reflect a dramatic upward trend in interest in complexity over the past 20 years which has contributed to a shift in thinking about intervention development.

Key ideas. Complexity theories are diverse and still developing. Some suggest that complexity has been observed by social scientists for a long time and is latent in many social theories. For example, agency effects are likened to Strauss' theory of negotiated order (Byrne & Callaghan, 2014). Similarly, practitioners working in social interventions are said to have an intuitive understanding of complexity based on their experience, but until recently have "lacked a vocabulary and methodological toolkit" to deal with it (Gerrits & Verweij, 2015, p. 484). Byrne (2005) writes that social scientists "are involved in a project of discovering complexity" (p. 101). He offers a working definition of complexity theory, which touches on a key reason for complexity theory to be of interest to program developers and evaluators, namely that it relates to understanding trajectory and patterns of change:

...[t]he interdisciplinary understanding of reality as composed of complex open systems with emergent properties and transformational potential. A crucial corollary of complexity theory is that knowledge is inherently local rather than universal. Complexity science... is concerned with the description and explanation of change and one of its most powerful terms... is trajectory – the account of the actual pattern of change of a system (p. 97).

Although a detailed review of complexity related to the social sciences is beyond the scope of this chapter¹, it is important to consider how actors are interpreting and applying complexity ideas to the design and evaluation of social interventions. This includes why complexity is thought to be relevant, what key features are important, and how they relate to practice.

Relevance of complexity to social interventions. Some have argued that society is becoming more complex as we become more interconnected and interdependent. This intensifies at least three sources of complexity for social programs: in their internal functioning, the problems they seek to address, and their intersections with other systems in the environment. For example, programs are increasingly structured around multiple collaborating organizations, and this contributes to their internal complexity. Many, if not most, social programs are involved in some way with multifaceted and intractable issues such as poverty, which resist precise definitions and consensus on solutions. These so-called “wicked problems” are said to be affected by our interventions in unexpected and sometimes adverse ways, leading to changes in the situation that feed back on our efforts (see, e.g., Finegood, 2012; Rittel & Webber, 1973; Williams, 2015). At the same time, program actors are dealing with continuous change in the surrounding environment due to, for example, rapid technological advances, shifting demographics and effects of other interventions operating (and potentially conflicting) in the same space (see, e.g., Finegood, 2012; Preskill & Beer, 2012; Williams, 2015).

A complementary view is that an evolution has occurred in the attention and ambitions of society to intervene in problems that are more complex. Rittel and Webber (1973) coined the phrase “wicked problem” more than 40 years ago when describing a shift from simple and complicated problems (e.g., sanitation, transportation infrastructure, access to education) to complex problems (e.g., poverty, inequity). This shift occurred as more of the simple and complicated issues had been addressed (the schools were built; the roads were paved). These authors also describe increasing recognition over time of interdependencies and feedback effects in complex open systems and the emergence of new questions about the definitions and sources

¹ For in-depth reading, see, for example Byrne and Callaghan (2014), B. Davis and Sumara (2006), Mitchell (2009), Stacey (2001), Urry (2005). For an overview of the development of complexity’s influence on the sciences, see, for example, Alhadeff-Jones (2008).

of social problems. The questions stemmed from frustration at the persistence of many problems, as these authors conclude:

...the cognitive style of science and the occupational style of engineering---have just not worked on a wide array of social problems. The lay customers are complaining because planners and other professionals have not succeeded in solving the problems they claimed they could solve. We shall want to suggest that the social professions were misled somewhere along the line into assuming they could be applied scientists--that they could solve problems in the ways scientists can solve their sorts of problems. The error has been a serious one. (p. 158)

These authors go further to argue that it is “morally objectionable” for those designing and implementing social interventions to “refuse to recognize the inherent wickedness of social problems” or to approach a wicked problem as if it were “tame” (Rittel & Webber, 1973, pp. 160–161).

Some authors take the view that social interventions are complex by their nature, if for no other reason than because social interaction is complex (e.g., Mowles, 2014; Snowden, 2011; Stacey, 2005), adding that the distinction we make between simple, complicated and complex represents an epistemological approach, not a difference in ontological states of social interventions (Williams & Van t’Hof, 2016). From this point of view, complexity is always relevant.

Key features of complexity. Overall, emergent and non-linear effects and non-equilibrium in complex systems are assumed (e.g., Callaghan, 2008; Mowles, 2014). Particular attention has been paid in this literature to the influence of agency, feedback, and the program’s starting conditions, warranting a brief overview.

Agency refers to volitional actions and interactions of actors such as program staff, funders and clients (e.g., Berta et al., 2014). Actors ‘self-organize’ within, and in spite of, rules, procedures and structures designed for control and may even shape the system itself within which a program operates. Patrizi, Thompson, Coffman and Beer (2013) remind us that programs are operationalized by people (see also Pawson & Tilley, 1997). Since most programs involve numerous stakeholders whose situations and interests differ, their reasoning and their

responses also differ, shaping the outcomes of the program. For example, although administrators (and evaluators or researchers) may try to control for variation in a program's operations, front-line staff often continue to “beg, borrow, steal and adapt” to try to improve delivery (p. 53; and see Hawe, 2015 for a case example; see also Pawson, Wong, & Owen, 2011).

Feedback also affects the intervention. As a program is implemented, aspects of the system ‘respond’, serving to inhibit, accelerate or redirect change (e.g., Midgley, 2006). As a program gains momentum and realizes successes, the very conditions that made it successful are changed: “[a]n intervention's success is always time limited since alleviating a problem always involves changing its concomitant causes” (Patrizi et al., 2013 p. 53). Other sources of feedback include parallel interventions, which can create 'turbulence' and interference.

Programs typically evolve from previous efforts. Their characteristics and their outcomes are influenced by their history and context. The program's starting conditions influence its trajectory so that the same program implemented the same way in two settings can have very different outcomes (Byrne, 2013), sometimes described in terms of path dependency (e.g., Lahtinen, Guillaume, & Hamalainen, 2017). This is partly structural in the sense that systems are said to ‘embody’ their history; past agency is built into structures, which in turn play a role to limit or bound our actions. Byrne and Callaghan (2014) outline Archer's theory of morphogenesis as a 3-phase cycle of “structural conditioning”, “social interaction”, and “structural elaboration” (p. 119). Structural conditioning provides the context for social interaction. Together conditioning and interaction give rise to structural elaboration, which sets up the next cycle. The process is ongoing. They quote Archer: “[c]urrent structures arise from the actions of people in previous cycles so that there is a 'temporal escape' of structure from past actions” (Archer, 1995, cited in Byrne & Callaghan, 2014, p. 119). They also draw from Compton to put it this way: “the majority of actors in the system are dead” and context is a causal force (Byrne & Callaghan, 2014, p. 119).

Implications for practice. The ideas above have important implications for program design and development. For one, a complex system is likely to be resistant to policy- or “top-down”-directed change (e.g., Edwards et al., 2011; Sterman, 2006). Also, although a complicated system is knowable (in principle) and, given sufficient information, effects are predictable, a

complex situation is considered unpredictable (Williams & Van t'Hof, 2016). For these reasons a social intervention should not be seen in terms of a “flawlessly pre-planned change based on accurate predictions of the consequences of action” (Midgley, 2008, p. 55). More appropriately, a program is seen to provide input or resources to agents active within a system to enable their adaptation, in order to contribute to some aim (Midgley, 2008). A program contributes to change: “through the workings of entire systems of social relationships” (Pawson, 2006, as cited in Byrne, 2013, p. 220); it inserts itself into an open system as a fresh input that generates changes in patterns (Pawson & Tilley, 2004).

In this direction, Davies, Nutley and Walter (2008) conclude that knowledge does not transfer or import to new contexts, rather “new knowledge is created through social interactions *in situ*” (p. 189-190, emphasis added). Practice cannot be externally mandated; it has a life of its own with a local and “improvisational logic” (Wenger, 2010, p. 2). Although *information* can be codified and transferred across boundaries, the *knowledge* that determines its application is necessarily locally embedded (Ferlie et al., 2012; Wenger, 2010). Stated another way, knowledge is “what the practice makes it” (Orlikowski, 2002, p. 250, paraphrasing Taylor, 1993).

There are related implications for our approaches to testing and scaling program models. Some argue that conventional methodologies, in particular controlled experimental designs, are limited because they are insufficiently sensitive to context, human agency and emergence over time (see e.g., Connelly, 2007; Marchal et al., 2013; Mowles, 2014). For example, Chen and Rossi (1983) note that findings of no effect may be due to any number of factors, including “poorly constructed causal linkages” about the program or poor implementation of the program, but that RCTs rarely provide the information needed to fully understand these factors (cited in Alkin & Christie, 2004 p. 27). Others raise concerns that a “no effect” finding can arise from challenges with application of the RCT in real-world contexts (see e.g., Cronbach, 1975 for a discussion of laboratory vs. field experiments). Either can lead to a mistaken conclusion that the intervention itself is ineffective (see e.g., Hawe, 2015) and its subsequent unjustified abandonment (Urban, Hargraves, & Trochim, 2014).

As an example, Jain et al. (2006) report an RCT to evaluate effectiveness of an active KMb intervention in hospital intensive care units. The study found no significant difference in outcomes on a number of measures of the active intervention compared to the control (passive

dissemination of information). The team undertook a follow up study and identified contamination between the treatment and control related to peer networks between practitioners, as well as multiple factors that affected how practitioners responded, such as workload, the condition of patients and role of opinion leaders. The authors write that the influence of these factors “seemed to differ by site and profession” and they conclude that KMb implementation “is profoundly complex” (Jones, Suurdt, Ouellette-Kuntz, & Heyland, 2007, p. 449). Byrne (2013) writes: “we are faced with the problem that methods devised for systems that are both simple and not permeated by purposeful human agency do not work for understanding complex social cause” (p. 221).

This is not to say that positive effects of programs have not been demonstrated through controlled trials. Information on such programs, called model programs, is shared to encourage and enable their replication in other settings. These are typically interventions designed to produce positive change for individuals (compared with, for example, systems), often in the fields of health, mental health and education. For example, the Nurse-Family Partnership program is a “model program” endorsed by multiple organizations for its effectiveness in producing positive outcomes for infants and mothers (e.g., California Evidence-Based Clearinghouse, 2019; Social Programs that Work, 2019; University of Colorado Boulder, 2019). Trials in three sites in the United States have reported positive outcomes from this program (University of Colorado Boulder, 2019).

When positive effects of a program are demonstrated, a new challenge exists for scaling the model outward (Weiss, Murphy-Graham, Petrosino, & Gandhi, 2008). For example, although the Nurse-Family Partnership is promoted as a “model program”, replication studies in Europe of the program have reported mixed or nonsignificant effects (California Evidence-Based Clearinghouse, 2019; University of Colorado Boulder, 2019). If a program design is less effective in the new context than where it was developed, or less effective for some groups of people, too much emphasis on fidelity to the model can lead to adverse consequences for program clients (see, e.g., D. Byrne, 2013; Hawe, 2015) and program staff (e.g., demoralization, see Mowles, Stacey, & Griffin, 2008). This presents an ethical dilemma. Yet we often privilege fidelity over other concerns. To illustrate, Hawe et al. (2009) describe an evaluation of a drug prevention program in schools. The evaluation was designed to assess fidelity of implementation

to the program design. The evaluators observed that the program was being modified by teachers at the local level. Hawe et al. propose that had a complexity lens been applied, these adaptations would have at least been explored for potential insights. But they were not. Instead, it was recommended only that teachers receive more training (p. 95). Westhorp (2013) describes another cautionary example, noting adverse outcomes of the Early Head Start program for some disadvantaged families, in spite of the program demonstrating positive outcomes overall.

Although some assert a fundamental contradiction in the reliance on methods for complex phenomena that depend on stability and are based on a linear model (e.g., Marchal et al., 2013), most arguments have more to do with when and how methods are applied and for what purpose. Concerns point to their use before a program construct has been well defined (Urban et al., 2014), and/or in situations when the program must continue to change. Concerns include the pursuit of a definitive conclusion on effectiveness of a program conceived as a single, contained, fixed entity, with insufficient attention paid to the contexts of its application (Kushner, 2015a). These concerns have less to do with the methods themselves. As Campbell (1970) argued, multiple modes and measures are compatible with experimental designs, as is a study of process; when well executed they can be helpful in complex programs. Cronbach (1975) argued for treating results of real-world experiments as hypotheses and interpreting them in context. This recognizes that explanations they provide are partial (“swamped out” by context) and they “decay” as the social world changes (p. 122). Contextualized scaling (i.e., tailoring a model to new contexts) is an option in place of replication (The World Bank, 2003)². From a complexity perspective, one important means to mitigate the limits on knowledge is the adoption of pluralism in choices of methodology, that is, a commitment to the integration of diverse methods and perspectives, from multiple locations in the system (Byrne, 2013; Davies et al., 2008; Finegood, 2012; Mowles, 2014; Williams, 2015).

Alhadeff-Jones (2008) describes complexity theories as potentially emancipatory, in the sense that they challenge old models and suggest new ways of thinking; invite critiques of traditional practices and assumptions of science; encourage multi-perspectival, nondualist framings, and transdisciplinary work. However – importantly – this author also cautions that

² Patton (2011) describes the use of DE to systematically support adaptations of established practices in new contexts.

complexity theories may enforce their own frames on the development of knowledge and may narrow or direct our thinking in their own way as we apply them across fields of social sciences and education. Their caution reinforces the argument for pluralism and reminds us of the need for ongoing research as these ideas make the leap from theory to practice-based applications.

Proposed practice applications. Proposed alternatives are thought to incorporate greater attention to process, avoiding an exclusive focus on outcomes (Byrne, 2013; Davies et al., 2008; Finegood, 2012; Mowles, 2014; Williams, 2015). They can enable actors to watch for, and respond to emergent outcomes as they arise (Lanham et al., 2013; Patton, 2011; Snowden, 2011) - called an ethical imperative (Jabeen, 2016; Lanham et al., 2013; Midgley, 2006; Pawson & Tilley, 2004; Snowden & Boone, 2007). They may also encourage actors to value findings that are plausible or ‘good enough’ for the decision or action at hand, because at some point the pursuit of accuracy wastes scarce resources. In changing environments, if a high level of precision requires the investment of many months or years, the result may be irrelevant (e.g., Lanham et al., 2013; Midgley, 2006; Pawson & Tilley, 2004; Snowden & Boone, 2007).

Some propose ongoing development of interventions through action learning approaches, using iterative cycles of inquiry, reflection, and action (see, e.g., Kemmis & McTaggart, 2008a). Mowles et al. (2008) sum up a rationale:

we can expect our plans to achieve some of the things we intend, but a lot of things which are unexpected and perhaps unwanted... This would suggest action, and continuous reflection on action to be at the heart of what it means to make sense of one’s interventions with others.... What this suggests is a process of emergent planning where the plan itself and the assumptions behind it are subject to the same kind of reflexive examination as the work (p. 816-7).

The ongoing development of social programs through “situated” (in context) adaptive learning involves iterative cycles similar to action research. For example, Lanham et al. (2013) describe a project to reduce hospital-acquired MRSA³ infection, which used a “bottom up”, positive deviance approach to identify where actors in local institutions were already innovating. These efforts were then discussed and interpreted by cross-functional teams to create action plans for

³ Methicillin-resistant *Staphylococcus aureus*

wider practice change aligned with the local setting and systems. A key component of the project was the support of researcher-facilitators. In a similar way, Mohrman, Gibson and Mohrman (2001) argue that to be useful, researchers need to contribute to the self-design processes of organizations (see also Van de Ven & Johnson, 2006).

Place-based initiatives are said to be distinguishable from conventional approaches because of characteristics such as: multi-sectoral collaboration in decision making, shared ownership and “entwined accountability” among stakeholders, and a focus on a particular geographic area or region with an interest in and intentional incorporation of local knowledge (Bellefontaine & Wisener, 2011 p. 7). They are described as open, complex adaptive systems, interacting continuously with their environment. This presents serious difficulties for evaluation, including challenges for identifying outcomes and attributing them to the intervention, as well as diversity in stakeholder conceptions of appropriate approaches to systematic inquiry (Bellefontaine & Wisener, 2011).

Yet, complexity thinking has been taken up in a number of approaches to evaluation, and especially in developmental evaluation (DE). A set of principles developed and established by Patton (2011) is meant to guide DE in practice. These principles are: developmental purpose; evaluation rigour; utilization focus (i.e., emphasis on use of the evaluation by intended users); innovation niche (i.e., leveraging DE’s strength to support innovation); complexity perspective; systems thinking; co-creation and timely feedback (Patton, 2016b). In addition to complexity, descriptions of such approaches often refer to adaptive learning, experiential learning and/or sensemaking. I focus the next section of the chapter on unpacking these and related terms to support the conceptual framework for this study (Chapter 2). I will then turn to the role of intermediaries.

Adaptive Learning

The idea of ‘adaptive learning’ has gained currency alongside increased interest in complexity. As the argument goes, when issues are complex, we cannot predict and control our way to social betterment. Instead, ongoing adaptation is a more viable option. For the study’s conceptual framework, I describe adaptive learning as a process that makes use of experiential learning and collective sensemaking, and which unfolds within a social learning system (see

Chapter 3). Experiential learning provides a helpful umbrella concept. Sensemaking research provides micro-level insights into how an adaptive learning process might unfold. These are outlined below along with related research on what might support or hinder adaptive learning, and the potential importance of intermediaries.

Experiential learning. Kolb's (1984, 2014) experiential learning theory is described as foundational to action learning approaches (Rezania & Blyth, 2009, p. 148). Drawing on the ideas of Lewin, Friere and Dewey, Kolb describes experiential learning as a “dynamic cycle” with a central dialectic of four modes of learning: concrete experience, reflective observation, abstract conceptualization, and active experimentation. Kolb (2014) states that experiential learning requires tension between these modes of learning. Kolb also writes of learning as transactional between individual and environment. Included in the conception of environment are other people with whom an exchange is taking place, or objects and tools in use (Kolb, 1984). There is also the implication that learning involves unlearning, which requires first surfacing existing beliefs and assumptions.

Sensemaking. The use of the term ‘sensemaking’ has proliferated in recent years (Maitlis & Christianson, 2014). It is sometimes used to refer to interpretation, or ‘meaning-making’, but covers more than this. Sensemaking is described as a “theory to explain what happens” in experiential learning (Rezania & Blyth, 2009, p. 150). Although there is no unified theory of sensemaking, a number of common threads can be found. Three generally agreed-upon components of a sensemaking process are: (1) attention, (2) interpretation, and (3) action (see e.g., Argote & Miron-Spektor, 2011; Maitlis & Christianson, 2014; Tsoukas, 2009; Weick, Sutcliffe, & Obstfeld, 2005). These are thought to occur in a cycle, with action resulting in changes in the environment, an alteration of trajectory and modification of cues for actors (Maitlis & Christianson, 2014). This is a rich body of research with substantial empirical contributions, drawing on diverse, largely qualitative, methods. These include single case study (e.g., numerous contributions by Weick), longitudinal multi-case study (e.g., Maitlis & Lawrence, 2007), ethnography (e.g., Gioia and Chittipeddi, 1991) and narrative content analysis (e.g., Weibe, 2010). Weick (2009) describes sensemaking as a process of conceptual ordering which leads to “plausible interpretations of ambiguous cues that are sufficient to sustain action” (p. 56). He draws parallels between sensemaking and complexity ideas, noting that complexity-

related concepts like self-organization, emergence and turbulence relate directly to sensemaking. Some key ideas in sensemaking theory, and particularly with respect to collective sensemaking, are noted next.

Disruption and continuity, confidence and doubt. Sensemaking can be disruptive in a way that is thought to be important for organizational learning, and it is also a process of (re)creating order (Weick, 2009). Maitlis and Christiansen (2014) describe a clear theme in the literature, writing that, “strategic change, learning, creativity, and innovation each involve breaking up the status quo and creating at least temporary disorder. Yet, sensemaking is concerned with the construction of new meanings... that underpin new ways of organizing and understanding” (p. 93). Whether and how such new meanings are developed is central to sensemaking.

Related to this theme of disruption and continuity is a necessary balance between doubting and knowing (Weick, 2009). A sensemaking process is often triggered by an experience of dissonance, which introduces doubt, which in turn motivates search. Multiple studies report that over-confidence can inhibit doubt and subsequent engagement in active sensemaking (Weick et al., 2005). However, at the same time, some level of confidence is needed for a sensemaking process to move forward. Weick (2006) calls this a capacity to respond or “capacity to act” (p. 1724). Belief in this capacity can affect what a group pays attention to and how they filter information, because, “[a] system’s willingness to become aware of problems is associated with its ability to act on them” (Westrum, 1993, cited in Weick, 2006 p. 1723-1724).

Time: Past and future thinking and their influence on group identity. Although sensemaking is usually considered a retrospective process, recent work argues for a prospective view (e.g., McMaster, 2013; Weibe, 2010). Maitlis and Christiansen (2014) write that intersubjective meaning-making involves creation of narratives and metaphors for and of the future, and these have an effect on actors’ interpretation of the issues they face. People also take action in consideration of the action’s plausible effect on this imagined future (see also Gioia & Thomas, 1996). Sensemaking’s temporal orientation remains a topic of active debate among sensemaking theorists and is described as an area in need of more research (e.g., Strike & Rerup, 2016).

Writing on complexity ideas, Byrne and Callaghan (2014) note that we internalize social norms about time. These norms enable a sense of continuity, which is important to collective identity, particularly in our use of shared tools to enable a collective (and selective) memory. This is important for collective coordination in the sense that "relative change and sequential ordering... allow actors to impose meaning and order" which is important for coordinated action (Byrne & Callaghan, 2014, p. 140). Gioia and Thomas (1996) note a relationship between sensemaking and identity and point to research that supports the view that organizational identity can be somewhat fluid or changeable, and that identity change is important for organizational change to take place. This is echoed in writing on complexity by Cilliers and Preiser (2010), who describe group identity as dynamic and formed relationally. This may be supported through collective sensemaking (below).

Direct experience and abstraction. Weick (2009) suggests that disengaging from the action to take part in analysis of a situation can inhibit sensemaking, writing that: "detached, atomistic analysis" can cause the actors to lose context and their thinking to be less situated (p. 58). Yet experiential learning is thought to involve both direct experience and abstraction, and both actor and observer perspectives (Kolb, 2014). Abstraction may also be necessary to enable collective sensemaking (below). The question of how sensemaking may successfully transition between direct experience and abstraction is unclear.

Collective sensemaking. Each individual has a history and social position that give rise to unique points of view, which may be integrated with other points of view through social interaction (Strike & Rerup, 2016). Collective sensemaking is often positioned as either a socially constructed process that emphasizes use of language, or from a more cognitivist perspective as a convergence of individual mental maps (Sandberg & Tsoukas, 2015).

A collective sensemaking process may be enabled through processes of articulation (e.g., into tools), requiring what is implicit to be made explicit. The process of 'surfacing' and articulating can involve 'naming' and categorizing (Weick et al., 2005, p. 413), during which actors potentially distance themselves from experience and gain critical insight; this may support a transition between individuals' direct experiences and collectively held abstractions. Weick and Sutcliffe (2006) caution that this carries risk because "emotional distance" and "generalizations"

may result in superficial categorizations (p. 520). Articulation, then, may bring “workability” but also distortion in a process of sharing and integrating perspectives.

Sensemaking can also be distributed across a system. This view proposes that multiple individuals *and* their tools each hold some pieces of information important to the group and contribute in different ways as the process unfolds (Engeström, 2006; McMaster, 2013). For example, sensemaking about appropriate treatment for a patient can occur across debriefings between nurses and a physician on rounds (Weick et al., 2005). McMaster (2013) links this with theories of distributed cognition. He places sensemaking as “in the world” and as a “property of the system” (p. 45).

The possibility that a sensemaking process may be distributed across actors who perform specialized tasks has been presented as a potential advantage, in part because actors in distinct roles can specialize, for instance by attending to different cues. For example, in an emergency situation requiring an ambulance, the data on location and optimal route to the incident are a priority for a dispatcher, while data to judge appropriate action to stabilize the patient is a priority for the paramedic (McMaster, 2013). Others suggest a disadvantage. Senge (1990) describes “organizational learning disabilities” that occur when members of an organization hold different “maps” of the organization and assess value and meaning of incoming information differently (cited in Cousins & Earl, 1992 p. 401). Sloman and Fernbach (2018) also describe a risk to individual sensemaking that arises from distributed cognition. They write that because knowledge is distributed in communities, individuals tend to overestimate their own understanding. This in turn can foster strongly held but unjustified beliefs. They found that asking research participants to interrogate their own understanding (e.g., to write down in detail how a public policy should be implemented) significantly reduced participants’ confidence in their understanding of the issue in question and led them to moderate their positions away from extremes (e.g., for or against the policy).

Our sensemaking is becoming more distributed with increasingly decentralized and globalized organizational structures (Maitlis & Christianson, 2014) and increasing dependence on technology (McMaster, 2013). Consider the rapid integration of artificial intelligence in our sensemaking systems (e.g., information flow through social media and online purchasing

platforms or via surveillance systems). Distributed sensemaking is therefore a priority topic for research.

Sensegiving. The idea of *sensegiving* was advanced by Gioia and Chittipeddi (1991) to describe situations where individuals or subgroups may actively work to influence a sensemaking process in a particular direction. It may occur when a gap is perceived in an organization's sensemaking process and an actor concludes that sensegiving is needed to compensate (Maitlis & Lawrence, 2007; see also Gioia & Thomas, 1996). For example, when a situation is especially ambiguous and involves multiple stakeholders, leaders may engage in sensegiving to bridge perspectives if they believe stakeholders will not develop a shared interpretation on their own (p. 77). Other stakeholders may also engage in sensegiving if they think leaders lack competency in the problem area. For example, McCullum, Pelletier, Barr, Wilkins and Habicht (2004) report a case in which there was apparent cooption of a social learning process by more powerful participants, such that changes in perspective over time shifted toward these participants' views, without the participants or observers reporting awareness of it (McCullum et al., 2004; see also Cornelissen, Mantere, & Vaara, 2014; Heaphy, 2017). When it occurs, sensegiving is important because it: "can determine the issues to which leaders and others attend and can help shape the kinds of accounts and actions that are generated" (p. 59). An actor's perceived expertise and credibility in the area have been found to be important to their success in sensegiving (Maitlis & Lawrence, 2007), making this an important issue for the study of specialized intermediaries.

Power. Maitlis and Christianson (2014) note that until recently power was neglected in the sensemaking literature, and interest has grown in "the political process through which some interpretations become legitimate while others 'evaporate'" (p. 98). Brown, Stacey and Nandhakumar (2008) and Sandberg and Tsoukas (2015) write of 'discrepant sensemaking' resulting in political struggle. Similarly, Maitlis and Christiansen (2014) describe "the tussles and tensions of organizational sensemaking, as different parties campaign and compete to shape meanings of and in the organization, gain acceptance for a preferred account, or subvert the status quo" (p. 98; see also Weick et al., 2005). Although the dominant perspective may be what is encoded, inconsistencies because of differences in experiences or interpretations of experience among members may be preserved below the surface (Levitt & March, 1988, p. 328).

Social learning systems. The research on collective sensemaking describes a learning process embedded within a “social learning system”, which Wenger (2010) describes as having “emergent structure, complex relationships, self-organization, dynamic boundaries, [and] ongoing negotiation of identity and cultural meaning” (p. 1). Wenger makes an explicit connection to complexity, as do B. Davis and Sumara (2010) who describe social learning systems as complex, with emergent effects arising through interactions. For example, they describe research on classrooms in which it is proposed that a class may be thought of as a self-organizing, adaptive system, in which students: “operate as coherent knowledge-producing systems - that is, as collective we's rather than collections of I's - around topics of shared interest” (B. Davis & Sumara, 2006, p. 121). They suggest that a collective can be understood in this way as “an intelligent entity” (p. 121). Some specific themes relating to social learning that are relevant to this study are noted below.

“Social learning” encompasses multiple theories, largely complementary (Muro & Jeffrey, 2008), including with respect to individual learning by Bandura (1977) as well as theories of collaborative learning such as situated learning (Lave & Wenger, 1991), transformative learning in organizations (e.g., Argyris & Schön, 1978, 1996) and related aspects of activity theories (e.g., Engeström, 2001, 2006). Although a review of social learning theories would require a much larger scope than possible for this review, it is important to note their range and breadth. Given the focus of this study, situated and transformative learning are of special interest.

Situated learning theory posits that participation in social groups can result in learning, but participation involves more than just taking part in events and activities, it is:

... a more encompassing process of being active participants in the practices of social communities and constructing identities in relation to these communities. This understanding of participation implies that it shapes not only what we do, but also who we are and how we interpret what we do (Muro & Jeffrey, 2008, p. 328).

B. Davis and Sumara (2006) note widespread agreement that members of a knowledge producing community must interact. However, they point out that it is not always recognized that the important “members” of the system are not so much the individuals or subgroups but their *ideas*:

“the neighbors that must interact with one another are ideas, hunches, queries, and other manners of representation” (p. 142). Muro and Jeffrey (2008) assert that this aspect of situated learning is particularly key in the context of “wicked problems” for which information about the context or potential solutions is ambiguous or conflicting (p. 329). Other ideas about social learning include proposed use of “social action” as a way to promote “social learning”. Authors suggests that interactive processes to make decisions and address problems are helpful to enable actors to engage their ideas in collaborative learning (Muro & Jeffrey, 2008, p. 326).

Transformative learning involves changes in perspective or of underlying assumptions that influence behaviour (see, e.g., Argyris & Schön, 1978). Similar to sensemaking theory, these theories suggest that an anomaly or a “disorienting dilemma” may prompt a process leading to transformative learning when the anomaly “cannot be explained by old ways of knowing” (Muro & Jeffrey, 2008, p. 330). Much of the emphasis in the above research is on individual or within-group learning, rather than multi-group collaboration, however. The latter is of greater interest to a study of social interventions. Engeström’s (2001, 2006) expansive learning theory takes transformative learning beyond the individual or group to consider the effects of interactions among multiple systems, how this is mediated by tools, and how the systems themselves also change through the process of actors’ learning.

In a KMB study about knowledge brokering, Ward, Smith, House and Hamer (2012) write that “knowledge exchange is a social and political rather than behavioural phenomenon which involves professional identities and norms in addition to individual beliefs” (p. 302; see also Berta et al., 2014). They suggest that when knowledge exchange poses a threat to organizational norms it is likely to be resisted. They also propose that actors may actively seek to disrupt organizational norms through knowledge exchange. These authors relate the social aspect of knowledge exchange to the idea of ‘ba’ introduced by Nonaka and Konno (1998) in the organizational management literature. “Ba” is described as “a shared space for emerging relationships... that serves as a foundation for knowledge creation” (p. 40). Ray and Little (2001) also make note of the concept of “ba” in work by Itami (1992), who describes ba as a bounded space in which people develop a shared “interpretation code” (cited in Ray & Little, 2001, p. 158). This code guides their interpretation of incoming information or cues. This echoes accounts of the importance of developing a shared “grammar of meaning” among members of a

heterogeneous group to enable their collective interpretation and action, as noted by others (e.g., Snowden, n.d.).

Social learning as a process or a theory is still underdeveloped, and the amount of empirical research is still small in relation to its popularity (Fam, 2017; Muro & Jeffrey, 2008). Moreover, descriptive theories about learning do not lead directly to useful prescriptive theory (B. Davis & Sumara, 2006; Muro & Jeffrey, 2008). Some empirical studies are available that relate to the topic of this thesis, for instance, qualitative studies on social learning processes in environmental initiatives that require collaboration among diverse groups of stakeholders (Fam, 2017; Mostert et al., 2007; Pelling, High, Dearing, & Smith, 2008). Muro and Jeffrey (2008) and others (e.g., Mostert et al., 2007) argue that a social learning process should be attempted only when there is reasonable likelihood of success. They note that there may be risk in an effort to help a group develop shared understanding, perspective and interests, particularly if some actors have greater levels of influence than others, and/or if the effort may result in entrenched conflict. Others point to “collaborative inertia” as a frequent outcome of attempts by multiple groups to work together on social problems (Vangen & Huxham, 2003). This involves substantial investment of resources for very little return. Mostert et al., (2007) report on a multi-case study indicating that independent facilitation of participatory learning processes was a critical factor influencing success, and they urge further research on the topic of facilitation. I return to this topic below.

Challenges for adaptive learning. Over and above challenges arising from political influences identified in collective sensemaking and social learning research, pursuing adaptive learning as a response to complexity seems to carry its own internal contradiction. This is because the complex conditions that underlie the rationale for its pursuit also challenge the learning process itself (Argote, 2013; Lam & Shulha, 2015). Senge (1990) called learning from experience a ‘delusion’ because we generally do not experience the effects of our actions – effects are often too distant in time or place, or in proportion or in kind from our expectations. Although learning is usually assumed to be a positive, conditions like these are said to lead to “superstitious learning”, in which “the subjective experience of learning is compelling” but relationships between actions and outcomes are misapplied (Levitt & March, 1988, p. 325). This

can have serious implications for program decision-making. An extensive literature documents challenges to experiential learning. A few examples are noted below.

Ambiguity is a recognized impediment to learning (Maitlis & Christianson, 2014). Causes may be ambiguous because of distance between cause and effect (geographic/spatial or temporal) (e.g., Argote, 2013) and conflicting interpretations of feedback. The effects (or outcomes) can be nonlinear to inputs and unexpected in kind (Maitlis & Christianson, 2014), influenced by lack of clarity about what determines success or failure (e.g., Levinthal & Rerup, 2006), or subject to changing conditions, interdependent intervening events and co-adaptation of groups. The ‘frame’ through which actors ‘make sense’ of situations can be efficient, but can cause us to miss important cues (e.g., Mark et al., 2000) or distort our perception of them (Weick & Sutcliffe, 2006). ‘Unlearning’ or discarding a mental model can be difficult, particularly in the context of powerful professional cultures (e.g., Berta et al., 2014). Weick’s (2009) case studies on organizational disasters due to sensemaking failure are poignant, and Patton (1997) describes the limitations of “natural sensemaking” leading to “beguiling distortion of reality, awesomely selective perceptions, [and] stupefying self-deception” (cited in Mark et al., 2000, p. 76). Zollo (2009) calls superstitious learning the “dark side of experience accumulation” (p. 905) and urges more research in this area.

Denrell (2008) argues that managers in complex organizations are likely to encounter many “spurious correlations” between actions and effects which are “a fertile source of superstitious learning” (p. 278, citing Lounamaa & March, 1987). If there is substantial uncertainty about outcomes and if experimentation carries a high risk, there may be a strong incentive to perpetuate practices associated (incorrectly) with positive performance. Denrell illustrates with a hypothetical example of sacrificing a goat each morning before dawn because this is believed to make the sun rise. As long as the sun rises, the perceived risk of alternative behaviour is very high (the sun not rising), and experimentation is unreasonably risky. Denrell (2008) suggests that superstitious behaviour may be acquired by actors external to the organization as well, which then results in positive feedback into the organization. The author describes the case of enterprises that were early entrants during the technology boom in the 1990s. Management practices reflected value systems common to the age group of early

entrepreneurs. The practices were probably coincidental to the success of the early ventures but were assigned causal influence by outside actors, which reinforced them and spurred replication.

A tendency for organizational actors to reference successes rather than failures can also provide feedback that distorts a learning process (Argote, 2013). Actors in leadership positions may be especially vulnerable because those in their immediate vicinity may avoid challenging them (due to fear), may be unrelentingly supportive (due to self-interest), or may not be heard (lack ‘voice’). This inhibits leaders’ doubt, considered necessary to a sensemaking process: “[c]ontinuous exposure to supportive information induces sensemakers to believe that they know what they need to know. They have little incentive to slow down and doubt their knowledge” (Weick, 2001, cited in Strike & Rerup, 2016, p. 881).

Enablers of learning. Ideas about what supports learning under complex conditions are also rich but tend to be more contextualized. For example, the degree of intentionality of a learning process is considered important (e.g., Zollo, 2009), including attention to outliers and failures (Argote, 2013), but only if not too heavily politicized (March, Sproull, & Tamuz, 1991). Also, ‘mindfulness’ in a learning process may come at a cost. For example, so-called ‘mindless’ routines in an organization can be valuable forms of ‘encoded intelligence’ that offer continuity, stability and efficiency. Moreover, ‘mindful experimentation’, when it takes the form of “tinkering” with existing routines, can have unexpected negative consequences in a complex system characterized by interdependencies (Levinthal & Rerup, 2006, p. 510). Unambiguous, rapid feedback is said to be essential (Rudolph & Morrison, 2008), but others assert it is important to wait for emergence; that too-rapid feedback results in ‘premature convergence’ on a solution (Snowden & Boone, 2007). Others caution that feedback must be timed and presented with sensitivity or it can generate resistance (Langlois, Blanchet-Cohen, & Beer, 2012) and strategically selected to limit noise for decision-makers (Strike & Rerup, 2016).

The research on enablers and challenges to experiential learning under complex conditions is relatively well developed, supported by numerous studies from multiple disciplines, and employing a range of qualitative and quantitative methods. Overall, the results of this research have implications for adaptive learning. They suggest there are multiple interacting factors likely to influence adaptive learning, which vary in their effects depending on contextual and process-related factors. As a social process, adaptive learning fits the definition of a complex or “wicked”

problem (Rittel & Webber, 1973).

Diversity. One of the most pervasive suggested enablers of learning in a complex system is diversity of perspectives (sometimes referred to as cognitive diversity or diversity of viewpoints, e.g. E. King, 2017). In a complex learning system, diversity is said to be the “source of its intelligence” (B. Davis & Sumara, 2010, p. 858). Research suggests that diversity in groups can increase creativity and engender a “healthy skepticism” that limits overconfidence (Zollo, 2009, p. 898). As an example, Weick (2006) describes the establishment of interdisciplinary teams including social workers as a breakthrough event for the recognition of child abuse by paediatricians in the 1950s. Weick sees this as “a story of imagining and speaking differently”, prompted by social workers’ perspectives (p. 1724). Roberge and van Dick (2010) caution, however, that the empirical evidence linking diversity with creativity and innovation is limited and contradictory. The link needs development especially in the context of innovation. This is because innovation goes beyond creativity; it includes the application of ideas, processes or products that are new in the context and which are meant to bring substantial benefits over what currently exists (West & Farr, 1990, cited in De Dreu, 2006). This application component is an important distinction and it relates directly to definitions of adaptive learning (see Chapter 3 and further discussion in Chapter 9).

One key complication for realizing the potential of diversity is conflict, which is often reported in studies of diverse teams. There is some research to suggest that conflict can both help and hurt innovation (De Dreu, 2006). De Dreu (2006) found a positive relationship between task conflict and innovation as long as conflict levels were moderate, supporting other research suggesting that innovation increases with moderate levels of task-related dissent. For example, conflict may provide impetus for differences of viewpoint to surface and be openly debated, and when ‘solutioning’ is an imperative for a team, this may spur creative efforts to reach an acceptable way forward. In contrast, low levels of conflict are associated with avoidance, and higher levels with reduced ability for actors to process information, distraction away from tasks to personal relationships, introducing relational conflict, reducing trust and inhibiting analytic thinking (De Dreu, 2006).

De Dreu (2006) suggests that a source of the contradictions in the research on conflict and innovation lies in the choice of outcomes measures of different studies. Research indicating

a negative relationship between conflict and innovation tend to rely on team performance indicators, for example, goal attainment. Research suggesting a positive correlation mostly focuses on learning processes, such as “the development of insight and understanding, and the ability to solve complex problems” (De Dreu, 2006, p. 84). He writes:

Learning and complex problem solving require effort, time, and energy. It thus follows that, at least in the short run, task conflict hinders efficiency of work processes and goal attainment. At the same time, task conflict may be beneficial for those subcomponents of overall team performance that are most directly affected by team members’ ability to learn, to develop and implement new insights, and to solve complex problems (p. 84).

In sum, moderate task conflict as part of efforts to innovate may impede short-term performance (see “collective inertia”, above) but may benefit “long-term survival” of the group (p. 103).

In his discussion of complexity and problem solving in diverse teams, Alhadeff-Jones (2008) echoes the potential importance of pressure on a team to arrive at a solution. He offers the example of interdisciplinary teams assembled by the British military in the second world war, which had notable success finding useful solutions to complex problems “[u]nder the pressure of war” (p. 70-71). Interdependency may be an important influencing factor, in other words the degree to which members of a group need one another to perform their tasks (De Dreu, 2006, citing Saavedra, Earley, & Van Dyne, 1993). Stevahn and King (2005b) echo this by suggesting that it is positive interdependence reinforced by cooperative goals (in contrast with competitive goal structures) that encourages joint problem-solving for collective gain. These authors state that the ability to “structure positive interdependence” into an evaluation (e.g., in tasks such as constructing evaluation questions, articulating program theory) may be one of the most valuable skills for use in a participatory evaluation (p. 423).

Roberge and van Dick (2010) suggest there may be multiple interacting relationships between types of diversity (e.g., functional and cultural) and types of conflict (e.g., task conflict, values conflict). These may be intensified in the context of cross-sectoral collaborations. Wenger (2010) writes that ‘engagement at the boundaries’ of a practice, where different perspectives meet, can offer rich insight and innovation, but can also result in misunderstandings (where values, commitments, use of language differ). In a similar way, Engeström (2001) wrote that

when more than one activity system interacts: “[i]t is a source of trouble and a source of innovation, demanding actions of translation and negotiation” (p. 136). Along these lines, Roberge and van Dick (2010) recommend study of the processes that may enable diversity to have positive effects. Jehn and Bendersky (2003) provide a model suggesting factors that strengthen positive outcomes and reduce negative outcomes of diversity. This model highlights the role of emotion and the presence of “interest-based third parties” (p. 222).

Emotion. Strong negative emotions (e.g., fear, anxiety) are believed to constrain sensemaking; they reduce cognitive capacity available for sensemaking, affect how people attend to signals in the environment, and how they interpret them (Maitlis & Sonenschein, 2010), and are believed to increase commitment to existing cognitive frames (Cornelissen et al., 2014). Negative emotions also interfere with social interactions, disrupting the social processes that are important for sensemaking to proceed (Heaphy, 2017). With respect to emotion, some ambivalence is believed to be most conducive, e.g., both hope and doubt (Vogus, Rothman, Sutcliffe, & Weick, 2014).

Heaphy’s (2017) case study on sensemaking mediation focuses on the impact of emotion on sensemaking, and efforts of organizational actors in a mediating role to support sensemaking. They describe a mediator “attempting to calibrate the emotions with different parties involved in the conflict so as to engage in sensemaking and sensegiving” (p. 663); this suggests that the two factors in Jehn and Bendersky’s (2003) model, emotion and the presence of a third party, can interrelate. The final section of this chapter highlights research on intermediaries in learning processes, specifically as described in sensemaking and in evaluation.

Intermediaries in a Learning Process

An assisted or facilitated sensemaking role is explicitly noted for program evaluators (Mark et al., 2000; Patton et al., 2016), and also in KMb (e.g., Lanham et al., 2013). However, there appears to be little empirical work focusing directly on this role. In this literature, the terms “facilitator”, “mediator” and “broker” are frequently used, sometimes interchangeably. A preference for the term “facilitator” is notable in the evaluation literature. This may reflect a view on appropriate boundaries for an evaluator’s work, or desired positioning (e.g., as impartial). As described by Preskill, Gutierrez and Mack (2017): “[b]y providing non-directive

leadership, the facilitator helps the group arrive at the understandings and decisions that are its task...The role is one of assistance and guidance, not of control" (p. 11). It may also relate to ambiguity or flexibility about positioning (discussed below) or simply to convention. In case accounts, a DE's role is usually described as facilitative, yet a clear definition of facilitation is rarely provided (Baldwin & Lander, 2018).

In the KMb literature, intermediaries are usually described as performing a brokering role. A problem is described in terms of gaps between people or locations where knowledge is held, and the solution is to hire brokers to bridge or make connections. Brokers are frequently described as performing a passive role as conduits of information in this literature, but others note how their work also filters and transforms information (see e.g., Wathan, Wyatt, & Harris, 2008). Greenhalgh and Wierenga (2011) note the lack of research in KMb on knowledge intermediation. They suggest that "independent facilitation" may help to prevent a KMb process going awry. For example, when there is a stated desire to integrate multiple perspectives in the development of new knowledge, an independent facilitator may intervene if specific interests are finding a privileged place in that process (p. 508).

A handful of organizational studies describe mediated sensemaking (Heaphy, 2017; McMaster, 2013; Stigliani & Ravasi, 2012; Strike & Rerup, 2016). Strike and Rerup (2016) note the lack of research in this area. They suggest this may be because sensemaking research has largely focused inside organizations and ignored how an individual from outside may influence the sensemaking process. They also suggest the ambiguity of the role might contribute to mediators being overlooked. Use of the term "mediator" in these studies is in part because the actors are not seen to be exclusively "outsiders", as facilitators are usually understood to be. Mediators "work with and influence the individuals who are sensemaking yet are also 'outsiders' who pay attention to the larger context in which sensemaking occurs and introduce new cues and perspectives" (Strike & Rerup, 2016, p. 881). They span boundaries but are seen as more than brokers; the reported range of their activities includes directional intervention. Sensemaking mediators may cross from supporting others' sensemaking, to active participation in a collective sensemaking process, and even engage in sensegiving (Heaphy, 2017; Maitlis & Lawrence, 2007; Stigliani & Ravasi, 2012). Although sensemaking mediation is seen as involving a higher

degree of intervention, the mediator is generally considered to lack any formal authority over actors, similar to a facilitator or broker (see e.g., Heaphy, 2017).

In their empirical study focused on the private sector, Strike and Rerup (2016) observed specific mediator activities, for example creating pause; inserting doubt; “mediating voice” and “expanding the frame” of attention and interpretation (p. 889). Stigliani and Ravasi (2012) also observed activities, such as , “noticing and bracketing cues” for internal actors in a for-profit firm (p. 1249). Strike and Rerup (2016) assert that practices important to adaptive sensemaking, particularly promoting the inclusion of a wide and diverse set of cues in the sensemaking frame, and consideration of alternative frames, do not happen or develop accidentally. These practices need to be cultivated (citing Cornelissen et al., 2014). Although Strike and Rerup (2016) explicitly deny a *sensegiving* role for mediators, Stigliani and Ravasi (2012) propose that influencing activities blur the distinction between sensemaking and sensegiving. Similarly, Heaphy (2017) describes individuals in mediating roles “pivoting” from sensemaking to sensegiving once a satisfactory account of a situation is assembled (p. 643). This is consistent with sensemaking theories that describe a transition to action once a plausible account is attained, with what counts as ‘plausible’ varying by group and situation. Sensegiving activities include communicating or retelling summarized versions of the account while using care not to amplify negative emotions (Heaphy, 2017).

As noted above, a handful of empirical studies focused on sensemaking mediation have emerged in recent years. These are all qualitative studies exploring a process and a role previously unstudied but potentially highly influential.

Evaluators as learning facilitators. A great deal has been written about evaluators performing multiple roles beyond a technical one. They have been described as educators (Greene, 2000b); facilitators, interpreters, mediators, role models (Preskill & Torres, 1999); innovation thinkers, project managers (Lam & Shulha, 2015), servant leaders (M. Langlois et al., 2012); strategy coaches (Dozois, Langlois, & Blanchet-Cohen, 2010), diagnosticians (Baldwin & Lander, 2018), critical friends, organizational developers (Cousins & Chouinard, 2012), and the list goes on. The many labels relate to the diversity of perspectives and disciplines present in the evaluation field (Greene, 2000b; Mark, 2002; Stevahn & King, 2005a) and especially diverse views with respect to appropriate functions and positioning of evaluators, which have evolved

over time (Greene, 2000b). The variety of proposed roles also relates to the highly contextual nature of evaluation work, in the sense that the roles that evaluators perform vary by context. Moreover, some argue that within any one context, evaluators may need to adopt multiple roles if they wish to have influence with program actors (Patton, 2007); whereas others have argued that the performance of multiple roles in evaluation may erode authority and influence (Greene, 2000a). Although frequently referred to, the various roles are seldom defined or delineated in detail in the literature (e.g., Ensminger, Kallemeyn, Rempert, Wade, & Polanin, 2015; for notable exceptions see Greene, 2000b; Ryan & Schwandt, 2002). Mark (2002) provides some definitional background on the concept of ‘role’ for evaluators: “[r]oles are commonly thought of as located in a sort of social nexus; you can step into and step out of a role... Role is largely an interpersonal concept, locating one social position relative to others, to a social structure, or both” (p. 20). Mark et al. (2000) argue for facilitation efforts by evaluators in stakeholder sensemaking on the premise that this role may help to overcome some of the challenges and “dangers” of “natural sensemaking” (p. 147).

Some have commented on particular challenges for developmental evaluators to define, develop and adhere to clear role boundaries through the DE process (Poth, Pinto, & Howery, 2012) and to manage their positioning in terms of relationships and impartiality (Dozois et al., 2010). Lam and Shulha (2015) made note of how a DE can provide momentum to program development through emphasis on “iterative learning” and “unlearning” (p. 365). Langlois et al. (2012) describe the role in terms of servant leadership, through which the evaluator establishes physical and temporal spaces (e.g., meetings or online platforms) to enable stakeholder reflection and insight.

Some authors argue that evaluator roles in general are poorly defined, unrealistic and assumed to remain constant throughout an evaluation (Skolits, Morrow, & Mehalic Burr, 2009). These authors call for more “realistic and useful conceptualization of evaluator roles, roles evaluators actually tend to assume” (p. 276), noting that research in other disciplines indicates that people are generally adept at performing and navigating multiple roles, even conflicting ones. Role fluidity is not necessarily unusual nor is it necessarily negative, and it may be a skill. Ambiguity with respect to the evaluator’s role may sometimes be an advantage (see below), but it can also obscure ways in which some roles may be synergistic or be in tension with each other

or with stakeholders' objectives for the evaluation process (see e.g., Greene, 2000a). The positioning of the evaluator as an insider or outsider is an example.

The insider / outsider debate. The “inside” and “outside” distinction in evaluator roles is perhaps the most long-standing in the literature (see e.g., Weiss, 1972, cited in J. King & Stevahn, 2002), with each assigned its own strengths and limitations. Authors have recognized that the distinction between inside and outside is not as clear as it may seem (for examples see J. King & Stevahn, 2002). In order to build trust, a developmental evaluator for example, is described as an ally who supports the values and vision of the team, and who is embedded in the intervention (Langlois et al., 2012; Patton, 2011). Langlois, Blanchet-Cohen and Beer (2012) describe their position in a DE as ‘insiders’ which: “meant that we were also participants, allowing for the joint meaning-making most relevant to action research” (p. 42). Yet these same authors also describe being able to ‘stay outside the fray’ and ‘feed data back into the system’, suggesting an outsider stance (p. 40-49). They also suggest that an individual in an evaluation role can be seen as having ‘no personal or organizational agenda’, unlike program actors (i.e., those who manage or deliver program services). This is said to be important to trust and credibility (p. 49). This same tension is reflected in other literature. In the context of knowledge dissemination, Rogers et al. (2005) describe individuals located at the edge of a system as ‘heterophilous’ with dual insider/outsider perspectives. Similarly, Strike and Rerup (2016) describe sensemaking mediators as positioned in ‘intersecting social worlds’ (p. 898), potentially with multiple roles or identities.

Status. Gamble (2011) suggests the developmental evaluator role be explicitly ‘named’. This seems to make it easier for actors to accept critical feedback (cited in Patton, 2011, p. 219). Similarly, Strike and Rerup (2016) suggest that a sensemaking mediator’s particular status gives access to decision makers, and permits them to challenge ideas, so that they can “say things other people may not be comfortable saying” (p. 892). It also provides freedom from structural constraints that might exist for people working inside the organization. This permission to challenge and freedom from constraints may be enabled by an ambiguous positioning.

Tools as mediators. Tools (e.g., visual aids) are likely to play a critical role in all aspects of a sensemaking process, not least as “boundary objects” for collaboration among heterogeneous groups (Akkerman & Bakker, 2011; Foreman-Peck & Travers, 2015; Oborn,

Barrett, & Racko, 2012; Strike & Rerup, 2016; Wenger, 2010) and for continuity of ideas over time (Oborn et al., 2012; Pandey & Srivastava, 2016). Pandey and Srivastava (2016) define boundary objects as:

objects that embody shared meanings and are of interest to each community involved (Brown and Duguid, 1998), help clarify the assumptions and attitudes of each community to others involved and to themselves as well (Buur and Mitchell, 2011). Moreover, they are also known to enable reflection and second-degree learning within communities engaged by them (Brown and Duguid, 1998) (p. 317)

B. Davis and Sumara (2006) note that in order for actors' ideas to interact the ideas need to be represented in some way. Tsoukas (2009) suggests the use of artifacts as mediators may explain the 'how' process of transition from individual to group-level sensemaking. Tools may also be more influential than generally appreciated. The role of tools to cognition and learning is a key component of activity theory (Engeström, 2006; Engeström & Blackler, 2005) which proposes that interaction between people and their context is mediated in part by their tools. Oborn, Barrett and Rocko (2012) point to research indicating the influence boundary objects can have on relationships and on practice. For example, they can support exchange of ideas, shared understanding, and be a site of negotiation among collaborating groups. A boundary object can also impede communication if not well designed (Bakker, Kent, Hoyles, & Noss, 2011) or promote misunderstanding. Patrizi, Thompson, Coffman and Beer (2013) provide some reasons why theory of change (TOC) diagrams and logic models can be counterproductive in practice (e.g., by obscuring complexity, creating an illusion of linearity or simplicity). Kushner (2015a) notes that logic models can actively suppress pluralism in perspectives on a program. Often such models bias thinking in that only positive feedback mechanisms tend to be included (Kushner, 2015b). In these ways an object can also serve a sensegiving role and may privilege interpretations of one group over another and/or serve to alienate members of a collaboration (Oborn et al., 2012). For example, Tsoukas (2009) describes a visualization that unintentionally excluded a particular subgroup's 'voice', creating tension in a collaboration (p. 953). Oborn, Barrett and Rocko (2012) emphasize the importance of care and skill when interpreting and integrating information into boundary objects.

Epistemic objects. A related role is proposed for ‘epistemic objects’ (McMaster, 2013; Miettinen & Virkkunen, 2005; Stigliani & Ravasi, 2012). Epistemic objects can help to elicit and “embed emerging knowledge in material form” (Khazraee & Gasson, 2014, p. 15832; see also Miettinen & Virkkunen, 2005), while also pointing to new questions or gaps in understanding to widen the inquiry frame. Drawing from activity theory, Miettinen and Virkkunen (2005) define epistemic objects as:

open-ended projections oriented to something that does not yet exist, or to what we do not yet know for sure. For this reason, they are also generators of new conceptions and solutions and can be regarded as a central source of innovation and reorientation in societal practices (p. 438).

The open-endedness of an epistemic object is important, because it enables a practice to become: “an object of inquiry in order to produce novel and alternative ways of acting” (p. 438). A TOC diagram can be used in this way, as argued by early proponents of theory of change as an approach in evaluation (C. Weiss, 1995).

Some authors highlight the potential role of such objects to reduce defensiveness among actors in a collaboration and promote a shared learning process (e.g., Saari & Kallio, 2011). For example, Baldwin and Lander (2018) report a DE case study in which the process of developing a logic model with stakeholders was “as important as the final product”. This was because it reduced tensions among the partners by giving them opportunity to discuss their positions and concerns and by making their collaboration “more concrete” through depiction in a visual model (p. 11). As another example, Barbrook-Johnson (2017) describes the development of a model of farmer decision making on soil conservation, which was used to “structure and focus conversation” among stakeholders on this issue. In his words: “[t]he willingness of participants to attack the (unfeeling) model also brought out sensitive topics for discussion that might otherwise have remained untouched” (n.p.) The author points out that the value is not in the model, but in the process, however it was important that the model was specific, had “intuitive appeal” and was neutral (n.p.).

This research highlights how tools can influence a collaborative learning process. It suggests that attention to tools, including how they are constructed and used, is likely to be important for a study focused on adaptive learning mediation.

Conclusion

Adaptive learning and its mediation are emerging areas of research. Foundations exist, but they are fragmented across disciplines. This presents a substantial challenge for researchers who seek to integrate and assess the available knowledge base in order to advance research in this area.

Experiential learning and sensemaking processes are two areas that have been subject to extensive empirical research. Gaps remain, however, particularly with respect to collective and distributed processes. The study of applications of complexity-thinking to social intervention design and evaluation is burgeoning but young. Some rich and suggestive case study work is now available. On the whole, however, this literature remains oriented to introducing and describing alternatives to conventional approaches through conceptual contributions. Among the areas of research reviewed above, the most diverse and extensive appears to be that which documents very significant challenges to experiential learning in organizations, along with potential enablers. Key factors thought to influence outcomes have been studied empirically through both quantitative and qualitative methods. This includes multiple studies on the relationship between diversity, conflict and innovation. Even on this topic, however, studies have produced seemingly contradictory results, raising new questions. Although the potential for intermediaries to perform a positive role is suggested, particularly in learning processes involving multiple, heterogeneous groups, until recently their presence and their influence has been largely overlooked. There is a rich and open space here for study of intermediaries, their roles and how they may influence results.

Social programs are directed at solving our most pressing problems. They are challenging to design and sustain in ways that optimize their impact, particularly in real-world situations where the effects of a program can be ambiguous and highly sensitive to context, and where the ability to predict future events is limited. Scholars in three distinct but related fields (evaluation,

knowledge mobilization and organizational learning) are proposing new ideas. The intersection of these ideas presents an enticing area of research. Although the literature, particularly empirical scholarship, is sparse, some suggest mediation in such a process has significant potential and deserves greater attention (e.g., Obstfeld, Borgatti, & Davis, 2014; Strike & Rerup, 2016). This study's purpose is to contribute to emerging research in adaptive learning. In the next chapter, I detail the conceptual framework for the study that guided it toward that purpose.

Chapter 3 Conceptual Framework

This chapter presents the conceptual framework for this study. This framework was built on the literature review in Chapter 2 and designed to respond to the research questions posed in Chapter 1. Here, I also provide working definitions of key constructs. I close the chapter with an outline of guiding assumptions influencing my approach.

My research questions ask about the role of mediation in learning for program design and implementation, and the implications for an adaptive learning process. Four important concepts are embedded in my research questions. The first two, developmental evaluation and place-based intervention, were defined in Chapters 1 and 2. The other concepts, adaptive learning and mediation, are discussed in the previous chapters but need clarification to establish working definitions for this study. I turn to these next.

Working Definitions

Adaptive learning. As discussed in Chapter 2, the term “adaptive learning” is often used, but not often defined or critically assessed. What is meant by adaptive learning is obviously a central concern for this study, but I was not able to locate a definition satisfactory for this use in the literature.

Authors note that defining group-level learning at an operational level is “excruciatingly hard to do” (Arthur & Aiman Smith, 2002, cited in Lipschitz, Friedman, & Popper, 2007, p. 5). For a start, in a way similar to “innovation”, “adaptive learning” implies some form of application if it is to be understood as adaptive (see Chapter 2). However, making a clear link between a learning process and any particular outcome is very difficult. Moreover, what counts as a ‘good’ outcome may be debated among actors and is likely to change over time (Lipschitz et al., 2007).

For the purpose of this study, I approached adaptive learning as a process. Important ideas guiding this understanding include B. Davis and Sumara’s (2006) description of an “intelligent collaboration” which “generates a diversity of possibilities and... has a mechanism for critically debating the merits of those possibilities” (p. 86). This builds on an understanding of an adaptive learning process being triggered by change that requires a response outside of

actors' current "repertoire" (Gary, 2005). I also drew on Lipschitz, Friedman and Popper's (2007) description of productive learning. These authors suggest that we can recognize productive learning when we observe action that is based on valid knowledge, where "knowledge is valid to the extent that it withstands critical evaluation and is not based on willfully distorted information or unquestioned interpretations" (p. 15). Here, these authors acknowledge that action is not intrinsically positive and recognize the importance of process (e.g., critical interpretation). I drew on sensemaking theory to provide a set of "sensitizing concepts" likely to be relevant to this process (see Figure 3.1, below).

An adaptive learning process also involves people. As Pawson and Tilley (1997, 2004) told us, a program is not a 'thing' but rather it is resources and ideas provided to people, and it is the people who make something happen. From this point of view, any learning process (adaptive or otherwise) that is taking place in a social intervention is embedded within a social learning system (Figure 3.1). An adaptive learning process is brought to life within and through such a system. In it, an ongoing negotiation takes place through interactions, and these affect the system's structure, relationships, boundaries and identity, often in ways we do not expect (Wenger, 2010; see also B. Davis & Sumara, 2010; Lemke, 2000).

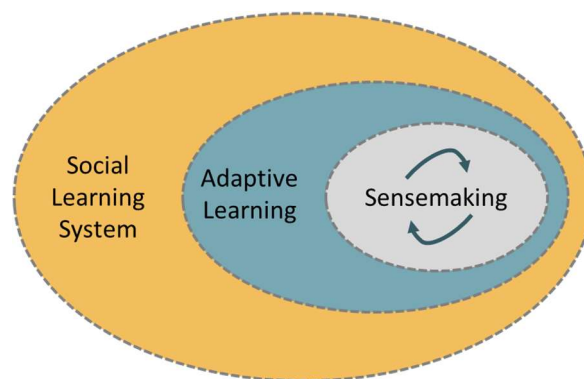


Figure 3.1. Adaptive learning in a social learning system

Hoped-for outcomes of adaptive learning may include increased flexibility and responsiveness to change (Muro & Jeffrey, 2008) in addition to or in place of improvement or betterment. From a complexity perspective, intelligence relates to the ability to "explore a range of possible actions" (B. Davis & Sumara, 2006, p. 86) and select from among them what is a good fit to needs and situation. These authors define learning as a process of becoming "capable

of more flexible, more creative activity that enables the unity to maintain its fit to ever evolving context" (p 92). This introduces resilience as an important potential outcome.

Lastly, I noted proposals in the literature that knowledge prove itself by its “functional fit” or ability to support “viable procedures” in a social program (Kushner, 1996, p. 197 with reference to Von Glasersfeld, 1991). Using functionality or viability as a reference point incorporates the idea of resilience. It also offers a platform for agreement among diverse perspectives in a pluralist-oriented inquiry in which values, for example, may differ among participants (Kushner, 1996). Drawing on these ideas: *an adaptive learning process involves generating data of range and balance, debate and critical interpretation, and informed, viable action.*

Mediation. Figure 3.2 introduces mediation, proposed to have an important role to play in adaptive learning, especially for collaborative interventions. For this study, *mediation is understood as an intervention that actively influences the relationships between components of a social learning system.*

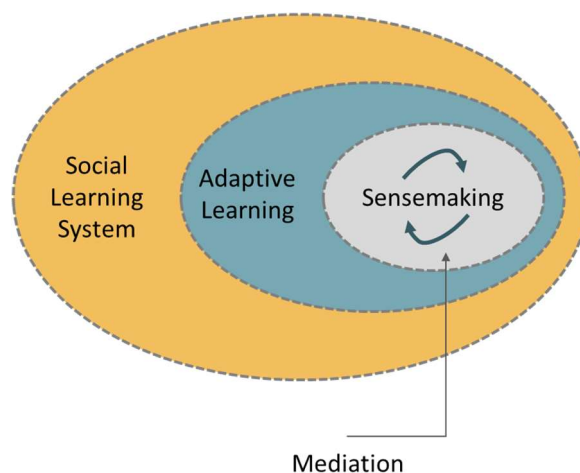


Figure 3.2. Mediation as an intervention in sensemaking within a social learning system.

In this study, I focused on the role of evaluators, however remained open to the potential for other actors and/or tools to play an important mediating role in a learning process (Miettinen &

Virkkunen, 2005; Stigliani & Ravasi, 2012; Tsoukas, 2009). Mediation in sensemaking may support “intelligent collaboration” (B. Davis & Sumara, 2010) and changes in the learning system. Sensemaking theory is important for understanding the micro-level process unfolding within, and deeply influenced by, that learning system in each case (see Figure 3.3 below).

Evaluation has been described as a learning system for organizations (Cousins, Goh, Elliott, & Bourgeois, 2014). From the perspectives outlined above, I positioned evaluation as a tool of social learning systems. This shifts the emphasis from an organization to the social system around an intervention and recognizes the limits of organizational control. The social system in a collaborative intervention may cross organizational lines and have a more fluid structure and identity.

System. Having introduced the idea of a social learning system, it is important to clarify what is meant by a system in this study. A system can be thought of as a “real-world thing”, as noted by Midgley and Lindhult (2017): “an organized set of parts, differentiated from their environment, giving rise to emergent phenomena that cannot be attributed to any one part, or sub-set of those parts, in isolation” (p. 3). However, for this study I adopted Ison’s definition, which presents a system as a thinking tool to help us make sense of a situation. In this definition, a system is: “a product of a distinction, formulation, or invention by someone, or a group... when a system is generated it is not a thing but a system-environment (or context) relationship mediated by a boundary judgment made for a purpose” (Ison, 2016, cited in Midgley & Lindhult, 2017, p. 5). This makes the system dependent on those making the boundary judgements, and it encouraged me to question my assumptions and framings as I proceeded with the study.

Conceptual Framework

At a more micro level, the conceptual framework for this study closely considered the sensemaking process in the inner circle of Figures 3.1 and 3.2. I drew from the work of multiple authors, particularly Argote and Miron-Spektor (2011), Maitlis and Christianson (2014), Tsoukas (2009) and Weick et al. (2005). Figure 3.3 incorporates three generally agreed-upon components of a collaborative sensemaking process: (1) attention, (2) interpretation, and (3) action. It also notes the influence of context, because experience is thought to interact with a group’s context as it becomes knowledge. This context is described as active (e.g., members, tools), latent (e.g.,

culture, structures) and environmental (Argote & Miron-Spektor, 2011). The mediator is included in the active context. Small arrows between the context levels denote interactions between them.

Also marked on Figure 3.3 (cycling arrows) are key transition points within the process. These are thought to be critical but not yet well understood, and include transitions between disruption and continuity, confidence and doubt, from individual to collective sensemaking (with the support of tools), between direct experience and abstractions, and overall, from a retrospective to prospective orientation (see Chapter 2). Arguably, each of these transitions relate to opening up a learning process to multiple perspectives.

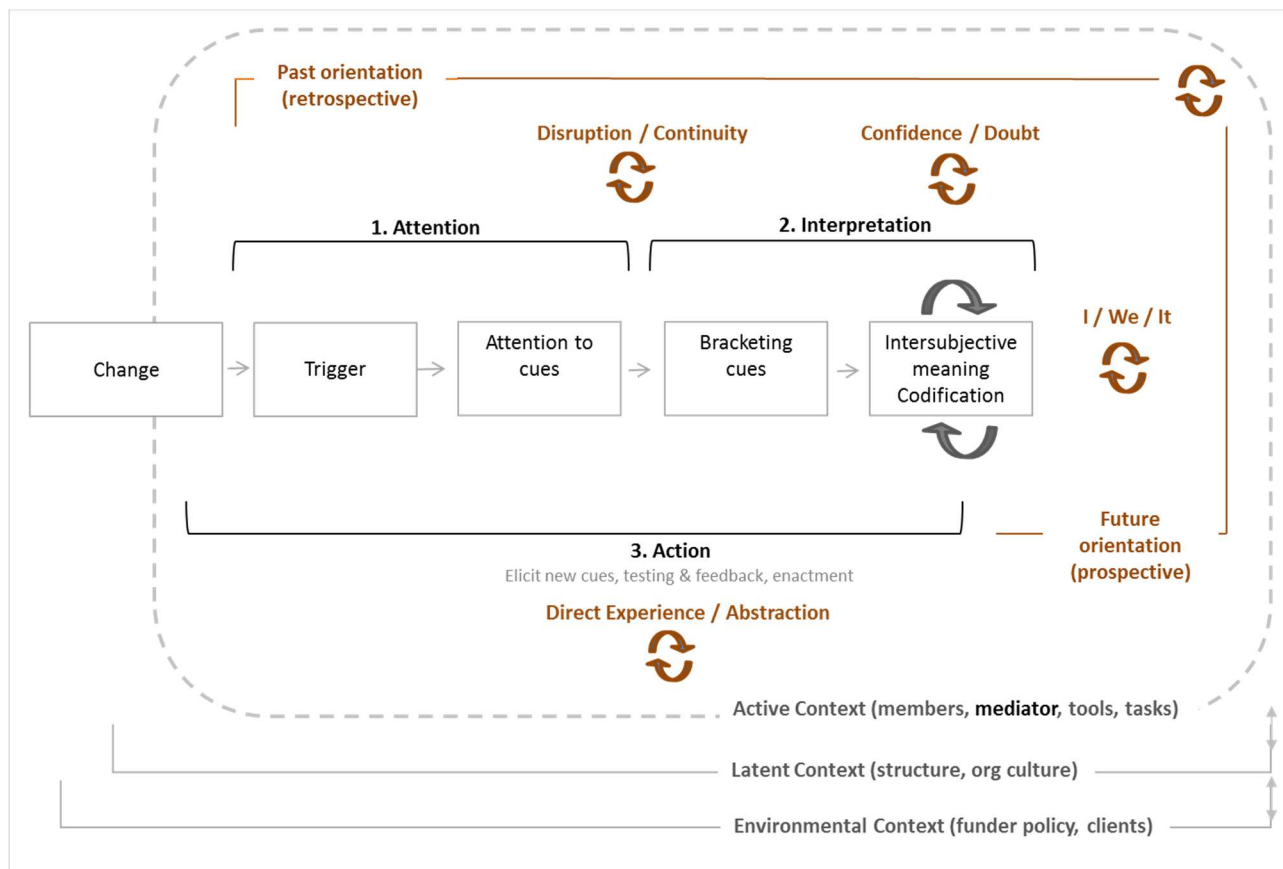


Figure 3.3. Conceptual framework for collaborative sensemaking in a social learning system

I propose that DE operates as an intervention by mediating a collective sensemaking process. This may include attempts to actively influence, by way of the evaluation, the relationships between components of the learning system in which the group is embedded. The learning system intersects with other systems, for example, those relating to funding, the implementing organizations, and systems of client stakeholders. An evaluation may span or connect different systems.

Guiding Assumptions

I approached this study from a ‘complex realist’ perspective (Byrne & Callaghan, 2014), influenced by pragmatism (e.g., Biesta, 2010; Biesta & Burbules, 2003; Dewey, 1930). Complex realism is a perspective that fits the topic and provides room for the study’s cross-disciplinary foundation. Realism assumes the world exists independently of our perception of it, and that our knowledge of the world is constructed (Byrne, 2005; Maxwell, 2013). Complexity theories describe reality as made up of open, nested and intersecting systems that can never be completely ‘known’. As a consequence, there will be multiple valid accounts from different perspectives (Richardson, 2004). The researcher is ‘inside the system’ not an outside observer (see also Byrne, 2005; Biesta, 2010).

These ideas sit comfortably with pragmatism, particularly on two fronts: the “transactional constructivism” described by Dewey, through which we construct an intersubjective world through interaction with others and our (real) environment; and also with the sense of openness and ‘flux’ in the world that present both limits and opportunities for the pursuit of knowledge (Biesta, 2010; Dewey, 1930; Strubing, 2007). For example, knowledge in a changing world may be temporal, and it is both developed and interpreted in context (Hall, 2013). Biesta (2010) explicitly links pragmatism with complexity thinking in this way, writing that: “[w]e are... participants in an ever evolving, unfinished universe... in which human beings are a creative factor and in which new things can emerge” (p. 110-113). Patton (2011) writes that pragmatism is the “epistemological foundation” of developmental evaluation, for which complexity thinking is an underlying principle (p. 285).

These ideas have influenced my approach in important ways. First, these perspectives encourage a pluralistic approach to inquiry (Biesta, 2010; Hall, 2013; Midgley, 2008; Rorty,

1982). Consequently, I approached this study with the intent to enable “conversations” among perspectives (Greene, 2007, p. 143) by designing the study to compare, contrast and, where possible, integrate multiple models and points of view (see Chapter 4).

This position also invites an approach to the study of causation in terms of how things change, for example through study of event sequences, transitions, and trajectory (Byrne, 2005) and the outcomes of transactions between actors and their environments (Biesta, 2010), and I took these up in my methods (Chapter 4). They suggest that causation may be multiple and contingent, particularly in the social world (Byrne, 2009; Marx, Rihoux, & Ragin, 2014). Causes may operate as systems or generative mechanisms (Pawson & Tilley, 1997; Ragin, 2009). The same outcome can come about through different mechanisms. Likewise, one mechanism can produce different outcomes, in part because causes are not independent of each other or their context. History and context, agency and feedback are therefore important to the understanding of case trajectories (Byrne, 2005). Within this, though generalizability is limited, ‘meaningful regularities’ do occur, and some are applicable in other contexts (Mark et al., 2000, pp. 153–154): knowledge is understood to be ‘local, but not relative’ (Byrne, 2005).

Third, a view of the world as evolving supports an approach to theory and practice as interdependent (Dewey, 1930; Johnson & Onwuegbuzie, 2004; Morgan, 2007), inviting study of ‘middle range’ theories, that is, within a defined scope or set of phenomena (Merton, 2012), that are approached as tools to be “developed in relation to the empirical” (Byrne & Callaghan, 2014, p. 109). Following this, I used my conceptual framework (Figure 3.3, above) as a ‘working framework’. Its features provided me with sensitizing concepts to collect and analyze data (Blumer, 1954). I later compared my empirical findings back to the framework to identify ways that these data support or challenge the theory, with a view to informing future research (see Chapters 9 and 10).

Lastly, pragmatism supports inquiry aimed at solving social problems (Hall, 2013; Johnson & Onwuegbuzie, 2004) and developing “understanding and knowledge in connection with work done, activities undertaken, and a consequent freer and more generously shared participation in... results” (Dewey, 1930, p. 79). These ideas motivate the use of participatory methods for a major component of the study (see Strand 2, Chapter 4), and the study’s overall focus on ‘the ways of doing’ of a living practice.

Given my professional background in evaluation, I identified with the commitments and the challenges of the evaluation practitioners and program designers whose work I describe in this study. Since initiating this study, I have also participated in three full-day training workshops for developmental evaluation practitioners, to increase my understanding of and sensitivity to the context of DE practice. I benefit from familiarity with the professional culture of evaluation, but I also recognize that this influences my perspective and my interpretations.

These first three chapters of Part I have situated the study and described its foundations. I used this work as a platform for decisions about methodology, described next in Part II.

Part II. Methodology

Part II provides details about the study's methodology. Here, I describe my approach to the research, including the study's overall design, methods and procedures. I note important characteristics of approach and design as they relate to the conceptual foundations of Part I.

Chapter 4 Study Design, Methods and Procedures

In this chapter, I describe the study's design and the rationale for the design. I then detail the methods and procedures I followed for each of the study's components and phases. I close with a description of steps taken to ensure quality and a note on challenges faced.

Overall, the study design and its methods reflect an interpretive and exploratory approach, consistent with my research questions and the nature of the topic of study. I revisit this further below after providing an overview of the design.

Study Design

I followed a mixed methods component convergence design (Greene, 2007) with two parallel strands. A component design retains relative separation of methods until the final stage, and distinct methods for each component are identifiable (Greene, 2007). A convergence design focuses the different methods on the same phenomenon to enable triangulation. I used qualitative multi-case methods for the first component, referred to below as Strand 1. For the second component, Strand 2, I used an integrated mixed-methods design for concept mapping. As an integrated design, Strand 2 used both qualitative and quantitative methods, which interacted over a sequence of steps to develop a concept map.

Strand 1 and Strand 2 were implemented concurrently and integrated in a final stage to produce overall findings, as shown in Figure 4.1. The two strands were treated with equal weight. More detail about the timing of implementation of the two strands and sequencing of their steps is provided further below (see Figure 4.4).

Outline of the design components. In Strand 1, I explored instances of mediation via DE through a multi-case study, with a focus on critical events, episodes, transitions and

trajectory. Strand 2 focused on components and structure of DE as a practice. I followed a group concept mapping (GCM) approach, which involved transforming qualitative input into quantitative data, then using results of statistical analyses as input to a qualitative interpretation phase (Goldman & Kane, 2014; Trochim, 1989). Findings from the two strands were integrated at the end of the study, following a dialectic approach described by Greene and Hall (2010). The methods for each strand and the integration phase are described in detail below, after a brief outline of the rationale for the overall design.

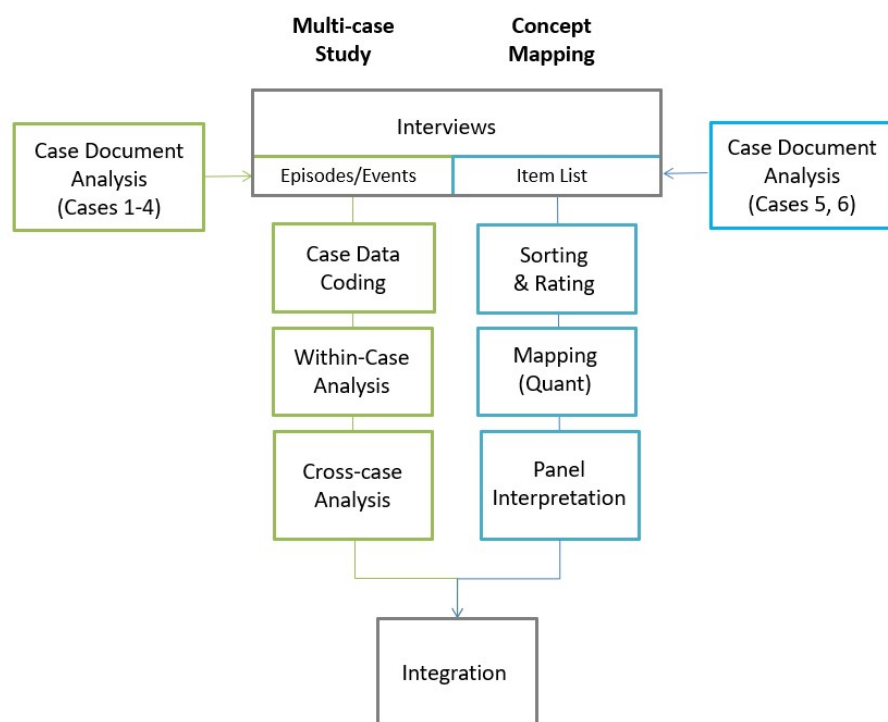


Figure 4.1. Study design

Rationale. I argue above (Chapter 2) that adaptive learning fits the definition of a wicked problem. Rittel and Webber (1973) assert that those who intervene in such problems have an obligation to recognize and work with the problem’s nature. Mixed methods has been promoted as an appropriate approach for complex phenomena (Mertens, 2015), in particular for its potential to improve understanding by systematically integrating multiple perspectives (Biesta, 2010; Greene, 2007) via different methods. As explained by Greene (2007):

The core meaning of mixing methods in social inquiry is to invite multiple mental models into the same inquiry space for purposes of respectful conversation, dialogue, and learning one from the other, toward a collective generation of better understanding of the phenomena being studied (p. 13).

A component design allows for more extended development of findings via different methods by keeping them separate through much of the research process. It also promotes the explicit juxtaposition and comparison of the two sets of findings.

In addition to method types, perspectives were compared in this design in other ways. For example, Strand 1 focused on process, and Strand 2 emphasized structure. Lemke (2000) proposes study of process as the 'fundamental unit of analysis' (p. 275, citing Bar-Yam, 1997) for phenomena that can be described as complex systems. Meindl, Stubbart and Porac (1994), note that studies that consider both the process and structure of thought (i.e., how information is used in action, and how 'knowledge' is organized) are rare, but can be especially rich if their relationships are linked in a meaningful way.

The study also incorporated multiple individual viewpoints within each strand. For Strand 1, I sought input from individuals in different roles in each case, and with different contextual experiences across cases. Strand 2 brought input together from several more individuals from different geographies and background experiences in DE, including positions in private consulting and academia. Finally, the study also incorporates diverse perspectives by drawing on research from multiple disciplines (see Chapter 2).

It is important to reiterate here that this study did *not* aim to describe DE as practiced generally in the field. Rather, I intentionally drew from "best possible" instances of DE (Ragin, 2009 p. 525, see Case Selection, below) and solicited perspectives of practitioners with a fairly consistent and in-depth knowledge of DE. This is in keeping with the study's purpose as described above.

Strand 1: Multi-Case Study

Case selection. A case is studied because it represents instances of the phenomenon of interest, it is a case *of* something (George & Bennett, 2005; Thomas, 2011). As noted in Chapter

1, DE projects were of interest because they seemed likely to provide robust instances of mediation for adaptive learning.

Careful attention to the construct of DE was a critical part of the study. DE's rapid development and growth in popularity in practice and in the literature over the last two decades (see Chapter 1) has resulted in an explosion of evaluations described as 'DE'. However, some vary in their features from what is described in prominent DE texts (Gamble, 2008; Patton, 2011; Patton et al., 2016). With this in mind, I drew from a set of 13 published accounts positioned as 'exemplar' cases of DE (Patton et al., 2016). These cases were recognized as consistent with DE principles and practice (Patton et al., 2016). They allowed for study of "best possible" instances of a phenomenon in keeping with this study's purpose. In essence, intentionally selecting what are likely to be positive cases (meaning, in which the phenomenon of interest is present and rich detail about it is available) is an important step in early case-based research, because it enables in-depth analysis of potential processes and features (Ragin, 2009; see also Eisenhardt, 1989). As Ragin points out, although for some forms of research (e.g., correlational studies) selecting based on values of a dependent variable is inadvisable, for case-based research on new topics of study, it is difficult to accurately identify negative cases without first having an in-depth understanding of positive cases. It is also very difficult to trace processes around a phenomenon when the phenomenon is not available or only weakly represented in a case.

I purposively selected four cases that had evaluator and program- or funder-based coauthors. The interventions described in these cases were all implemented by multiple organizations working in collaboration, and each involved a social intervention explicitly aimed at system-level change. As noted in the next chapter, in all of these cases, the actors involved were unable to identify an existing program model that matched what they envisioned, and therefore chose a developmental approach. The case collaborations all involved several 'inquiry-action' iterations spanning multiple years of work. I approached these cases for their capacity to demonstrate instances of mediation in a process aimed at adaptive learning. I focused on the cases as interventions in a social learning system (as described in Chapter 3), bounded temporally by the involvement of the evaluator(s).

Within-case data collection. I collected data from the published case accounts, other publicly-available documents about the initiatives (print and video), and semi-structured

interviews with at least one evaluator and one program member per case (in total, 9 interviews, 121 documents, 13 video files). Sources beyond the published case accounts were important for triangulation, to provide a more complete historical account and extend the data beyond the time of the writing of the published narrative, and to explore aspects of the case in greater depth, as described below.

Data collection followed a 'critical incident' approach with focus on turning points and transitions, the role of the mediator in these 'moments', and their relationship to case trajectory. I used this strategy to ensure that data collection would be focused on answering my research questions about the mediator role and implications of it for learning, important activities and points of intervention, how it is enacted and how the system responds, and, lastly, implications for the learning process overall.

To develop the interview guide for each case, I began with content analysis of each of the four published accounts. I coded the content of each account in NVivo following a predefined structure and combined this with a 'situational mapping' approach (Clarke, 2003) to identify major components and relationships at play in the system of each case (see template in Appendix A). I also constructed a timeline of major milestones. These data were used to construct the interview guide for each case. I solicited input on the interview guide for Case 1 from two other PhD Candidates with backgrounds in evaluation. I then made revisions and piloted the new version with one of the original reviewers (see Figure 4.2) and adjusted the guide a second time based on the pilot. I used this guide as a template for each of the subsequent cases, tailoring the guide for each case based on the above content analysis (see Appendix B). Each guide otherwise followed a consistent overall structure to support cross-case analysis. Tailoring each guide was important, because it allowed me to extend the data from the published account and to explore turning-point events in depth; a number of these events were either not covered by the published account or not described there in sufficient detail for the purposes of this study.

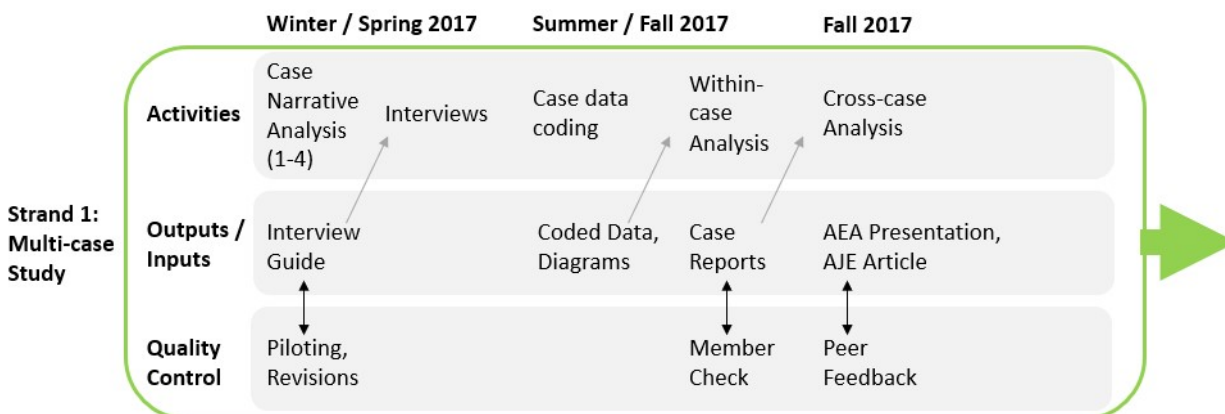


Figure 4.2. Steps in the multi-case study data collection and analysis.

Within-case analysis. All four of the selected cases were multi-year collaborations. This raised concerns about participants' ability to recall events other than those presented in the published account. This was a particular concern for the case with the longest duration (9 years), however coincidentally both interview participants had recently completed a review of their records for the full term of the evaluation for use as a teaching case. In addition, documentation was extensive for the periods of interest (e.g., media reports, interim evaluation reports, videos of presentations, program newsletters). During data analysis, I compared the accounts of different participants and triangulated these with available historical documentation.

I coded the data separately for each case in NVivo. Coding was inductive within an overarching, common framework across cases⁴; and informed by the conceptual framework (Chapter 3) and the research questions, supporting focus. The coding framework was also intentionally broad so that inductive coding within it could explore "what is going on here?" or "what is the story?" for each case, in its own right (Green et al., 2015, pp. 401–402). This follows Eisenhardt's (1989) advice for theory-building (exploratory) case-based research, namely to "understand each case individually and in as much depth as is feasible... [allowing] the unique patterns of each case to emerge before investigators push to generalize patterns across cases" (p. 539-40). The 'rich familiarity' that this develops with each case is important for cross-

⁴ Consisting of 'broad buckets': Starting Conditions (e.g., relevant latent context, trigger events), Enacting (e.g., structuring, strategies, methods), Unfolding (e.g., emergent events and turning points), 'End Game' (e.g., learning outcomes, current status of the initiative and evaluation).

case comparison (p. 540). I followed a process coding approach (Miles, Huberman, & Saldana, 2014) for the inductive coding, reflecting the words used by participants as much as possible. Process coding uses gerunds and attends to elements of action in the data, often as they relate to dynamics of context and time. For example, codes included “seeing the unexpected”, “reassuring people” and “walking the talk”.

I combined the NVivo coding with use of case trajectory diagrams, informed by event structure analysis. Event structure analysis applies case data to develop a path diagram. By visually tracing sequences of events, patterns relating to influences on trajectory and path dependency may become more apparent (Brueggeman & Boswell, 1998; see e.g., Stevenson, Zinzow, & Sridharan, 2003). In this study, I sketched a trajectory diagram for each case as a way to summarize the overall ‘narrative arc’ of the case⁵, including links between interdependent events, interventions and turning points over time. An example of one of the diagrams is provided in Appendix F. This helped me to keep the overall structure of the case ‘story’ together in a high-level visual summary, at the same time as the data were being ‘broken up’ through NVivo into discrete thematic codes.

I compared the preliminary themes generated in NVivo with the case trajectory diagram in an iterative process to develop higher-level themes. The two different ‘perspectives’ on the data - a time-bound and relational diagram, and a theme-based perspective from the NVivo coding – gave me different ways of thinking about the data. As an example, in one case, tensions between interests of actors at different levels of the evaluation appeared to be influencing the overall trajectory. This did not emerge in the NVivo coding but was strongly suggested in the diagram. The diagram prompted me to return to the coded data to confirm a possible pattern. Conversely, the diagram aided with the interpretation of emergent NVivo themes in the overall context of the case trajectory. Overall, comparison between the two analyses strengthened my confidence in the account and helped me avoid becoming “lost in the data” as I reconstructed the data into an in-depth summary for each case.

⁵ I adopted this term from comments by interview participants. I used the diagram to remind myself of the purpose to sketch an overall pattern of the unfolding and enacting of the process, similar to event-structure analysis.

When each case summary was complete, I revisited the original published account of the case (Patton et al., 2016) to look for contradictions. None were found. I then returned the summaries to interview participants for their feedback and received responses from at least one participant in three of the four cases. This led to minor adjustments in two summaries. The summaries were then fed into my cross-case analysis.

Cross-case analysis. For the cross-case analysis, I used matrix coding with reference back to the original data. I refined and extended the emergent themes using exploratory partially-ordered meta-matrices (Miles et al., 2014; see also Eisenhardt, 1989). A partially-ordered meta-matrix is used to lay out and summarize data in a visual display as a way to identify patterns aligned with themes and subthemes. I began with three matrices, one for each major category of themes developed through within-case analysis, and that were evident in multiple cases. I used a column for each case, and organized subthemes under these categories in the rows. Beginning with the case summaries (Appendix I), I recorded condensed data into relevant cells. For the first theme, I worked directly from the summaries to the matrix. For the subsequent two themes, I first coded the case summaries by subtheme in NVivo, and then worked from the data organized in NVivo to the matrices.

After review of the initial results, I separated each matrix by thematic clusters or dimensions to deepen the analysis. For example, the initial matrix for a central theme “Making Space” was subdivided into two matrices, one for “Why” and one for “How” space was made. I then revisited the raw data as needed, for example, to investigate gaps indicated by empty cells. I further condensed the data and reorganized rows to refine subthemes. For example, subthemes were eliminated if they did not hold well across cases. Some subthemes were collapsed into higher-level categories. Summarized versions of these matrices are provided with the cross-case results in Chapter 6.

Strand 2: Group Concept Mapping

One thing a person cannot do, no matter how rigorous his analysis or heroic his imagination, is to draw up a list of things that would never occur to him. (Attributed to Thomas Schelling, 1987, cited in McLinden, 2016)

Overview of GCM. Group concept mapping (GCM) is a mixed-methods approach that integrates qualitative data with statistical analyses to visually represent ideas on a topic of interest (Goldman & Kane, 2014; Trochim, 1989, 2017). The approach used in this study differs from some other approaches to concept mapping applied in the field of education, for example the development of node-link networks as knowledge maps for instructional purposes or student assessment. Applications include the use of a referent map created by domain experts for comparison with maps created by students, who create their own maps using pre-set concept sets and a software interface (e.g., Clariana, Koul, & Albright, n.d.; Trumpower, Filiz, & Sawar, 2014). The approach used for this study involves generating a set of statements about a topic in response to a focus question. The individual statements in the set are sorted into conceptual groups by participants for aggregate quantitative analysis using multidimensional scaling and cluster analysis. A concept map visualization is created for interpretation by participants as a final step. GCM is considered a conversion design because qualitative data is converted or ‘transformed’ into quantitative data. The map that is generated is not a final product, rather an opportunity to present the data back to practitioners with expertise in the topic area for “a fresh interrogation” (Clarke, 2003, p. 560) in a qualitative interpretation phase. The aim is to “create usable knowledge for complex practice issues” which is an identified challenge for researchers (Petrucci & Quinlan, 2007, p. 39 citing Kirk & Reid, 2002), by drawing on practitioner experience in a systematic way.

In the process followed for this study, participants independently generated idea statements, and also sorted and rated them independently. In this way, GCM provides a means to assemble and aggregate multiple independent points of view, which is believed to be important for ensuring a fulsome contribution of diverse and unique perspectives to study findings (McLinden, 2016, citing Surowiecki, 2004). GCM can also be used to avoid imposing a preconceived conceptual framework on participants, which can be important when theory in a topic area is weak (Behfar, Mannix, Peterson, & Trochim, 2011), such as for this study. The consistent reliance on practitioner input through all steps of the process also helps to reduce the influence of researcher presuppositions or assumptions on results (Michalski & Cousins, 2000). GCM can be a valuable input to research in complex topic areas by eliciting tacit knowledge of individuals with practice experience to ‘map the conceptual terrain’ from the perspective of those with deep experiential knowledge (Petrucci & Quinlan, 2007). The choice of a spatial method

was also based on considerations for quantitative approaches to the study of complex phenomena (Michael Wolf-Branigin, 2009).

Procedures followed. I used an approach to GCM described by Trochim (1989) since his approach has been well studied and well documented and now appears in a wide range of empirical research publications. I carried out Strand 2 over a sequence of four steps, summarized in Figure 4.3.

Overall, 21 participants contributed to the development of the concept map, in 3 non-overlapping groups. Participants in all three groups were purposefully selected for experience with DE in support of novel and/or complex interventions. The first participant group was made up of both program-based and evaluator members of DE collaborations and were the same participants who contributed via interviews to the case studies in Strand 1. Members of the second and third group were all experienced DE practitioners, having conducted evaluations for initiatives in different sectors (e.g., criminal justice, education, international development, community development, health care), in different regions of North America and abroad. The groups were mixed gender and represented a range of ages.

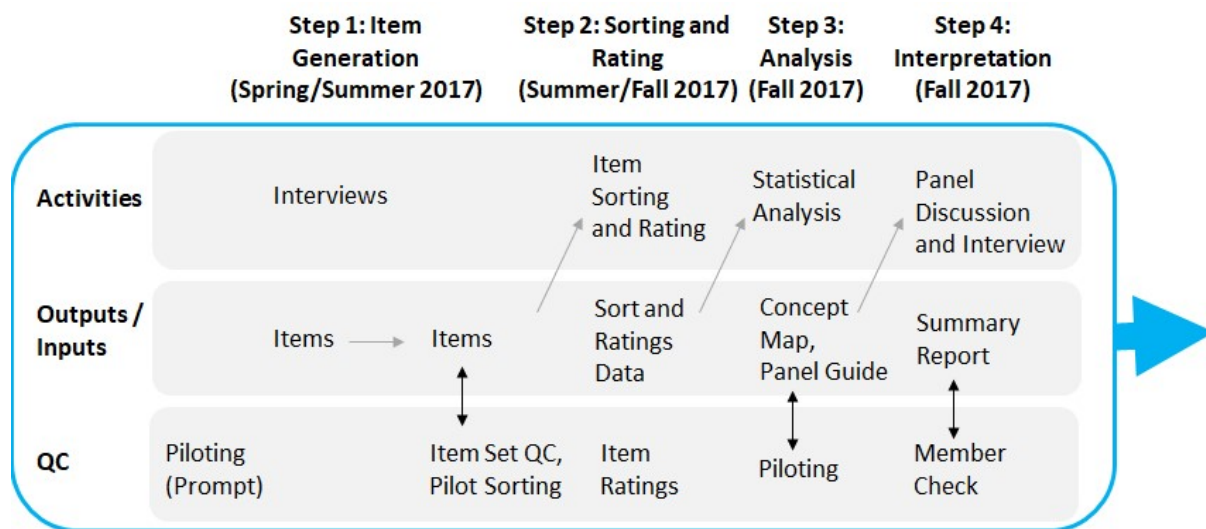


Figure 4.3. Concept mapping steps

Step 1: Generating the list of ideas. The interview participants in Strand 1 generated the initial set of idea statements, and I supplemented the list through content analysis of published

DE case narratives (Petrucci & Quinlan, 2007). The interview participants all had recent experience with developmental evaluations assessed as high quality (see Case Study methods, above). By involving these participants, I was able to include the perspectives of both consultant evaluators (n=4) and program staff (n=4) in the initial input to the concept map. All participants are based in North America but work in different regions and sectors. The participants were asked: “What are the most important things a developmental evaluator does?”. This prompt question was presented as the opening question in the interviews in order to keep Strand 1 and Strand 2 as independent as possible (see Interview Guide Template, Appendix B). Altogether, these 8 individuals provided 99 statements in response.

The appropriate number of participants to involve to generate statements depends on the purpose of the study. GCM studies have involved varying numbers, from 1 to over 200 participants, and Trochim (1989), who is credited with introducing GCM in evaluation, describes 10-20 participants as typical in his work. To ensure adequate coverage of ideas for the map, I analyzed two published ‘exemplar’ case narratives of DE, guided by the same question prompt. These case narratives were selected from a pool that had been qualified for inclusion in the Strand 1 multi-case study but had not been used for that part of the study. An additional 33 items were added to the statement list from content analysis of these case narratives, for a total of 132 statements.

I reviewed all statements for clarity and redundancy. Statements that contained multiple ideas were split so that each statement contained only one idea (Kane & Trochim, 2007). Redundant statements were eliminated through use of a “keywords in context” approach to content analysis (Kane & Trochim, 2007; Krippendorff, 2004). Through this approach, keywords are associated with each statement, and then the statement list is sorted by keyword to identify items that are similar. The items grouped by keyword are then reviewed for redundancy. Statements were also edited to remove ambiguities, for example, abbreviations and professional ‘jargon’. After the statement reduction process, 97 statements remained.

The reduced set was then reviewed against the original list by a second researcher (a fellow PhD candidate) who is familiar with DE, to ensure that no pertinent ideas had been lost, that the statements were clearly worded, and that each represented only one idea. Over two

rounds, the second researcher made suggestions for editing statements and recommended the reintroduction of one statement; I admitted 98 statements to the piloting phase.

The statement set was tested with two individuals who are not evaluation specialists. The pilot testers sorted the statements independently of one another, and each suggested additional edits to statement wording for clarity. The testers also provided feedback about the sorting process that resulted in eight statements being eliminated. These statements were found to be either too ambiguous for the pilot testers to sort meaningfully or too similar in meaning to other statements. All proposed changes were discussed with the second researcher who had provided the quality review and consensus was reached on revisions. The final statement set contained 90 statements. This number is consistent with recommendations for a concept mapping study using this approach (Kane & Trochim, 2007), considering participant burden for sorting and rating and the risk of attrition or spurious sorts as sets become larger.

Step 2: Sorting and Rating the Statement Set. A second set of participants sorted the final statement set into conceptual groups. Participants were purposefully selected for having knowledge and experience of DE similar to the group who generated the statements (i.e., were “reasonable proxies”, see Befahr & Trochim, 2011; Petrucci & Quinlan, 2007), but were not involved in Strand 1. Participants were either authors of published case narratives in the pool qualified for (but not used in) Strand 1 or were recommended by authors from this pool (snowball sampling). All participants are based in North America, although in different regions and working in different sectors.

Thirteen individuals agreed to participate, but only eight participants completed the activity⁶. This is a relatively small number of participants for a GCM study, but within an acceptable range when the purpose is to “assess the validity of the statements based on expert knowledge rather than for generalization purposes, and [when] the number of experts may be limited” (Petrucci & Quinlan, 2007, p. 30), as is the case for this study.

⁶ One participant contacted me after completing the rating activity to say they did not feel confident that their DE experience was sufficiently comprehensive. Their data were removed from the set prior to analysis. Four others who withdrew did not provide any data: two were unable to participate in the time available due to other commitments and two gave no explanation. Of note, in their review of GCM studies, Rosas and Kane (2012) report that on average across 69 studies, close to ½ of the participants who agreed to sort did not produce useable data. Therefore, I conclude that the rate of withdrawal in my research was not cause for alarm.

The recommended number of sorters for a GCM study has been a topic of conversation in the literature; frequently cited recommended ranges are 10-40 participants (Kane & Trochim, 2007) or 20-40 participants (Rosas & Kane, 2012), although much smaller samples are not uncommon (Donnelly, 2017; Rosas & Kane, 2012). A few points are worth noting relating to these recommendations. Trochim's (1993) reliability study for GCM found a positive correlation between sample size and reliability, however found that smaller samples (as small as 7) can produce maps with similar measures of 'fit' to much larger groups. Donnelly (2017)'s review of 108 GCM dissertations reported a negative correlation between stress (a measure of 'fit') and the number of sorters ($r=-0.30$), indicating larger samples tended to produce better 'fit'. In a pooled analysis of 69 GCM studies, Rosas and Kane (2012) reported a positive correlation between number of sorters and estimates of reliability, but based on the strength of the correlation, they conclude that the differences "may be minor" (p. 242). The number of sorters in these studies ranged from 6 to 90, median 20. These authors found no linear relationship between stress value and the number of sorters. Their recommendation for 20-40 participants was based on secondary analysis of randomly selected sorts from the largest studies in their sample (with sample sizes >60).⁷

The composition of the sorting group is an important consideration for determining an acceptable number of participants, but this was not a factor assessed in the above studies. GCM studies frequently recruit participants to represent multiple roles (e.g., social workers and family members of clients, see T. Davis, 2009; or external consultants, company management, and staff, see Michalski & Cousins, 2000). Reviews of GCM studies are highly likely to include studies of this type. As noted above, participants for this part of my study were purposefully selected for relative homogeneity, and my purpose was expert assessment, not generalizability to a heterogeneous population or comparison among subgroups. As reported below (Chapter 7) the

⁷ These authors performed a secondary, more detailed analysis, with results suggesting higher variability in stress value when the number of sorters is below 15. This led them to recommend 20-40 sorters. However, their finding is based on analysis of the five studies in their sample with the largest number of participants (>60). Although information about the participants in the five studies was not reported, it is reasonable to expect a study with more than 60 participants to have a fairly heterogeneous set. Consequently, we would expect a higher variability in the sorts of a small number of participants randomly selected from that set. (With each of the 5 studies included in their secondary analysis, they selected participants at random and, beginning with the aggregated results from two participants in each study, they calculated a stress value, then added the results of one additional randomly-selected participant at time, calculating stress at each step until the results of all participants were included).

stress value for my study fell within the acceptable range, despite the smaller sorting sample.

Participants sorted the statements working independently via a specialized online software program, CS Global Max⁸. Use of the online platform meant that I could recruit over a wide geography, which was important as the number of people with experience in DE in any one location is limited. The statement set was randomized before they began. They were asked to sort the statements into groups in ‘whatever way made sense to them’ (Petrucci & Quinlan, 2007), and to provide a label for each group of statements that best described the contents of the group, when their sorting was complete. Participants then rated each individual statement on a 5-point Likert-type scale for how often they found the item to be a priority for their DE work⁹.

Step 3: Analysis and Visualization. Each participant’s sorting and rating results were checked for quality (completeness and ‘reasonableness’, i.e., in accordance with instructions, no indications of spurious response) (Rosas & Kane, 2012). For example, sorts were reviewed to ensure the groupings were based on meaning, not other categorization such as a ‘not important’ group. Review of rating results looked for reasonable use of the scale, for example all items not rated the same value (Kane & Rosas, 2017). The sort data were then analyzed with specialized software¹¹ in the following sequence of steps.

Nonmetric multidimensional scaling (MDS). Nonmetric multidimensional scaling (MDS) is a multivariate, non-distributional technique, meaning that there are no assumptions made about the distribution of the data, which allows the inherent structure of the data to “emerge” (Petrucci & Quinlan, 2007, p. 31). The software used for this study applies an algorithm that aggregates each individual participant’s sort data in a similarity matrix. This matrix identifies the total number of times each statement was sorted with another statement by participants. Based on the aggregated similarity matrix, a distance matrix and a point map visualization were produced by the software (see, Figure 7.1), which provided a spatial representation of the relationships between the statements, based on the participants’ sorts. Statements most often sorted together appear closer to each other on the point map. Statements rarely sorted together are placed further apart. All participant data were included. I did not apply any filters in the MDS process.

⁸ The Concept System Global MAX™ (Build 2016.046.12) [Web-based Platform]. (2016). Ithaca, NY. Available from <http://www.conceptsystemsglobal.com>.

⁹ 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Always.

Anchoring/bridging and spanning analyses. I conducted and documented anchoring / bridging and spanning analyses to aid with the later assessment of the cluster analysis results. These analyses are automated through the concept mapping software, which provides values for each statement ranging from 0-1; a low value indicates an ‘anchor statement’. This is a statement which is similar to statements close by and not similar to statements elsewhere in the map. A high value indicates a ‘bridging statement’ that *is* like statements in other areas of the map, and potentially links different parts of the map conceptually (Befahr & Trochim, 2011; Kane & Trochim, 2007). Spanning analysis helps to identify the bridging relationships through visual display of the number, diversity and relative strength of these relationships. For each statement, the display identifies all other statements with which it was sorted, both within and across clusters, and by how many sorters. This aids with interpretation, for example by indicating relationships between particular ideas across the map. I used the outputs provided by the software to identify bridging and anchor statements prior to proceeding with cluster analysis, marking statements with values at the high and low end of the range on the point map (see Chapter 7).

Cluster analysis. I generated a 2-dimensional concept map using hierarchical cluster analysis. The software used in this study applies Ward’s algorithm in an agglomerative process to compare the distances between statements in clusters that are candidates for merger, in an iterative process that begins with each statement in its own cluster and merging 2 clusters at each step. The point map is the basis for the cluster analysis. A 2-dimensional solution is easier to interpret and allows for a 3-dimensional version to also be produced with the third dimension displaying information on the priority ratings provided by participants (see Figure 7.5).

With this approach to GCM, there is no single correct solution with respect to the number of clusters. Instead, a map is developed to provide a strong basis for subsequent participant interpretation; the number of clusters is analogous to resolution or level of focus (Kane & Trochim, 2007). I therefore followed a procedure described by Kane and Trochim (2007) to select a 7-cluster solution for presentation to participants for interpretation (see Chapter 7). This procedure prioritizes the purpose of the project and what is likely to best support participant analysis of the data. First, I judged an appropriate ‘practical’ range of clusters to be between 9 and 4 clusters. Starting at the upper end of this range, I reviewed the contents of the two clusters

to be merged to determine whether a merger makes conceptual ‘sense’, in other words, whether the contents of the two clusters are similar in a meaningful way such that merger is appropriate. I repeated this review of contents at each subsequent iteration and identified a ‘pivot’ point at 6 and 7 clusters. I then used the prior anchoring/bridging, spanning and rating analysis as inputs to select between these two solutions. Information about the 6-cluster and 8-cluster solutions was retained for the interpretation phase.

My analysis of the locations of statements with the highest bridging values in the 7-cluster solution resulted in my relocating two statements to adjoining clusters where they had better qualitative ‘fit’ to improve the map’s interpretability (Petrucci & Quinlan, 2007; Rosas & Kane, 2012). I moved statement #82 from Cluster 4 to adjoining Cluster 7, and statement #47 from cluster 1 to adjoining cluster 2.

Step 4: Interpretation. A third set of participants assisted with the interpretation of the resulting map. This step is common in GCM studies, to allow for interpretation of the results in greater depth, drawing directly from practice expertise; it foregrounds practitioner interpretations over those of the researcher. The involvement of expert practitioners at this step is also meant to identify any ‘fatal flaw’ in the results (Stewart, Shamdasani, & Rook, 2007, p. 44).

As with earlier groups, these participants were purposefully selected for their experience as developmental evaluators, based on publications of their DE work and/or information provided in an online directory of developmental evaluators published by the McConnell Foundation, which is recognized for its use of DE, and Social Innovation Generation¹⁰. Seven participants agreed to participate, but two withdrew due to scheduling conflicts. Of the five participants who provided input, four are located in North America, one in Oceania.

Four participants participated in a 1-hour, facilitated online expert panel discussion, and one participant provided input via a 1-hour semi-structured private interview¹¹. These methods allowed for direct interaction as I anticipated a need to answer questions or offer clarifications given the abstract nature of the map. The online format was chosen to allow for the participation of individuals located at a distance; this strategy is recommended when a specialization is

¹⁰ <http://innoweave.ca/en/modules/developmental-evaluation/coaches>

¹¹ This participant was unable to participate in the panel at the scheduled time.

relatively rare and dispersed (Stewart et al., 2007, p. 104). The panel discussion was chosen to allow for participants to respond to one another, providing data on points of agreement/disagreement or differences in interpretation, and so that participants could work together to clarify concepts that were likely to be ‘fuzzy’ or ambiguous, particularly with respect to the meaning of clusters and their interrelationships (Stewart et al., 2007, p. 104). Focus group discussions are also believed to be useful because synergies can enable more extensive exploration of ideas (Stewart et al., 2007, p. 104). The interview provided a helpful complement and comparison to the panel discussion, because this participant was not influenced by the ideas of others while reviewing the map.

I conducted both events virtually using GoToMeeting, supported by Powerpoint slides that displayed visuals of the map and related analyses, delivered via the platform’s screen-sharing feature. The map provided to participants included broad, draft descriptive statements for each of the clusters to seed the discussions. These statements were based on comparison of the labels assigned by the sorting participants with the list of statements in each cluster. I piloted the platform, supplementary materials and focus group guide in advance, once with two nonspecialists jointly, and once with a colleague who is familiar with DE (who also provided review and input to earlier phases of the study). The focus group guide is available in Appendix H, with the initial map provided to participants for interpretation.

All participants received two briefing documents at least one week prior to the sessions; one was a brief introduction to GCM, and the other provided the cluster map, the rating map and the statements in each of the clusters. At least three participants conducted some preliminary analysis of the map independently, prior to the sessions, based on these materials. This was valuable, because it increased the likelihood that independent ideas were brought forward in the group session. The same set of questions was used for both events, and these were also provided to participants in advance. Participants were asked about the resonance of the solution with their experiences, to consider the relationship between the entities depicted on the map, for example, sequences and dependencies, and to identify possible gaps or omissions, and implications for DE and its role in learning.

Both events were audio-recorded, and a chat log was generated with additional typed comments by panel participants. I transcribed both audio-recordings verbatim and incorporated the related chat comments in the transcription.

I coded the transcribed data in NVivo using a preset coding framework¹². Codes corresponded to each cluster and the interrelationships between the clusters. I also coded inductively by theme. I used a data display matrix (Miles et al., 2014) to organize the data corresponding to each cluster to ensure each cluster was fully developed in terms of proposed meaning, importance to practice, temporal relationships, and to identify any gaps. A summary of preliminary results was sent to the sorting and rating participants, inviting them to provide feedback. The summary included visuals of the map, the items in each cluster, and key points from the preliminary interpretation. I also sent a full draft of the results back to the participants of the interpretation phase. Feedback from participants was positive, no changes were suggested.

The two strands of the study were implemented concurrently as shown in Figure 4.4.

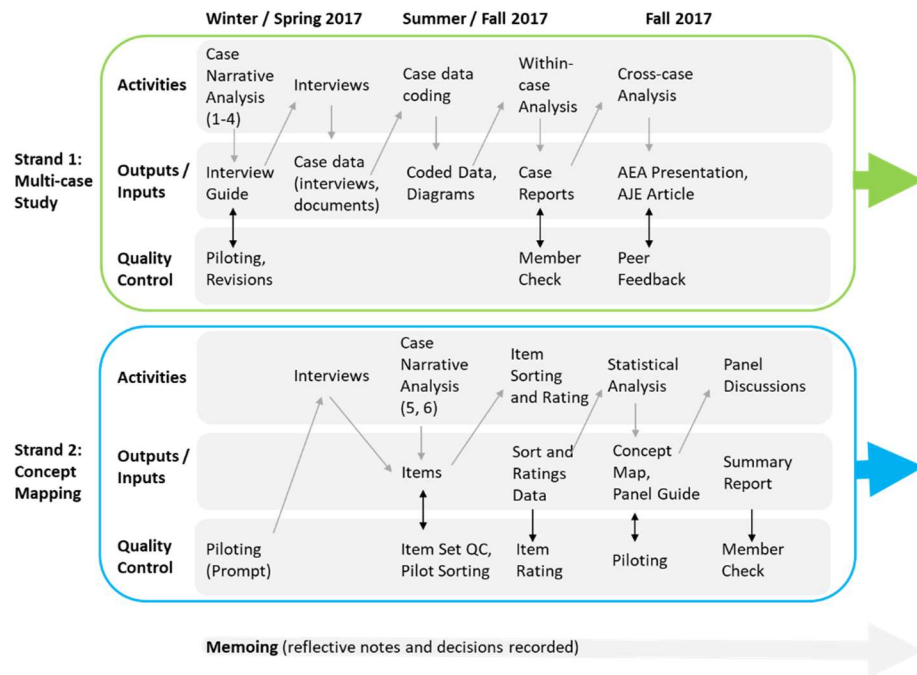


Figure 4.4. Implementation of the two strands over time

¹² In three major categories: Components (discrete elements); Conversations/Interrelationships; Overarching Themes

Although implemented concurrently, the two strands were not entirely independent: participants from Strand 1 contributed to the statement set for Strand 2, and the strands were also connected through me as the sole researcher. This is noted among the limitations of this study in Chapter 10. Working alone, I was not able to implement each step of each strand simultaneously, rather I worked back and forth between the strands as time permitted. For example, while participants were sorting GCM statements, I was working on within-case analysis

Integration of the Strands

The two strands of the study were integrated at the overall interpretation stage, consistent with a component mixed methods design. As noted above, the purpose of the integration stage is to compare and contrast the two study components as a way to further develop and extend the findings from each separate strand.

I drew from Greene and Hall's (2010) dialectic approach, which is intended to elicit insight as well as to assess convergence through juxtaposition of methods and perspectives (Greene, 2005, cited in Greene & Hall, 2010). As described above, the two components (Strand 1 and Strand 2) offer different perspectives on the research questions, a process-oriented perspective using qualitative multi-case methods and a structure-oriented perspective using a mixed methods conversion design.

I approached the integration analysis in three steps. The first two were aimed at answering the first research question about the role of mediation in the context of DE. The last step focused on my second research question about implications for an adaptive learning process. For steps 1 and 2, I used themes derived from one strand to analyze results from the other strand. As suggested by Greene and Hall (2010), analyzing data from one component of a study under the "structure of meaning" derived through the other can be a useful way to generate rich findings (p. 129). It can surface previously unconsidered connections and contradictions and serve as a check on understanding. Each step is outlined next.

Step 1. I used findings from the multi-case component of the study as an organizing framework for the first step. In this step, I focused on roles performed by evaluators as mediators identified through cross-case analysis (see Chapter 6). I used these findings on roles as a coding framework in NVivo and coded the GCM statements against this framework.

Step 2. For the second step in the integration, I used findings from the concept mapping component of the study as an organizing framework. Specifically, three “domains of activity” identified in the concept map (see Chapter 7) provided a coding structure. I coded the cross-case findings against this structure in NVivo. I used the same framework to organize a data display matrix, into which I entered the condensed data from the NVivo coding to compare and contrast with the GCM findings.

Step 3. In the final step, I compared my results against my original conceptual framework to consider their implications for an adaptive learning process. I reviewed the findings from both strands to the components of my working definition of ‘adaptive learning’, and key elements of sensemaking, and discuss implications with reference to relevant literature.

Quality and Validity

Validity is understood to refer to the credibility, trustworthiness or quality of the interpretations and inferences arising from research (Maxwell, 2013). Quality of inference in a mixed method context is understood to depend both on quality of the design and ‘interpretative rigour’ (Greene, 2007), and I incorporated a number of procedures to support both (Maxwell, 2013; Miles et al., 2014).

As expressed by Poggi (1965): “A way of seeing is also a way of not seeing” (p. 284). Overall, the comparison across different participant perspectives, data sources and data types in this design have provided for ample triangulation of results. Mixed methods designs also benefit in general from complementary strengths of qualitative and quantitative approaches. For example, the depth and richness of detail available to case study are strengths of Strand 1, but the case studies are also highly contextual, and depend heavily on researcher interpretation for synthesis and integration. Although less context-deep than the multi-case study, Strand 2 incorporates more points of view, allowed for independent points of view in analysis (e.g., independent item generation, sorting and rating), and provided a mechanism for aggregating perspectives into a high-level summary (McLinden, 2016, citing Surowiecki, 2004).

I maintained a practice of memoing over the evolution of the project, for reflexivity (i.e., to surface and consider how my assumptions were influencing my interpretations), to aid with insight, and to provide an audit trail (Corbin & Strauss, 2014) (see Figure 4.4, above). I had

accumulated over 125 pages of memos by the time of writing this chapter. Memos were maintained in topic files in NVivo and reviewed and synthesized at intervals over the course of the study.

Other quality procedures are noted in the relevant sections above. In brief, they included:

- attention to construct validity in casing and participant selection
- pilot testing of all data collection instruments
- personal transcription of audio and video data
- member checking by participants or quality check by a second researcher after completion of each data synthesis step, prior to feeding the results into the next phase of analysis, and
- soliciting peer feedback on findings from each of the strands before their final integration.

Challenges. Mediation and adaptive learning, which are concepts at the core of my study, are difficult to define. This makes for challenging research. I established working definitions to guide the study and these concepts were frequently the subjects of reflective memos. They were an explicit focus of attention in data analysis. As advised by one of this study's participants, I 'held my ideas lightly' about these concepts and tried to be receptive to empirical challenges to the framing I took up from my literature review. An explicit critique of the framework was part of the integration step and is documented in Chapter 9.

Part III. Strand Level Findings

In the next three chapters, I report the findings from each strand of the study. The multi-case component, Strand 1, is presented first. Summary findings from within-case analysis of each of the four cases in Chapter 5 are followed by the results of the cross-case analysis in Chapter 6. I report findings from Strand 2, the concept mapping component, in Chapter 7.

Chapter 5 Within-case Findings

Each case in Strand 1 has its own story to tell. I analyzed each in depth to understand actors' activities, the challenges they faced, how they responded, and what this suggests about mediation of a learning process. The analysis provides "rich familiarity" of the cases to support the cross-case analysis (Eisenhardt, 1989, p. 540). In this chapter, I provide a summary of my findings. The complete reports used in cross-case analysis are available in Appendix I.

Common Features of the Cases

Each case involves multiple organizations who are collaborating to develop and implement a social intervention. None had an established program model for what they wanted to do. The initiatives all have a place-based orientation, that is, solutions are developed through learning in the local contexts over time.

The evaluations were commissioned by the principal funding agency. Key actors in the funding organizations were described as open to innovation and experimentation, but DE was new to the organizations involved and effort was needed to build capacity for the approach. The DE work extended over multiple years (see Table 5.1). This gave actors an opportunity to experiment with a range of approaches, methods and tools over time.

Although each of the cases were analyzed independently and largely inductively, with variation in themes emerging, the overall coding framework provided a common general structure (see Chapter 4). Each of the case summaries below are presented accordingly in four parts: (1) background history leading to the start of the DE, (2) underlying strategies and/or

principles influencing how the DE was designed and carried out, (3) outline of implementation and case trajectory, and (4) learning outcomes.

Table 5.1. Description of the cases

Case	Duration	Sector	Objectives	Geographic scope	Evaluation Commissioner
1	4 years	Social services, education	Agency capacity building, youth education outcomes	Canada, urban area	NGO funder
2	4 years	Education	Post-secondary education for low-income students	USA, urban area	Foundation
3	3 years	Arts	Urban renewal	Canada, urban area	Foundation
4	9 years	Agriculture	Agricultural innovation, food security	International	Foundation

Case 1. A Better System for At-Risk Youth

Background. An NGO in a North American city had been working to improve outcomes for at-risk youth. The organization had identified fragmentation among service agencies as a barrier to outcomes. When the NGO received a very substantial donation of funds, it decided to develop an initiative to foster knowledge sharing and learning among city agencies. The goal was to change how the system functioned on behalf of youth.

This was a new way of working for the organization (d0, c12, c11)¹³, described as without “manuals” or “roadmaps” (d0; c11; c12). The initiative was considered risky because funding was usually directed to concrete client services. An actor described it as a “difficult sell”: “to say actually we want to use this money to create a better system... we had to tie that to real change” (c12). Stakeholders had high expectations for outcomes, and program actors knew that there was a “keen desire to demonstrate progress on stated goals quickly” (d0).

Actors at the NGO worried that their conventional approach to evaluation would not fit this initiative. They would need information quickly (d0/c12), and the initiative was already “in

¹³ Sources have been identified with codes that are cross-referenced to full identifiers in my files. Items with ‘d’ are documents, ‘v’ are videos, ‘c’ are case interview participants.

action even while underlying ideas [were] in fledgling state” (d0). In the words of one actor: “we knew it would be fluid and unpredictable to try to make sense of” (d0). In addition, participating agencies were concerned that the initiative would be used to “scrutinize” them (d-er). The initiative depended on the agencies being willing to openly share and discuss challenges together with the funder at the table, even though some of them competed for funding. They needed to develop trust, suggesting a collaborative approach to evaluation. A DE was commissioned when the initiative was 1 year old.

Strategies and principles: Making it part of the work. The evaluation was shaped around making it part of everyday activities. A team member said: “we tried to make sure that evaluation was the work and the work was the evaluation” (c12). This was motivated largely by a desire to build capacity for evaluation. To this end, many of the evaluation activities were conducted directly by funder staff, with the support from the consultant as a coach (c11, c12, d0). Agencies also took part in data analysis (c12, c11, d0).

There was also a goal to form ‘a continuous loop of evaluation’ (d-fr, d-v), so that data collection and analysis would better fit key decision moments (d0) and be sustained over time. Actors included simple data collection at regular events, for example: “we just had people write on a sticky note the answer to one question” (c12). A participant noted:

[Evaluation] should be part of their work more than this add-on thing... the more that’s a mantra you take, the more you have something useful and sustainable... the more [evaluation is] an extra thing the greater the likelihood that it gets dropped. ‘Oh, we ran out of time to do that today’ ... (c11).

Making it part of the work was also motivated by a commitment to collaboration, reciprocity and simplicity, an affirmative approach and a view that practice and theory are interconnected.

Collaboration, reciprocity. A commitment to collaboration and reciprocity is visible in how the evaluation was framed and modeled (c11). For example, a participant said: “instead of saying, ‘I’m going to collect some info’ and then disappear, [we say] ‘we’re going to collect some info together and we’re going to use it here now for something that’s probably relevant to you’” (c11). The team prepared data ahead of agency events so that members could participate in analysis, ensuring joint access to and involvement among the agencies in use of the data (c11).

Findings were shared back with agency members, which helped to create ‘a sense of shared ownership’ (d0). A learning agenda was developed with agencies each year. This helped to shape the evaluation around their learning goals and was said to help develop trust and ‘openness’ over time (d0).

Simplicity and opportunism. Keeping evaluation in reasonable proportion with the needs of the initiative was another principle (c11), for example, by emphasizing simplicity and leveraging opportunities as they arose. A team member said that a valuable part of the evaluation ‘coaching’ was “helping us learn... to think about all the opportunities when we could be gathering data in very simple ways” (c12), for example, “always thinking about how we can use this upcoming event as an opportunity to gather information, what is it that we need to learn now?” (c12). The team did invest in more intensive data collection and analysis, for example, a survey, focus groups, and case studies, but according to a respondent “we didn’t do a lot of that, we tried to make sure that evaluation was the work” (c12).

Affirmation, permission to fail. Accepting failure as an opportunity to learn in this initiative was said to contribute to a ‘different kind of conversation’. A participant expressed it this way:

I see all the time when people come together, they come together to talk about what doesn’t work, and how we need more money, totally legitimate. This is a different kind of conversation... building relationships around learning that are more egalitarian, that are more inquisitive as opposed to judgmental, helps to build that sense of shared ownership. It’s not about whether we get it right or wrong, it’s ‘oh that was a total failure, what did we learn from that?’” (c12).

This was expressed in the initiative and also in the evaluation. For example, a data collection instrument created by staff that ‘did not go well at all’ was approached as an opportunity to learn technical skills and to think differently about the initiative because the experience illuminated some important priorities (c11). The instrument was redeveloped. The new instrument worked well and later became an important tracking tool. The ‘failure’ and how the team learned from it was credited with sustained use of the tool for the duration of the case. A participant noted that

tracking tools are often abandoned, but in this case they “really stuck with it”, “maybe because it was a struggle, and then this moment of clarity and usefulness early” (c11).

‘Doing is part of theory making’. Actors described drawing a connection between doing and learning as a fundamental part of this effort. Emphasis was needed to integrate intentional learning in day-to-day work of actors who otherwise had an ‘action-oriented’ culture. Speaking of the sector, a project member stated: “I don’t think we value reflection and learning. I think we value doing, and we see those things as separate, and I don’t think they’re separate” (c12). Overall, a message that “doing is the creation of the learning” was said to lend energy and momentum to the initiative over the years (c12). Bringing the ‘doing’ and the ‘learning’ together required skill. For example, a participant described how they identified theories about change implicit in actors’ everyday decisions while “framing up emerging working models” as the initiative unfolded (c11). They explained: “for a lot of practitioners they’re not wired to think ‘this is our theory of change’ but it’s there and its sort of visceral and you can bring it out and its helpful to them” (c11). They clarified that: “the way not to do [it] is to stand up... with a whiteboard and a marker and say, ‘ok what’s your theory of change?’ Because it’s torture for everyone”; more effective is asking questions, seeking their interpretation of data and applying their input back to a formalized model (c11).

Implementation: Making space for learning. Over the four years of this case, actors worked to hold ‘space’ for a learning process, so they could develop the initiative. A participant reported needing “some time under our belt”. Later, key stakeholders were “completely totally excited about [the program]”, but “it took time to get there” (c12). Coaching, questions, continuity, and credibility were important to making space to develop the intervention.

Coaching. A participant described how important the staff leader’s role was to sustain the approach to developing the initiative and prevent drift back to the way the work was “always done”: “[the team leader] held that space and normalized it” (c11). The team lead took time to reassure staff one on one. The consultant reinforced their efforts with ongoing coaching about DE (c12). The effort was needed in part because of a tension between DE’s reflective learning approach and existing norms. For example, a participant suggested that people can feel threatened and unsettled if they are used to action-oriented work (c12). In this case, staff:

...didn't have a history of working that way, and I think they felt quite threatened by that... this constant asking about the work was interpreted by them as meaning 'I'm not doing the right thing', and 'when can we ever just, like, land and say this is the way we're going to do it, why do we have to keep talking about it' ... I think that for people who aren't used to working in complexity and... are really good at implementing and getting stuff done, having to keep opening that conversation and revisiting [it] was not comfortable for them, it created a lot of ripples and it took time to work through. (c12).

It was said to become easier over time: "because the team saw value and benefit and it just became part of their practice" (c11, c12). It was said that DE itself created space for learning (d0, c12). A participant called this, "the greatest value that DE brings to my work... it creates a container or space for regular reflection, for relationship building, for learning together about the work" (c12).

The importance of questions. A participant said that DE is: "about asking good questions and helping [people] ask good questions and helping people to be curious about their work" (c12). Questions were used to give structure and for framing – to enable both focus and exploration.

Questions as structure. Questions were reported to be important for creating 'a kind of structure' that supports collaborative analysis and reflection; to help the team to make sense of what was happening and 'avoid running headlong into implementation of the initiative' (d-er/c12). A question framework guided reflective practice sessions, held 2-3 times a year. The general format was: "what have we been up to, what have we learned, what do we do going forward?". These questions were described as "simple, but the thing that's often missing from organizational work" (c11). They reportedly helped the team to rethink the initiative's theory of change (d0) and proactively "hone strategy and course correct" (c12).

Framing. The power of questions for framing an issue for focus was a key theme. For example, good questions were said to help evolve the group's understanding of the nature of the problem they were targeting, to help the team to consider what success would look like, and to help frame the intervention itself (c11). A turning point was credited to a set of good questions. When these questions were used at an event, the result was: "outstanding, like really a lot of

insights... [It] gave us direction on how to respond, injected a lot of energy and momentum into the work, and... [this led to] some really good strategy decisions” (c11). A participant said that questions sometimes ‘helped people to see better’ (c11).

Questions were credited with helping to reframe this project’s goals. A project member made a contrast between conventional questions that ask whether what was planned had been done. The questions they posed in this case led them to learn that: “[a]ctually... we couldn’t do that. So what we learned was... that’s not the way we need to frame this” (c12). For example, the initiative was expected to impact graduation rates:

...through DE we were able to have very honest conversations about ‘how doable is this?’... We learned about attribution and contribution and... through the development of our theory of change ...[w]e were able to learn to be much more specific about what we’re doing and what our contribution is and manage expectations about that (c12).

Over the course of the evaluation, new questions arose as a result of changing conditions. Effort to keeping the thinking frame wide helped the team to recognize and incorporate these new questions. For example, agency membership in the initiative grew rapidly over the first two years, reaching nearly ten times the expected participation. This led the team to ask about the value and feasibility of continuing to grow. Working through this question led to a decision to continue welcoming members because that aligned with the data and the purpose of the initiative (d0).

Continuity. Space for the approach was also held by creating a sense of continuity for actors even though the initiative was changing. Continuity was maintained through documenting the path taken and outlining the path forward.

The path taken. For example, documenting their progress enabled the team to ‘describe the rationale, basis or evidence for every key decision’ (c11, also d0). A data collection tool was used at events 2-3 times per year over the four years. The team also used a structured document to record their reflective sessions. In the end this ‘revealed to us the whole arc of the experience, how the thinking evolved over time, what we figured out, how they responded’ (c11).

A theory of change (TOC) was developed and updated over time to document actors' understanding about the intervention. The first time the team recognized faults in their theory, it was experienced as: "oh my gosh we actually need to rewrite our theory of change!" (c11). But over time, iterating the TOC became normalized and a 'source of pride', because it showed their learning: "in the end, we were quite proud we were on version 9 of the theory of change" (c11). The TOC was illustrated in visual form and used for external communication, as a boundary object, "it helped people talk with others about [the initiative]" (d11). It was also said that: "The sense that [the initiative] was evolving kind of lived there", meaning, in the TOC (d11), and "it clarified something that was ambiguous" (c11).

The path forward. The TOC was said to have been used for 'making good choices' (c12). Described as 'watching for deviations, reminding of the rationale, questioning shifts, pulling people back' (c11); 'using data to say, are we on the right track?' (c12), or 'pointing out if decisions were being made without or in spite of the data' (c11). The TOC was also a platform: "to check what [program actors] were experiencing against their assumptions and expectations" (d-er). The team avoided locking in to the TOC however, saying that: "we never assumed that what we were doing was the right thing" (c12). Overall, the approach "enabled us to hold our ideas about the initiative lightly" (c12, d0).

Credibility. The actors needed to explain and defend the developmental approach over time. The staff lead's role as a champion was described as key because they actively helped leadership understand the work and what was being learned as it was unfolding (d0). The ongoing documentation provided reassurance and transparency. A second external evaluator was also asked to review the evaluation. Their report included contributions from academics with expertise in complexity thinking (d-er). These were: "trying to give [DE] legitimacy and currency but also some explanation around what this is", so that readers could "situate it" because this was not a traditional evaluation (c11).

An expert panel was also formed. The panel was described as 'an evaluation intervention' (c12, c11). Its members had expertise in different sectors and disciplines, which participants described as important to help make sense of a complex situation (c11, c12). They were referred to as 'independent experts' in public documents (d0, d-er) and they "had an awful lot of credibility" (c12, also c11). The embedded structure of the evaluation and the evaluator's role as

coach presented a ‘risk in perception’. This was offset through the panel; their input gave weight to the team’s recommendations (c11). Moreover, the panel “saw [DE] as a legitimate approach and people then within the organization said: ‘if they’re on board, then we are too’” (c11). The panel provided an assessment that was key to a decision to continue funding the initiative four years on (d0).

Learning outcomes. Accounts of this case describe learning at various levels. This included evaluation capacity, such as technical skills, for example, in survey development (d0). Agency members of the initiative cited evaluation as one of the types of learning they acquired through their participation in the initiative (d-er), reporting expanded use of evaluation in their other programs, earlier use of evaluation in program lifecycles, and adoption of methods they had experienced in this DE (d-er). In addition, participants reported learning about theory (e.g., attribution/contribution) and gaining capacity for collaborative data analysis. This last was a point of particular emphasis:

...rigorous methods, valid survey design, interview process, yes that’s important. But there are diminishing returns. And I would rather see a group get really good at shared analysis and the ability to take conclusions from that analysis and let that inform how their work is unfolding... That’s the missing piece for a lot of organizations... a lot of organizations put all their marbles in the data, and don’t invest in building the capacity around analysis and use (c11).

Multiple references in the data point to collaborative work leading to instrumental use of findings, that is, concrete decisions and actions informed by the evaluation. As an example, when data showed that front-line practitioners were underrepresented in the initiative, the team initiated a process of idea-generation with agency members. This was followed by small-scale experimentation and observation to assess effectiveness of the ideas for engaging this group. This process fed improvements to the initiative which facilitated front-line practitioner participation (d0).

Culture shift. Fostering a culture of evaluation or a practice of inquiry was attributed to participation in the evaluation (d/c12, c12, c11). A team member said, “it was very much like culture building” (c12) and explained:

...it shifted our culture... because [DE] is much less about having to prove that I'm doing the right thing or list all the activities that I'm doing and hope that you're happy with them, and much more using information and data to say 'what does this mean to us collectively?' and 'are we on the right track?' (c12)

Developing theory. Accounts of turning points highlight unexpected outcomes being recognized and incorporated into the intervention's theory in this case. For example, the TOC had outlined a path to systems change that began with participation in the initiative by individuals, leading to individual learning. Individuals would then transfer that learning to their organizations, and organizational learning would lead to systems change. However, when the team tested these links in the TOC, the data indicated a problem with this pathway (c11). A participant said that they brought this to the participating agencies: "saying here's what we're trying to figure out... [and] they came back to us with some really useful analysis ... we needed to evolve our expectations and change what we were doing" (c11, also c12). In effect, agency members described the intervention 'shifting power dynamics in the system' through network effects that ran counter to organizational silos and hierarchies, and that these were changing how people worked and seeding new initiatives across organizational boundaries (c12, c11). As a result, in the words of one participant: "we were able to make some really good, well-informed recommendations to the board and... to take some shifts in direction in the years that would come after that" (c12).

Epilogue. The evaluation is credited with supporting a decision to extend the initiative beyond its initial 2-year term, and again at 4 years (d0). After the fourth year, some DE work continued, focused on remaining gaps in understanding (d0). Otherwise the evaluation became more formative in approach. The change was attributed by participants to greater comfort with initiative (c11, c12) which allowed for a shift to a less intensive, refinement-oriented evaluation process (d0).

Case 2. The ‘Game Changer’: Higher Education for Low-Income Children

Background. A community foundation had been investing in local education initiatives for decades, yet continued to witness declining enrollment, graduation rates and test scores (d0, v11, v12). When a study found no correlation between its past investments and student achievement, the foundation reached a turning point (d0, d35, v12). The foundation joined with a local school district to co-develop a new intervention, designed to establish the conditions for student educational success (d0, v12b c23, d6, d5).

The initiative was designed to be “multipronged”, to include health, academic and financial supports, and to foster a ‘college culture’ among students. It was described as “a game changer” for students, families and the entire community (d3, d7, v1, v5). It required a very significant and long-term investment (d0), called: “a huge risk”, in which the foundation was “putting [itself] out there in a way [they] never had before” (c21). In addition, actors were faced with a shifting political, economic and social context for implementation (d0, v12).

The foundation committed to funding an evaluation, but actors doubted that a traditional evaluation approach would be effective (d0, c21). Concerns were linked to the “fluid” strategy they were adopting, the complexity of the initiative, and the “fragile relationships” involved (d0). Actors wanted “an approach to evaluation that would support and inform the initiative as it matured” (d0, c21). A DE team was hired. The team started work the same year the initiative was launched.

Tension and glue: Two interdependent organizations. The foundation and the school district are in a close and interdependent partnership, and this relationship is at the heart of the intervention. However, a participant explained, they were: “two entirely different organizational cultures trying to come together” (c21, also c23, c22). Their differences affected actors’ positions toward the evaluation. A highly political context surrounding the school district meant that actors there tended to view the evaluation with concern about how it would be used (c21; c22). The district’s previous experience with evaluation was that it was “punitive, judgmental, very often finding fault”. This made evaluation a sensitive issue for the partnership (c21). In contrast, the foundation is described as having an adaptive culture and a more secure position from which to embrace evaluation (c21, also c22). Although differences were substantial, a high level of

interdependency had been built into the initiative's design (d38, c22, v8). Data suggest this helped to keep actors at the table over time. Moreover, the DE process was described as a “glue” that advanced the partnership (c21).

Strategies and principles. Some overarching strategies and principles influenced how the evaluation was executed. These included co-creation, responsiveness, raising voice, grounding the work in data, and systems thinking.

Co-creation. The team's “first crisis” (c21) occurred early. After months of direct consultations to develop the evaluation plan, it was rejected by one of the partners (c23, d0, c21). A participant said it was ‘a blank spot for us’ that the newness of the partnership meant ‘there wasn't enough trust in the room’ (c23). In their response, the evaluators were credited for their “facilitative and negotiative skills, being able to hear both sides of the dialogue” in this situation, and willingness to invest in “getting it right” (c21). An intensive two-day working session, described as much more of a co-creative process, helped actors find common ground (c21, c23). The event was said to have “changed the dynamic” in the group from concern that the evaluators were there to “judge what the players were doing” to seeing the evaluation as “in service to us”; and about “stuff we're curious about” (c23). The experience was said to have instilled co-creation as the “mantra” of the program team (c21, also d0, v12).

Being responsive. Being responsive to stakeholder needs involved “iterating the [evaluation] design... and being flexible with it” (c23, also v12, d0), said to be key to maintaining the relevance of the evaluation and to building trust and engagement (c22, d20). Fitting the evaluation to the initiative as it changed (c23) was said to allow the evaluation to go deeper and enable the use of findings (d20). For example, the evaluation questions were revisited every 6 months to ask, ‘is this information still making sense?’ (c22); participants came together annually to: “figure out ‘what are we interested in learning this year?’... never assuming we'll continue activities” (c22).

Raising voice. Bringing multiple perspectives to the evaluation was described as important to reduce uncertainty, since the initiative was unfolding in a complex and changing context (c22). It was also important to “make the findings real” for program actors (c22) and to “model a commitment to co-creation” (d35). Raising voice was said to increase data quality

because it demonstrates respect for informants' needs and perspectives and supports engagement when findings are shared with participants (c22); bringing findings back to the people who are informing the evaluation ensures that they "see their voices being raised up" (c21). A participant elaborated: "Otherwise they're not going to see the value in talking to us. We're just extracting information from them" (c22). The way questions were framed was also important to how people engage: "if you are giving people power, people can see that. Like if you ask... 'what are the things you haven't been able to achieve', instead of asking 'what's your vision for how programs should look like in your school?'" (c22)

Systems thinking. Systems perspectives are evident in how the evaluation was structured and implemented. As examples, systems mapping helped actors broaden their perspective of the initiative. Of note, supporting alignment among actors in various locations in the system was a key role for the evaluation. This role increased with time as a "web" of diverse relationships with organizations throughout the local community was developed, said to be important to "building a movement" around the initiative, but also limiting actors' ability to control it (c21, also d35, v8).

Implementation: helping them have conversations. Data-driven conversations were called a 'primary mechanism' for strategic learning in this case (d0). A participant described the evaluators "trying to come to the table with data to help [stakeholders] have a conversation and make sure that it's a productive conversation" (c22). Formal activities that supported conversations included setting up structures, 'grounding in data' and building capacity and buy-in. Informal activities were also important. Accounts describe positioning and 'giving cover', keeping 'an ear to the ground' and 'walking a fine line', helping to see, and demonstrating a commitment to learning together. Overall, the evaluation supported conversations among stakeholders that shaped the initiative's trajectory (c21, d0).

Setting up structures. The evaluation was structured to create protected space for conversations among participating groups. These groups included the core team from the foundation and school district, joined by the evaluators, a larger primary stakeholder group that were involved in some of the 'nitty-gritty conversations' (c23, also d0), and an advisory group representing the wider community (d0, c22, c23). These groups met regularly face to face to have conversations focused on evaluation data about the intervention. The team structured how feedback was delivered and adapted their approach over time in responsive ways. An example

below illustrates this: the team tailored feedback processes to fit program actors' capacity to absorb and interpret data, so that conversations could be productive.

Data collection and feedback to the program's core development team were initially structured in 3 steps: on-site data collection, a memo two weeks later, and a teleconference to discuss. This was repeated quarterly. The memo provided a 'candid' report of what the evaluators were observing (c22). It noted patterns, identified urgent issues that had been "completely off our radar", as well as less urgent things "that might require a little fix" (c21, also c22). The process was supplemented with periodic 'check-ins' (d0; c21; c23) and formal reports (v12).

This process became overwhelming for some program participants and resulted in pushback (c21, c22), in part because this was a new and unfamiliar way of working (c22). Challenges processing the volume of feedback in addition to their regular workload, trying to 'make sense of it on the fly' and uncertainty about expectations for action produced stress (d0), especially for the school district (c21, also d0). A school district member: "had this sense of anxiety that... [there was an] expectation that everything that needed attention would be checked, sort of 'check, check, check' – that [they] would attend to all of those things" (c21), which led to "a really uncomfortable conversation" (c21).

The team added an in-person, informal 'debrief' after each round of data collection (c21, c23, d0). This helped actors to do some initial processing of the findings (c21). The informality produced less anxiety than the written memo, and: "when the learning memo landed in the inbox it wasn't a shock" (c21), noting that "when something is in writing it feels a little more high-stakes" (c21); the debrief also provided the evaluators with more context. It: "gave the team a chance to sort of assess or say, 'is there something hot button here?' (c21).

The evaluators also became "more selective" about the information being delivered: "to make sure that we were not overwhelming them... [and] paying attention to things that were the most important" (c22, also c23). The memo was reformatted to point directly to what was working well, and what needed attention (d0, c21, c22). Identifying priorities for action was described as a skill on the part of the evaluators. They considered what could be changed in the near term and what would involve a longer-term process (c23, d0; v12). A participant compared

data about a gap in awareness to data on attitudinal barriers. Awareness is a technical problem but changing attitudes takes substantial time and effort as a process (c23). The foundation and district team members also began to hold a regular ‘now what?’ conversation on priorities, separately from the evaluation team; this conversation was less formal (c21). A participant noted that the new process changed the dialogue “from a place of being defensive” and “attacking the data itself”, to “a ‘so what does this mean?’ and ‘now what?’ conversation” (c21).

The new sequence iterated between informal and formal communications in a way that the two could support one another (c22). It reflects the importance of pacing, filtering and framing to align with actors’ capacity (d0). The use of a ‘runway’ was described by a participant as a way of introducing activities in steps, that is, meeting people where they were and staging activities to warm people up or prepare the ground for more intensive or abstract work: “to get to a point where you can have a meaningful conversation” (c23). This participant gave another example of using a runway from systems mapping: “it’s hard for people to just come into a room fresh and just think about a system. Systems are so amorphous and big and there’s lots of things going on” (c23). This participant described having individual conversations and soliciting input on draft visuals: “going down the path a little bit... in terms of the thinking process” for people, before holding a mapping session with a large and diverse group of stakeholders (c23).

Grounding the work in the data. The importance of data was stressed for multiple reasons. The data helped to provide ‘on-track/off-track’ feedback (d0) and to bring more certainty, so that they were “walking on a little bit more solid ground” instead of acting on assumptions (c22). Data were used to discern trends, as a participant explained: “when you are innovating or building something new, sometimes it’s really hard [to see] what are the early wins, the early outcomes” (c22). When findings were controversial, the data helped reduce tension: to be able to go “back to the facts and [say]: ‘this is what we heard, and this is how many people we heard it from’” (c23).

In the first year alone, actors conducted three surveys and gathered interview and focus group data from more than 70 participants (d0). The school district provided administrative data on student progress (e.g., attendance rates) (d23). Data were reported regularly to stakeholder groups and the wider community, including local media (e.g., d3, d23, d31). Local media

reported data that identified areas of focus for improvement (e.g. d28). Regular public reporting of the evaluation data reflected a commitment to transparency and utilization (e.g. d3).

Building capacity and ‘buy-in’. Helping stakeholders have conversations also involved capacity building. A priority was placed on building “capacity among... stakeholders to receive feedback and to use information to drive decision making” (d20, also v12). For example, at stakeholder meetings participants were guided through detailed evaluation findings so that they could make sense of them collaboratively (c21). A participant said: “[t]hose were extremely useful meetings ... where the rubber met the road in terms of... really talking about what actions, and how does this change our plans, and what should we think about for next year” (c23). We “put the stakeholders in a room and had a sensemaking day for the whole day, and the report was informed by the sensemaking” (v12). Another participant described: “giving them a table with a lot of data where they can solve some of the things that were coming up with the program and just make the program better.” This participant said: “the reactions and changes in strategy, that was really meaningful” (c22). These meetings were also said to help reassure stakeholders by increasing the transparency of the evaluation:

...all of the school principals and superintendents and foundation staff... we would review the data then with them and nothing would come out without everyone buying in and everyone just asking questions about the findings and trying to get a little bit of a better understanding of why certain things were happening (c22).

Another participant credited the productivity of the meetings to the practice gained by stakeholders as the meetings became “habitual” over time. It took time because: “just the act of looking at data and thinking about implications is not a natural thing for most people, right? It’s not how most systems work” (c23).

Capacity building included developing understanding of DE (d0, c22, c23, v12). Stakeholders approached DE with interest, caution and skepticism. Some actors were “conditioned to think evaluation has to be a particular way” (v12). The evaluators needed to work to establish understanding of the approach (d0, c23) and develop trust (c21, c22, c23). For instance, the development of evaluation questions was said to have been carefully guided by the evaluators to ensure their ‘nature and tone’ focused on understanding the initiative and how it

was developing rather than on measuring impact (d0). As an example, a participant noted that often actors expect to be able to measure a program's direct impact, for example, on graduation rates. This participant described the need for: "kind of talking people down from that" when the program was at a developmental stage (c23).

A commitment to learning together. Accounts describe the importance of the evaluation's responsiveness and its support to successful program changes, because this demonstrated a commitment to collaborative learning. For example, in the third year, front-line school staff raised concerns about the program, and the evaluators brought their concerns to leadership. Evaluation activities helped actors work together to develop a solution, which involved reconfiguring the program and securing substantial additional resources. The solution was supported by donors with the help of evaluation data (c21). 'Successes' like these were reported to keep actors engaged with the evaluation process because they could see value in it (c21, c22, c23).

Positioning and 'giving cover'. The position of the evaluators played a key role in these conversations. They were understood to be "at the table" with program implementers and "on their side" (c23), but at the same time acting as the "outsiders in the room" (c21). On the one hand, being an 'insider' was important. It involved: "cultural competency and the awareness of power structures" (c22). Another participant explained that: "you are 'on their side', right? [The evaluation is] not to put anybody down, it's not to point fingers, it's not to say 'gotcha'. It's about for everybody to say: 'how do we make this better?'" (c23). This position was important to the DE team's ability to "have difficult conversations... and help move through when there's discomfort" which is "different from someone who's purely in the land of controlled trials, where it's just very scientific" (c21).

At the same time, an 'outsider' position was also useful: "because you have the external status and you can play that card and you can ask the questions you want to ask" (c23). The evaluation team was described as a 'neutral broker' and a credible source of potentially unwelcome information (c21). For example, evaluation results sometimes "validated things that we were [already] thinking... [and] provided a vehicle for those kinds of things to be put forward by a third party where we could say, 'yes'" (c21). A participant described a problematic situation

that risked damaging relationships in the collaboration. The evaluation process brought the issue forward and “gave cover” for actors to address it (c21).

In interviews, evaluation team members referred variably to the initiative as ‘our initiative’ and ‘their initiative’ or to the students as ‘our kids’, for example, even though the evaluators did not live in that community. This corresponds to the ambiguity in their positioning. At the same time, they expressed awareness and deliberation about boundaries in their role. For example, they commented on the importance of protecting the scope of evaluation activities and their own impartiality. In terms of scope, a participant said the evaluator’s job is to provide information, but: “it’s very clear in my mind boundaries about who is responsible for putting in a solution. ... [W]e could go as far as, you know, putting people in the conversation who could say, ‘what do these findings imply for action on your part’ but they still have to figure out what the action is and take that action” (c23). Another participant called it being “a neutral advocate”: “of course I want this program to succeed, but at the same time I’m neutral. I’m going to be rigorous in my data and I’m going to tell you what I’m finding, even if it’s not positive” (c22). Participants were careful to point out that flexibility and responsiveness in their work does not reduce the importance of rigour (c22; d0; c23): “We have evaluation tools that we can use [and] we have to give them justice. We need to be able to stand behind our data for everything we’re saying” (c22).

Keeping an ear to the ground. ‘Keeping their ears to the ground’ (d0, v12) was said to increase the team’s understanding of the local context, to maintain the evaluation’s relevance over time and to build trust (c21, also c22, c23). The evaluators were helped by a trusted advisor living in the community (v12). A ‘check-in’ email to staff every few months elicited answers to a few short questions (d0, v12). They also spent informal time with stakeholders, ‘eating tacos’, participating in community events, and following local media (c22; c23). This helped the team respond when there were sudden changes in context that required a shift in the evaluation’s direction, which a participant said happened: “all the time” (c22). One example was an unexpected “investment boom” in the community, which media attributed to this initiative (d11, d13, d32, v1). Rising housing costs were brought up in evaluation interviews by parents as a serious concern, and this was communicated to the program team (c22). Potential consequences of housing issues for student participation in the intervention became a priority (d23, c22, v12).

Data collection was extended to cover the issue, which led to funding support for low income housing; described as a “big turning point” in which the evaluation “provided the initial solid grounding for [actors] to start making decisions” (c22).

Walking a fine line. Conversations sometimes needed skilled navigation, as explained by a participant:

...an important role an evaluator plays is a little bit like a mediator or – I think you have to be really aware of power dynamics – and you need to be really careful in the way in which things are communicated and how you are asking the different questions (c22).

Critique of “pieces that people hold dear” can provoke defensiveness, and “sometimes people are more ready to hear some messages than others” (c23). One actor said that navigating this “had a lot to do with how the evaluator is able to convey information”. They explained: “the message you want to convey can get caught up in peoples’ mental models, defensive stances... some of those conversations were trickier than others, and we just kind of had to work through it” (c23).

The evaluation’s scope was restricted from the start to not include the partners’ internal operations. Bounding the DE this way was challenging because the work of the partners and the work of the initiative were so interconnected; as described by a participant: “we had to be very careful, we were always walking a really fine line” (c22). Most comments on this theme related to delivery of feedback, but ‘walking a fine line’ was part of other activities such as how questions were constructed (c23) and being attuned to organizations’ capacity limits for data collection (c22).

Helping to see (differently). The evaluation was credited by program actors with ‘opening our eyes’ to how complex the problems were, and how actors could approach problems differently (d0 - c21). For example, some stakeholders were said to be: “very much head-down, putting out fires day to day” (c21) and the systems mapping helped: “to kind of get their head up and see how... what we’re trying to pull off here... is not just about what’s happening right in front of you right now” (c21, also d0). The map was described as a “medium that [people] can interact around as a group” that “creates a different kind of understanding” (c23). This was described as ongoing work: “sometimes clients [are] caught in the day to day operation of the initiative... what we’re trying to do is to continuously say: ‘... remember that this where we’re

going” (c22). The data collected through the evaluation helped identify patterns that are difficult to see, and to help stakeholders see the intervention from multiple perspectives, for example, “how teachers are perceiving different aspects of the program, [and] administrators... how are different groups and different activities getting to full implementation” (c22). Actors used multiple methods to access and incorporate different perspectives, including creating structures to support inclusion (d0) such as the development of a network to engage parents who were harder to reach (c21, d23).

‘Helping to see’ also involved clarifying constructs. This was described as important because fuzzy concepts are difficult for actors to “see”. A participant offered the example of the aim to create a “college culture”: “for a traditional evaluation, you just need to facilitate a conversation so you can have a better understanding of what they’re doing, but in this case ... *they* needed to have a better understanding of what that meant” (c22). This participant explained that once a shared understanding took shape, they were able to move on to concrete discussions about implementation, for example the supports that would be needed and the people who would need to be involved, to create a college culture (c22, also c21).

The evaluation was also said to ‘provide signals’ about trajectory and raise awareness of “issues that were off the radar” (d0). For example, data indicating low parent awareness of the program were brought forward and actors responded in time to meet a critical program milestone (d, c21, c23). Another example had to do with an equity issue related to the program design: “when we started surfacing the data... [it] became a big issue, the... foundation took action and developed a solution for students” (c22). In this instance, the evaluation was soliciting input from key actors who were still upstream of the program implementation, which meant the team could respond with design changes for that phase before adverse effects were felt (c22, c21).

Learning outcomes. Actors described outcomes of the evaluation that included development of norms around data use and changes in perspectives about evaluation, changes in understanding about the intervention, program changes, and sustained public support. Participants also describe learning the need for patience.

Norms. A participant spoke about establishing norms in the group that created space for conversations. Although some actors were thought to be still operating in ‘compliance mode’ in

other areas of their professional lives, for this collaboration: “I think what we got to was at least setting a norm for the conversations we had with them... having a conversation about implications based on data in a productive non-defensive way, we managed to carve out a space to have that” (c23).

Accounts credit the evaluation with strengthening working relationships and increasing capacity to use evaluation (d0, c21, v12). Participants mentioned changing team members’ view of evaluation, described as a: “slow gradual shifting... moving from being really sort of fearful and apprehensive about this process to embracing it” (c21). This participant added that this change was because of the DE process, not its findings, and elaborated that the “relational part” of the DE: “benefited our work together and our ability to work together” in the collaboration (c21). Asked if the evaluation was able to increase trust, a participant was definite but said that it took a lot of time (c22). Another participant agreed, saying it took 2 or 3 years to shift how people responded to data, from asking: “‘is this going to be in the newspaper tomorrow?’” to “‘I’m looking forward to the insights from your visit’” (c23).

Understanding the intervention. The evaluation was said to have changed how actors conceptualized the intervention, for example by clarifying the meaning of ‘college culture’ (d0, c22), and influencing actors’ strategic thinking about the program. This in turn was said to have influenced decisions and actions (d0, also c21). Examples of actions are noted above, such as the program reconfiguration (d17, d19, d23, d28, c22, c21); investment in affordable housing (d26, c22) and allocating extra support to students whose progress data indicated need (d23, d20, v4). A participant said: “we’ve had incredible findings that have been harvested from the DE that have absolutely shifted how we’re doing our work” (c21).

Public support. The ability of the program development team to garner public support is visible in the sizeable funds raised for the program over time (‘exceeding expectations’), the team’s ability to make program changes that required the support of other organizations, and the consistently positive tone of media reports, even when data suggested slow progress on some indicators (v9). Data were integral to this. Public communications about the intervention released by program partners highlight the evaluation work taking place (d3, d9, d23), the local media report progress data (d31, d8) as do other documents from external organizations (e.g., d35), indicating it was widely shared. Data helped secure additional funding to make program changes.

A program participant stated: “we were able to point to the evaluation findings to say ‘look’ there’s data to support this... If you want to talk about a win, that was huge, huge” (c21).

Learning patience. Constraints on the evaluation and the use of findings did exist, relating to issues of trust and political operating environments of some actors. Where evaluation findings suggested changes, some were actioned more than others, related in part to the sphere of operational control in which change was indicated, differences in actors’ capacity to act, and because of the ‘dynamics of the situation’ (c23). The three interview participants expressed some regret that faster progress could not be made. They mentioned structures, budget constraints, and other less tangible forms of capacity; “I feel we would have gone so much farther, you can only go as fast as – this is a huge leap for [some]” (c21, also c22). DE was described as sometimes ‘intrusive’ (c22). Some actors were described as pushed ‘outside their comfort zone’ (c21), and it was recognized that different people need to travel different distances to engage with DE (c23). “I think it takes exceptional leaders and different thinkers to break out of it, and it takes time to build trust” (c23). “[I]t teaches you a little about patience and just some appreciation for just giving yourself some grace, I guess, and to others, and a little bit of wisdom around when to push, what to push around” (c23, also v12).

This program brought funding to area schools and is credited with increasing enrollment in schools that were previously experiencing decline. Increased enrollment translated to increased government funding (v11). These benefits constitute strong motivations for partners to support the initiative, but they do not entirely explain their engagement in the evaluation or how their engagement was reported to have increased over time. Arguably, the data suggest a broad pattern in which the evaluators intentionally built capacity for data use through productive conversations. The actions taken were seen as successes. The successes were instrumental to strengthening the collaboration, in turn enabling more productive conversations.

Epilogue. This evaluation was strongly developmental in Years 1 and 2 (c23). After Year 2, it was ‘narrowed down’ as the initiative stabilized, and stakeholder interests changed (c22, c23, d). In Year 3, the evaluation took a hybrid developmental / formative character (c22; c23). By Year 4 the model ‘had sort of come together’ and the formative aspect became primary (c23). The evaluation is ongoing and projected to “remain a key source of learning” for the initiative for the future (d20).

Case 3. Transforming Cities through the Arts

Background. A foundation believed that engagement with the arts can positively affect how people relate to each other and to their city, and that this in turn can contribute to positive systemic change. It launched a program of arts-related projects over a period of 5 years in a small city in economic and social ‘transition’. The foundation collaborated with local and non-local artists, non-profit organizations, and other funding agencies.

The foundation was only about 3 years old at the time, but it had a clearly articulated statement of purpose and a set of principles in which inquiry and complexity thinking were prominently featured (d35, c31, c32, d0). The foundation also had a pre-existing formal mechanism for internal learning. This included annual 2-day retreats for staff, board, advisors and guests to undertake ‘deep analysis’ and reflection on the foundation’s work (c31). However, actors were skeptical about evaluation (c31, c32). The foundation had no prior experience with evaluation. Advisors to the organization were familiar with DE and they suggested it to the foundation as a way to further the foundation’s learning (c32, c31) and ‘tell the story about how the initiative could have impact’ (c31). Foundation actors were interested in developing their TOC (c32), and DE’s iterative approach seemed to align with the experimental nature of the initiative (d0, c31). This initiative involved other funders who had evaluation requirements; this was thought to act as ‘a bit of a catalyst’ to try evaluation (c32).

Strategies and principles. Some broad purposes informed the evaluation strategy. These included prioritizing a learning agenda, integrating evaluation in the projects, and navigating a tension between evaluation and the arts.

Privileging a learning agenda. The program was designed to generate learning for the foundation as well as benefits for the city. The foundation aimed to ‘connect with the city’s transition’ (c31, also d0, c32) to learn if, by engaging in these projects, “does that help arts flourish or help a place know itself better, and what does that lead to?” (c32). The program was expressly positioned as an experiment to develop an innovative approach (d, c31) with risk and failure a necessary part of learning (d35). There were no specific targets for outcomes (d0). This position was thought to keep stakes low, and ‘create space for experimentation and adaptation’ (d0).

The foundation's aim "to figure out" the role of the arts for individual growth and social transformation (c31), included "discovering" the foundation's "evolving role in part through doing" (d35). The experiential learning agenda is a central theme. It connects to the foundation's direct involvement in programming. As described by a participant:

...we needed to be working on the ground, we needed to have our own stories to tell, we needed to be able to articulate an idea of how we thought that the arts could be transformative and how it could connect to interventions in problems in society, before we could become more of a voice for that (c31).

Although the initiative was place-based, learning was expected to have relevance elsewhere (d31).

Integrating evaluation. Actors structured the evaluation to be deeply participatory (d0). Staff and collaborating artists were involved in the evaluation design, data collection and interpretation (d0; c32). The consultant was described as a 'coach' and was "embedded in the projects with the artists" (c31). The intent was for the evaluator to be "inside of the project" (c31) so that the evaluation:

...was part of and not separate from the artistic experiences. So, it wasn't this kind of like, weight on the side, it was 'how do we creatively try and integrate collecting information and doing analysis with people, often in a very participatory way, as part of the experience?' (c32)

This was motivated in part by a concern that "too much evaluation that didn't feel like part of the [artistic] experience would actually be compromising to the experience" (c32). The consultant worked collaboratively with the artists to make the evaluation a part of their projects (c32).

Each of the funded artistic projects in this initiative were unique. This was intentional to provide a variety of experiences from which the foundation might learn (d0); they were designed as "experiments into the transformative power of the arts" (d31, also d06), with the city seen as a 'laboratory' (d30, also d06). This meant there was minimal intention for development at the level of the artistic projects. Rather, although the DE was positioned in the projects, its purpose was to

help answer higher-level questions relating to organizational role, theory of change, and knowledge generation about the arts and social transformation generally (c31, c32).

In addition, each of the arts projects was intended to have: “a different DE methodology to create a kind of array of possibility” for using DE (c32). In effect, actors aimed to experiment with DE in tandem with experimenting with the arts projects; to explore a variety of ways in which DE might be applied (c32, d0). This connected with interests of a collaborating funder to potentially develop a set of tools for evaluation with artistic projects (c31).

Navigating tension between evaluation and the arts. The evaluation approach was guided by a view that some forms of systematic inquiry are incompatible with artistic endeavours. For example, describing core principles, an early foundation document asserts:

...artistic creativity embodies values and attributes diametrically opposed to the narrower concepts of efficiency and rationality that have contributed to the modern crisis, concepts reflected in the prominence of words like objectivity, calculation, measurable, predictable, quantifiable, replicable... rational, and linear (d35).

The implication is that arts outcomes defy measurement through approaches and assumptions common to the field of evaluation (d0). A participant explained: “there’s been this struggle in the field, in the sector for a long time, how do you evaluate arts organizations and artistic projects? What we’re trying to say is it cannot be the superficial measures of time and money spent. It needs to be about how meaningful those experiences are as well” (c31).

The foundation was willing to experiment with DE provided that the evaluation could be sensitive to and aligned with artistic endeavours. This was about avoiding alienating artists and participants. It was also so that the evaluation could generate meaningful insight for an organization with a well-developed vision of art and the role of art in society (c32).

Implementation: stretching the boundary of DE. Actors described ‘stretching the boundary of DE’ (c31, also c32) in creative ways, to find a fit for the projects, the program, and the foundation’s interests. Tailoring to the projects, engaging participants and crossing levels of inquiry led to opportunities and challenges related to stretching this boundary.

Tailoring to Each Project. The evaluation was carried out over five projects. The variety of methods used and how they were applied speaks to an intensive tailoring effort. In one project, for example, the evaluator collaborated with artists to plan a one-day event at several sites in the city. Opportunities were built into the event's structure so that attendees would gather together over the course of the day and share their experience through interviews, drawing and storywriting (d0, d15). This was described as a way to "layer evaluation into what was happening in a way that would enrich the program experience for participants" (d0), to facilitate the 'sensemaking experience' of participants as they engaged with the event, as well as to 'capture information and perspectives' for learning by the organization (c31, c32, d0).

Although each project's evaluation was tailored, there were a few common features. For example, each project evaluation involved documenting the experience to generate learning (d0, d35), engaging participants at the project level and crossing levels of inquiry.

Engaging actors at multiple levels. An actor described the process of working alongside the artists in each project to design an embedded evaluation (c32). For community members, participation involved: "creating the right questions and ... creating the right environment in which those people can feel comfortable contributing" (c31). For instance, providing opportunities for people to communicate their experience through various modes: conversation, drawing, video, etc. (c31, d0).

The foundation's directors and advisors were not as engaged in the evaluation as the artists and staff. Early on, the consultant introduced DE to these stakeholders at an annual retreat and provided input based on the data (c32). In later years, however, reporting at the retreats was handled by staff leadership (c31). As a small organization, the foundation was said to be able to use less formal communication flows (d0) and as noted above, the foundation had a pre-existing structure for reflecting on its work, which was already supported by staff briefings on program activities (c32). Guest participants were routinely invited to the retreats for the foundation to: "get different perspectives" (c32). Overall, the evaluation was just one of a number of inputs for learning, mediated largely by staff (d0, d31).

Two other funding agencies were involved in the initiative; however, this DE was directed at supporting the foundation and its scope was bounded there (c32). Although one of the

funding agencies “had very different evaluation needs” (c32) and a different understanding of the initiative (c31), the tensions resulting from these differences were negotiated and addressed by foundation staff, separately from the DE process (c32, c31).

Negotiating differences. A number of the turning points in this case related to challenges in collaborative work. Negotiating differences in perspectives and organizational cultures is difficult, but important for this foundation, which describes itself as a “convenor” and a “broker”, that intends to “take a transdisciplinary, cross-boundary, cross-sectoral and inter-generational approach to its work” (d35, also e.g. d27) with “capacity to be mutable and to build bridges” (d30).

Collaboration was described as easiest and most successful with the artists involved (c31, c32) but more difficult elsewhere. The foundation was new to the city and found it difficult to gain traction among local organizations. This was thought to relate to the initial collaboration approach and aspects of local culture (c31, d0, d33, d07). The foundation struggled to find common ground with a partnering funding agency (e.g. d07). A participant noted that the complexity of this work is challenging to communicate to those working in other contexts and with different underlying assumptions (c31). This foundation is described as a ‘very unique’ organization, with a firm commitment to honoring its purpose, principles and world view in its work (d30, also c31, c32, d25). Ultimately, the foundation adopted an approach to collaboration in which they encouraged people who identified with the foundation’s mandate and “way of working” to come forward and self-organize around their community’s needs (c31).

Within the evaluation, the consultant, artists and foundation staff collaborated intensively to adapt evaluative inquiry to the projects and to the foundation’s culture. A participant noted the importance for evaluation work to adapt to “an organization’s culture and how it thinks and what it likes” (c32). For example, in early attempts to create a TOC, actors tried the usual visual format, but: “visual diagrams failed to capture the nuances of the organization's thinking and experiences” (d0, also d32). The foundation was said to “really struggle with this format” (c32). Actors developed a long-form (essay-style) narrative, written collaboratively. This was described as “very unique” for a TOC (c32). Collaboration was also focused on finding common ground between evaluation's strength in serving operational decision-making and the foundation's other interest in answering more theoretical questions.

Crossing levels of inquiry. The evaluation was pursuing multiple purposes: at the operational level (i.e., for program and organizational use) and at a more theoretical level (i.e., about the role of the arts in society). At an operational level, efforts were made to bring lessons forward from individual projects to inform the foundation's programming approach (d0). For example, staff leadership were interviewed multiple times over the 3 years as a way to facilitate cross-project reflection and 'sensemaking' (c31). The interviews were described as promoting 'dialogic thinking', which brought out new ideas for programming (c31).

New operational questions arose over time and were taken up by the evaluation, for example, when the foundation encountered challenges with local-level collaborations. Efforts to raise levels of engagement among pre-selected agencies were 'burning up a lot of energy' (c32). The foundation opened invitations to a broader range of actors, to work with whoever came forward with interest and willingness to engage (c32, also d0, d33). The evaluation provided feedback on the efficacy of these relationships that supported a shift to a more emergent approach to collaboration for the long term (c32, also d33).

The balance between operational and theoretical questions (d0) appears to have shifted over time. For example, the earliest evaluation work looked at aspects of project design and project outcomes. Later, focus turned to collaboration approaches (d0). Midway through the 3 years, questions began to shift toward the second purpose, for example: "to gain a better understanding of how the arts can be meaningful" (d0) the creative process of the artists (d0), and "the contemplative experience of creation for the artist" (d0). Evaluation activities were scaled back in the final two projects, attributed to experimentation with smaller scale DE designs (d0). At the same time, the evaluation was moving into the domain of research (c31, c32).

The "rhetoric of research" was used throughout the period (c31) and is evident in documents in which the DE was referred to as research and the consultant referred to as "our researcher" (e.g., d15). Framing the work this way was thought to be important to the artists' participation (c31). As described by a participant:

If [the artists] understood it as us bringing an evaluator in so that we could rethink our organization or redesign our organization, they would probably have been less interested in it. They were interested in the idea that it was research into how people will make meaning

of their work that they were experiencing, so they were very open to having an evaluator participating in the conversations... It wasn't that we were trying to evaluate whether their project was 'good' or 'bad' or should have been done differently, it was more research into the meaningfulness of that project, what it means to participate in that work (c31).

Stretching the boundary of DE. Participants recognized that questions of interest to the foundation were leading the evaluation toward research (c32, d0, c31). A participant described the ability to bring the two "in relation" as one of the successes of this case (c31). However, two challenges emerged from stretching the boundary. First, there was a tension in trying to fulfill both purposes. The foundation was said to be "closer to the research end of the spectrum in their interests and needs" (c32), which was echoed by another participant who described the foundation being primarily interested in "the big questions" about the meaningfulness of the arts experience (c31, also d31, d30). However, the evaluation tended to lean to operational questions, such as: "how this was working for [the foundation] as an organization" (c31). This tendency was attributed in part to organizational development being the usual 'mode' for evaluation (c31). The foundation's desire to "think and work at more of a cultural or systems level" and to answer these larger questions, was said to be challenging for evaluation (c32, also c31). This was described as a push and pull in areas of interest, and a 'tension', that in some ways limited what they achieved (c31).

Ironically, an actor also described it as a challenge to link what was learned back to applications for the organization, such as 'bringing it all together in a final analysis' for program and foundation use (c31). The 'dialogic thinking' and reflective work that took place via interviews during the evaluation process was one activity that helped staff to make that link, described as a: "process of questioning" that brought up ideas for staff; that 'forced them into a frame of thinking' that led to developments in programming (c31).

Learning outcomes. The evaluation ended after 3 years, and the foundation withdrew from this city about a year later. Although the foundation chose not to continue with evaluation or this initiative, accounts describe learning outcomes from the experience (c31, c32, d0). As noted above, the interviews with staff were said to have used a 'depth of questioning' that brought out ideas and 'prompting a particular way of thinking' which was described as one of the 'intangible things' that the evaluation added (c31). More tangibly, the foundation made concrete

changes to programming elsewhere, for example adopting a less ‘centralized’ way of ‘driving’ projects and moving to “building a local capacity” in other cities (c31). As one actor explained:

...that shift from more centralized approaches to decision making... to more of a platform-based approach... That is a result of this evaluation largely. [It] showed us that when people take more ownership and have more of a role in determining what is happening in their place, it has more meaning, it has more power for them (c31).

Actors were described as ‘really keen’ about developing the theory of change (c32), and articulations of the foundation’s theory are evident in its annual reports and other documents over the period of this case. The evaluation was said to contribute to “an improved conceptual understanding and an increasingly clarified change model” (d0), said to have evolved out of “[t]he critical thinking, engagement and careful observation of its activities” (d0).

Rethinking its approach. The foundation had hoped the DE would help them: “tell the story of how the arts could be transformative” (c31). In the end, a ‘story the DE told’ was that the foundation needed to engage differently with people, in other words, to rethink their program design (c31, also d0, c32, d30). A participant said, “the experience of doing the evaluation taught us that we needed to take people on a much, much longer trip with us before they could articulate more clearly” about their experience (c31). Described as a “logical development” from the DE, program actors began considering capacity building and pedagogical tools (d30).

Epilogue. The foundation transitioned from DE to a program of action research in other cities that reflected the above shifts and was described as an outcome of the evaluation work (d32, d30, also d20). Ultimately, however, the foundation did not see a need for evaluation. The foundation wanted to go deeply into the meaning of arts experiences (c31), and there was less interest in ‘purposeful iteration and experimentation’ at an operational level – where the strengths of DE lie (c32, c31). In terms of evaluation to measure impact, a foundation document explains:

At times we glimpse evidence of the transformative power of the arts—we find great stories to point to—but measurement as a form of evaluation is a social construct. We need to create an alternate vocabulary and value system with which to discuss the immeasurable

impact of the arts... We must talk about the arts in ways that aren't reductive... The value of the arts is not in answering questions, but in inviting people to ask them (d32, also d27).

This foundation was not required to meet external accountability demands through evaluation, nor did it see a need for evaluation in its internal learning process. It had independence in funding, planning and implementing its interventions, and overall few constraints on its forward journey. The intended outcomes remain open as it: “strive[s] to create the conditions for everyone to experience the possible” (d15).

Case 4. A System for Agricultural Innovation

Background. A foundation seeks to improve food security in developing countries through funding and capacity building support to agricultural research projects. Their initiative operates at three levels: the overall program (global), regional communities of practice (multinational), and individual projects (local). After some 20 years in operation, the program reached an “inflection point” (c42, d0). A leadership restructuring occurred around the same time as an influx of funds doubled the program’s scope (d0, d05, d29, c42, c41). New people brought new ideas to the program, including “what the program should be about and how it should be implemented” (d0).

The foundation is described as open to innovation (c42, d0), with a long-standing commitment to learning from people who are “working on the ground” (d0, also d13, c42). It cultivates long-term relationships with grantees with a view that impacts may take substantial time to accrue (c41, d05) and it has a commitment to accumulating a knowledge base to support its work (d13, c41). Until this point, however, evaluation at the foundation had been limited (c41, d29, d0, d05). Some actors were wary of evaluation. There were concerns about reporting being onerous for grantees (c41). Some were uncomfortable with the idea of evaluation and/or had negative past experiences with it (d29). Finally, it was unclear how to balance aggregated learning with a deep commitment to “place-based knowledge” and “decentralized innovation” (d0, also d41, d05, d29).

The influx of funds came via a grant from another foundation. This introduced a commitment to share data on program activities and results (d29; c42), which encouraged the foundation to go more “data deep” in its evaluation efforts (c42, d35). A new program director was hired. They were chosen in part for having an evaluation background (d05, d41, d42), which ensured a prominent role for evaluation in the program (d29). The director brought experience in participatory, utilization-focused and systems approaches (c42, d29).

Strategies and principles. The evaluation approach was underpinned by systems and complexity thinking, commitments to utilization and to collaboration.

Systems thinking. The program was changing rapidly. This and other conditions in the program reinforced the director’s view that systems thinking would be useful:

...[a]s I settled into the new job at the foundation, I clearly saw that all the transition in [the program] called for a systems-based, emergence-oriented approach. The program was changing even as we were trying to design the evaluation; it was a bit like building a bicycle while you are riding it” (d0, see also d05, c41).

The director recruited consultants with a background in complex systems approaches (c42, c41, d0). The team chose an adaptive action framework (d05, c42, d07, d29), which they felt would help establish greater coherence and a process for data-based decision-making and could respond to a context that was very dynamic (d0). The team proposed six-month cycles of data collection, analysis and interpretation supported by structured reflective practice (d0, d07, d05, c42, d29). The framework was to operate at all three program levels (d0, d29), “with lower level results feeding into upper level conversations” (d0).

Prioritizing utilization. The overall intent was to design a system to support the program over a long term with data and evaluative thinking (d0). From the start, the foundation held the view that: “the only evaluation system that is worthwhile is one that is actually used” (d30), and case accounts consistently depict a utilization focus. As an example, actors assert that less data, if ready in time and in a useful form, is better than collecting more or higher quality data but being unable to analyze it in a timely way to foster use (d0, d30).

Getting there collectively. The team had a strong commitment to stakeholder participation for both methodological and ethical reasons. The program was viewed as complex, and from this perspective, multiple viewpoints from different parts of the system would be important to understanding (d07, c42). Accounts assert that participation by stakeholders increases ownership of the evaluation, and therefore improves quality of data (d30) and its interpretation (c42, also d07). There was a strong sense of interdependency (d29, d07). For example, as noted by a member of the foundation, the evaluation depended on good “evaluative relationships” with grantees. Without good relationships, “we’ll have less certainty and trust in the information we receive... We’re dependent on the groups that we support” (d30).

Added to this, a document asserts that participation, “has an ethical dimension, because people have the right to be consulted and involved in activities conducted on their behalf” (d07): a participant explained: “we are committed to co-creation, and that means we all bring something to the table to help understand better... it’s also a commitment to honouring all different forms of knowledge” (c42). These commitments took shape in extensive efforts to collaborate and share results across program levels, including with grantees (d29).

Getting started: making a paradigm shift. The proposed evaluation approach was unfamiliar to many program actors, who represented a range of perspectives and inquiry cultures, especially from disciplines in agricultural research (d0, c41, c42, d05). In addition, at the leadership level, differences of opinion persisted about the appropriate form and function of evaluation (d0). The team faced resistance when they tried to implement the evaluation. They responded by changing tactics to build stakeholder capacity, establish trust and enable a ‘paradigm shift’.

Resistance. The team piloted the chosen framework and the idea of reflective practice sessions at a regional meeting. This did not meet participants’ expectations for evaluation (d0, d41) and the language being used lacked credibility. It was heard as “empty jargon” (d0). Pushback was attributed to distance between the proposed approach and the existing inquiry culture, combined with a lack of trust in the evaluation team (d0). Not only was the team new to the program, the lead evaluator did not have an agriculture background. Consequently, “those who had lived, eaten, and breathed agriculture over the course of long careers were skeptical” (d0, also c41).

The team took the approach to a leadership meeting a few months later (d29). The response was cool, again attributed to unfamiliar language that lacked currency, and discordance between the approach and expectations for evaluation (d29). Participants at the meeting also raised concerns about inadequate consultation in the evaluation's preliminary program review work (d0). Moreover, the evaluation's goal to build coherence and alignment among program levels was not a universally accepted priority. An account states:

...[w]hile the foundation felt a need for greater coherence, the members of the regional teams who had been working on the ground for the past few years did not feel the same need. For them, the program was working fine the way it was; why were we stirring the pot...what made us think we knew better than they? (d0).

Regional autonomy in the program was highly valued. An effort to build coherence was a concern if it would limit that flexibility (c41). The team had been working intensively for months, and this experience was described as a “demoralizing” setback for them (d0, c41).

Restructuring, changing tactics. The commitment to utilization and collaboration meant that the evaluation had to happen ‘with’, not ‘to’ program stakeholders. Also, accounts stress the importance of mutual trust and credibility “across the program” in order for an approach like this to work (d0). Therefore, although the team was convinced that the approach fit the program's complexity (d0) they began to adapt it (d0, d29), while seeking ways to build underlying support for its essential elements. The team believed that developing such support would require “a paradigm shift” in the program (d0). To enable this shift, they increased their focus on building capacity for systems thinking, providing opportunities for actors to try out elements of the approach, and building trust. These three levels of activity were intertwined.

Capacity for systems thinking. The team worked to help actors ‘expand their frame’ of thinking about the program at all levels. At project levels, for example, actors were encouraged to stretch understanding of the scope of local research studies to consider the local context in which findings were to be used, as part of the research process (d0). A document provides an example of research leading to a more nutritious crop variety that has an unintended effect of increasing milling costs, which undermines the envisioned use and benefit of the crop for local families (d29). It was noted that researchers were not usually trained to consider the systemic

context of their research (d29, also c41). A participant said: “we spent a lot of the first few years really pushing this idea that they had to be thinking more systemically” (c41, also d07, d30).

Using tools. The team developed tools to be used in common by actors in different program areas to improve alignment and anchor systems thinking in the program (d0). The development of common tools and frameworks were a central feature of the early evaluation. Some were used across the program as ‘boundary objects’ for multilevel communication. The team took to heart that the way tools were communicated was critical. An account states: “[n]aming something can be empowering; however, it can be alienating if names used don't resonate with people... [who then] 'throw the baby out with the bathwater’” (d0, also c42). When developing tools, the team struggled with the desire to ‘embrace the heterogeneity in the program while building coherence’ (d0). The challenge was described as ‘systematizing a dynamic process without losing its heart and soul’ (d0).

There are a several examples in this case of the important role of tools to advancing the evaluation and the care taken by the team to develop them. Accounts of developing a theory of change (TOC) provide one example. TOC development was a mechanism to engage the foundation’s leadership with the evaluation in a productive way following the difficult introduction of the initial plan. The TOC was intended to promote “shared meaning making” among actors, for example by clarifying how components of the program interconnected and why (d24, also d11, d29). Developing the TOC was called: “a transformative moment, when, for the first time, the entire leadership team wrestled with the what, why and how of [the program]” (d0, also d29). It provided an opportunity to think about the program with a systems lens (d29) and contributed to the coherence objective (d0). It was also aimed at promoting learning and collective action (d29).

Collaborative process + concrete tool. A document describes the TOC development as “a conversation that took time, inspired passionate discussion, and even caused some disagreement” in part because multiple perspectives needed to be negotiated while new relationships were being formed, including multiple “scientific disciplines” brought together in the group (d05). This was echoed by a participant, who described resistance in part because of diverse perspectives and low levels of trust: “you move at the speed of trust they say. We were just in those early stages of building trust” (c42). Ultimately, the program TOC was described as

a breakthrough because of both the collaborative process and the concrete result: “really getting people engaged, even when there was sometimes pushback... that process and then having something in hand, give a lot of traction” (c41). The TOC became central to the evaluation (d29, c41, c42, d0, d14, d23, d24, d25, d33). It was tested and refined as conditions changed (c41), and commitment increased over time: “Now people love the theory of change... It’s interesting because there was a lot of pushback initially on it. The theory of change has been really important to ground [work] in at all levels” (c41). It was described as the biggest single initiative giving the evaluation traction (c41, c42).

The team later introduced the idea of project-level TOCs to grantees. The response was described as ‘polite confusion’, and sometimes “stormy” (d0). The team arranged for capacity building support and encouraged the foundation to mandate compliance. This was based on an expectation that actors would value TOCs with practice: “a lot of just forging ahead and believing in it until people could grab on... In the beginning people just did it because they had to, and now I think they actually see value in it” (c42). Grantees were required to develop a TOC at the start of their project, to develop a linked monitoring and evaluation (M&E) plan, and to refine it over time (d07, d03).

It took years of active support to ‘build the habit’ of using the TOCs as an input to decision making (c42, d0). The challenge to foster use of the TOCs at the project level is attributed in part to the ‘institutional environment’ within which many of the grantees work. A foundation actor is quoted as saying: “Evaluation is not [usually] something that’s really about reflecting ... It’s more about, “Oh, I said I was going to reach 5,000 farmers, can I check that box?”, because the focus is on “doing what they said they were going to do” (d29). This was said to relate to focus on the ‘doing’ of research without “investing our time and resources in learning how they are doing, and what the results they’re getting mean” (d29).

Use: thinking, centring and acting. Accounts describe the TOCs being used in multiple ways, including to foster collective debate and understanding. For example, project-level theories of change were called “a springboard for conversations” among team members (d07, also d22, d26). Sometimes project members: “show up at the inception meeting without knowing what is written in the proposal... [And i]t is important for collaborative partners to be on the same page with what the project is going to do and intend to achieve” (d29). The foundation’s support

teams are also said to use the TOCs to help grantees “identify neglected assumptions” referred to by some as the “miracles” in the project’s theory (d29, d07).

The program TOC, together with reflective practice, was said to have helped to build agreement on program scope and provide a basis for testing new ideas. This was important because of the diversity of actors involved in a highly participatory leadership process. For example, the TOC was said to help “bound the creative energy” of the leadership team, so that ideas introduced by its members “remained somewhat within the agreed-upon scope” of the program (d0). The TOC has acted in this way as a centering tool.

The program TOC was also described as a tool for enabling action (c42). As an example, a framework to guide research priorities was “formalized” by embedding it the TOC (d05). A foundation actor is quoted as saying that this: “helped to translate the now defined framework into an action plan” (d05). Some of actions were difficult for actors and needed active support. For example, the new priorities framework had the implication of ending funding to some long-term grantees (d07), a decision which seemed to pull against the foundation’s commitment to long-term relationships (c41). This was said to have been uncomfortable for people, taking a long time to work through. The evaluation team supported actors to navigate this by asking questions and helping to clarify the rationale behind the process for making decisions (c41), providing a mechanism for critical reflection to keep the program on track.

Establishing evaluation as part of the work. It took 5 years for the evaluation system to become, finally, ‘no longer a debatable thing’ (d29, also c41). The team sustained their efforts to integrate the evaluation into program activities and structures and make it part of the work of the program. In doing so, their focus on capacity building continued while they navigated the role of the evaluator and tailored the work to context.

Building capacity through integration. The team saw integration as important to build capacity for evaluation, and especially for use of evaluation data. A participant stated:

some people are going to be more adept at evaluation skills, and that’s fine there’s a place for that, but for people to hold more of a stance of inquiry and be able to make meaning and have that inform next steps is a skill everybody should have (c42).

Another participant explained: “every meeting was an evaluative meeting. It was an evaluation moment... It’s this idea that it has to be integrated into the work that they do, and they have to become facile in the way that they do that” (c41). Meetings at various levels were used as opportunities to introduce and discuss evaluation concepts (d0, d29). The foundation also invested in building staff capacity (through consultants) to identify and ask good evaluative questions, as part of building a ‘culture of inquiry’. Notably, having well-constructed questions was said to be key to focusing and avoiding ‘information overload’ (d07). Accounts also highlighted that the way that questions are framed can change the perspective of the work in important ways. The difference between questions like “Did we make the mark?” and ‘How can we learn in order to adapt, so we can make the difference that is needed?’ (d34) can lead to very different inquiries. A participant summed it up their approach to capacity building for evaluation use:

...evaluative thinking... if you break it down [is] just a way of working... In order to do that you have to have skills, right? You can have simple reflective skills, but you have to know how much data to get, which data, how to make meaning with it. It’s really the notion of embedding the what?, so what?, now what? at all levels, and what are the skills and tools to do that well? (c42)

Evaluator roles. Evaluation functions were shared by various actors in the program, with members of the evaluation team providing support and facilitation (d29, d0). External evaluation consultants were contracted from time to time to look deeply into specific questions, particularly summative questions (c42 see also d05, d17, d20, d30, d34) when impartiality of the embedded team and collaborating program actors might be an issue. The lead evaluator’s close work with the program made them very “involved in helping to shape how things are happening” (c41). Contracting for summative questions was described as a combination of ensuring impartiality in ‘real’ terms, and in the minds of stakeholders: “it’s the idea of being able to recognize your biases. There’s the reality of that, and then there’s the perception of that, and you have to take both into consideration”; “it just feels better to have someone come in from the outside and look to see how well those things are working” (c41).

However, the use of specialized evaluation consultants did not work well when they were hired to provide ongoing support to grantees. Local evaluators were hired to provide culturally

and context-aware, on-the-ground support to project members (d31, d0). However, project members saw these consultants as ‘separate’ from their teams (d29) and this contributed to the perception of evaluation as a separate activity - the opposite of what the team wanted (d0, c41). Moreover, the consultants were accustomed to conventional approaches. Retraining them for DE required more time and resources than expected (c41, d0), and the outputs that they provided were not well used by program actors (d29). The use of consultants for internal support was abandoned, in large part because it was inhibiting the integration of evaluation into the work.

Tailoring to context: evaluating quietly. The evaluation team also built structural capacity and embedded processes for integrated evaluation (also d31, d29). They often took care to do this in an unobtrusive way, tailoring closely to context and capacity. For example:

...we worked with reframing the concepts of complex systems into language that did not feel foreign, exclusionary, or overly precious. Sometimes, we learned, it is not important to name things; it is sometimes enough simply to integrate the concepts into the work, without calling attention to them (4d0).

Accounts describe preparing feedback data in a way that facilitates use. The team experimented with providing raw data to actors for interpretation, “[a]nd that didn’t really work well” (c42). The team instead typically provides partially analyzed or synthesized data for groups of program actors to interpret and discuss (c42). A participant gave an example: some preparatory data analysis was done by a small team in advance of a meeting. Regional teams were given a question to answer along with these data. Each group then prepared a presentation to deliver their findings at the meeting. Then, “we took their thinking along with other thinking and then had a small group actually lead a discussion session that took us deeper into what that means” (c42). Having a series of steps like this for group analysis and collaborative interpretation reflects a form of pacing, or a ‘runway’ to enable more meaningful and intensive participation (c23).

Making space for DE. Successfully integrating DE as an embedded system took years. As noted in a document: “[i]mplementing an evaluation that is very different from what people are used to can provoke doubt, resistance, pushback and even (occasionally) hostility. We experienced all of these at different points” (4d0). The team needed to actively create space for

the DE. They needed to create space to experiment, fail, recover, adapt, build capacity, establish legitimacy, and earn trust.

The *opportunity* to make space for DE is attributed to factors such as the foundation's culture and its willingness to invest in long-term solutions (d13, also c42). This was identified as key among the factors that "kept the door open" for the developmental approach (c42, also d13). There was growing interest in systems and complexity thinking in foundation circles generally over this period of time, and this was also identified as important (c42, d34). The evaluation team was made part of the program leadership team (4d0, d22, c42), and they had champions at the leadership level. For example, as described by the lead evaluator: "having a program director who didn't need convincing and was fully on board felt like a gift" (4d0). Two members of the evaluation design team were recognized experts in systems thinking and brought significant credibility; they continued to advise the foundation after the evaluation was launched.

Responsiveness. The *doing* of this work demonstrates a high degree of flexibility on the part of the evaluation team to respond to stakeholder needs, as well as patience, persistence and opportunism. They expressed intent and willingness from the beginning to fit the approach to the needs and context of the program (c41, d0). This stance placed some limits on their commitment to particular methods. For example, team members reported being unsure about the extent to which DE might be useful to the program, or what components of the program might need DE and which might be better served with other approaches (d0). Hybridizing with summative studies is noted above. Overall, "as the foundation became more clear about what they needed and we became more clear about what was helpful, we let the evaluation become what it needed to be" (c41, also c42).

Patience, persistence. Patience and persistence are dominant themes. The trust that 'belief would follow practice' led the team to create opportunities to practice via various means, supported by building capacity, enabling participation, and mandating compliance. "Over time people could see... the case for this working and started understanding and buying in" (c42). For example, over the early years the team consistently introduced the use of the 'what?', 'so what?', 'now what?' sequence of questions. Eventually: "[l]ike magic, the questions stuck. The leadership team members began to use the framework explicitly in [program] work, and the

regional teams used it with their projects" (d0). As program actors began to see this as effective, it also helped establish the credibility of the evaluation team (d0). A participant said it this way:

...my experience with evaluation is that very frequently ... people hate the evaluation until they get the data and they have something they can sink their teeth into and they realize 'oooooh, this is very interesting, this is telling me a lot about my program, and this makes me feel like I'm accomplishing something' ... I tell people, 'let's just get through this first round and get some data in your hands (c41).

A participant asserted that important to this persistence was having trust in "what we were trying to create - a complex and adaptive global grant-making program that used evaluative thinking at its core" (d0): this vision fueled their years of effort.

Opportunism. Persistence worked together with opportunism. One actor wrote: "It sometimes felt like for the first 3 or 4 years we learned more about what worked with the evaluation by initially getting it wrong and by crawling through windows of opportunity" (d0). As an example, the team encouraged the foundation to support project grantees to come together in face to face meetings for planning and evaluation (c41). Building support for the idea involved bringing the idea forward at moments when decisions were easier to make. It also involved preparing the ground:

...you can do little things that kind of plant things... It's not like you just hold the idea and then you jump. You're suggesting things and bringing things up and doing things and then when there's an opportunity to make a shift... everybody's sort of ready" (c41).

Learning Outcomes. This case covers 9 years in which a new evaluation 'system' was designed, implemented and integrated into a program. The data suggest the evaluation has had impact at multiple levels. Use of various kinds is described above, and includes conceptual use (e.g., thinking more systemically) and instrumental uses (e.g., program changes). Some broader impacts noted in the data are touched on below.

Foundation level. This evaluation effort is said to have, over time, changed the way that the foundation thinks about evaluation. Having started with concerns about what it might involve, the DE increased commitment to evaluation at the leadership level (c41). This change is

thought to have come about through experiencing evaluation that “was rigorous, [but] that wasn’t antithetical to [their] mindset” (c41). Six years in, the foundation adopted an adaptive action framework “across the entire foundation” (d05). Around this time, the grant that had enabled the program to expand in the first year of the evaluation was ending. The funding partner renewed its commitment with another 5-year grant (d28; d04) based on a proposal prepared by members of the evaluation team (d33). A representative of the partner described the DE work as innovative: “[it] helps teams understand the implications of results and map out change. It impresses me as a process that is cutting new ground” (d33).

Program level. ‘Adaptation’ as an objective is said to have become embedded in the program and this is attributed to the evaluative work. Near the end of the period of this case, leadership articulated a commitment to supporting grantees to adapt their projects over time as they learn; contrasted with approaches by other donors who sometimes penalize change (d29). Doing and using evaluation to enable adaptation has been embedded in the program’s guiding principles (d06).

The evaluation system has also been used to help resolve a tension between stakeholder interests in ‘scaling out’ program results and the program’s commitment to a place-based approach (d0). The system provides support to contextualized scaling (d07), which includes development of ‘adaptive capacity’ as a measure of success (d0) and uses the program’s operating principles and systems thinking as ‘fidelity measures’ for planning and evaluating the scaling process (d0, d07, d29).

Project level. After 9 years, grantees are said to be considering the context of their projects more, as part of their research process. This ‘shift’ was attributed to capacity building supported by evaluation activities (d01). An independent survey of grantees was conducted in the sixth year of this DE. It reported that grantees ranked this foundation’s evaluation process high relative to other foundations for strengthening their organizations and their projects (d17).

Epilogue. This program continues to encompass “a very complex and dynamic set of initiatives” (v01). After almost a decade of change and growth, the foundation remains committed to developmental evaluation to carry the work forward.

Closing Note

These four cases have their own stories: bringing together action and learning for youth services in Case 1, having productive conversations between partners from different parts of the education system in Case 2, stretching boundaries to adapt evaluation to the arts in Case 3 and balancing local and global food security efforts in Case 4. They speak to ways that collaborative learning for adaptation can be mediated over time. In the next chapter, the cross-case analysis pursues key themes that emerged from this phase of the study, such as making space for adaptive learning, navigating a forward path, and overall, the role of mediators.

Chapter 6 Cross Case Analysis

In this chapter, I report findings from the cross-case analysis. These findings speak the role of mediation in an adaptive learning process. Specifically, they describe important activities and points of intervention, how the role is performed and how it unfolds. The findings are developed further with the integration of the two strands of the study in Chapters 8 and 9.

While unique in their own respects, the four cases in this study have key features in common. All describe multi-organizational collaborations whose purpose is to develop innovative social interventions. In all cases, evaluators participated from early stages. *How* they participated, in terms of activities, functions and roles, is the focus of this chapter. The findings are organized by three primary themes: making space for adaptive inquiry, learning forward, and performing roles.

Making Space for Adaptive Inquiry

“Making Space” is a central theme emerging from the cross-case analysis. It is about efforts to create capacity for actors to engage in an adaptive inquiry *as a collaboration*. The metaphor of “space” has been used before in the DE literature, for example to describe the importance of establishing physical and temporal spaces such as meetings or online platforms where program actors can interact (M. Langlois et al., 2012). McKegg and Wehipeihana (2016) write that developmental evaluators need to “carve out space” for people to collaborate in evaluative inquiry, especially in complex initiatives (p. 284). Participants in this study also used the metaphor of “space” in the second, broader sense. Interview participants described having to “carve out space” for people to participate in DE in a productive and non-defensive way, for example when “compliance mode” was a normal way of working (c23). They also described a need to “hold space” for experimentation, analysis and learning, for example when organizational culture was strongly action-oriented (c11). The data indicate that efforts to ‘make space’ were important to the trajectories of all four of the cases. I start this section by outlining why there was a need to create space, and I follow with description of strategies that the evaluation teams used for this purpose.

The need to make space. Productive collaboration was necessary in these cases at two levels: for the evaluation process and for the functioning of the interventions themselves. The

need for collaboration in DE is driven by its strong utilization focus and principles of co-creation and complexity thinking (Patton, 2016b). In addition, in at least three of the cases, reciprocity (mutual benefit for participants) was reported to be an underlying principle (Cases 1, 2 and 4) which increased the importance of multi-level engagement. In general, as explained in a case document, collaboration is considered: “essential to the success of [DE] methods” (4d07).

Moreover, the initiatives themselves were structured to require collaboration in order to be implemented. The relationships between the collaborating groups were highly interdependent in three cases, both in the way the initiatives were designed and because choice of collaborating partners was either fixed or limited in the context (Cases 1, 2, 4). Interdependency was weaker in Case 3 where some flexibility was available to change partnerships in response to tensions. Collaborative relationships were nonetheless important to actors and were a focus of years of effort. The case data suggest interdependency kept actors at the table in spite of challenges in Cases 1, 2 and 4; and a commitment to collaboration in principle was important in all four cases.

In all cases, doubt and uncertainty were evident with respect to the interventions when the DEs began. Doubt is described as a necessary precursor to an active sensemaking process (see Chapter 2). However, in these cases it was not sufficient for the collective inquiry process because actors faced challenges to engage (see Table 6.1, below); this became a core area of activity for evaluators as intervenors in these systems. As observed elsewhere, new collaborations frequently include members “who share a history of conflict, misunderstandings, benign neglect, or have little experience working collaboratively” (Foster-Fishman et al., 2001). Challenges identified in these cases included insufficient trust and mutual understanding among actors, uncertainty and anxiety about the interventions, mistrust of evaluation, and differences in norms for learning. These drove a need to make space.

New relationships, need to build trust. In all four of the initiatives, relationships between the organizations were either new or recently redefined. Relationships involved low levels of trust or low mutual understanding in all cases and were described as “fragile” in one case (Case 2). In three cases, different organizational norms, interests and cultures of inquiry came together in the collaborations (Cases 2, 3, 4).

Table 6.1. Factors affecting engagement

		Case 1	Case 2	Case 3	Case 4
Relationships	Relationships within collaboration	Existing but redefined Funder/grantee relationships (power) Grantee competition New people	New, 'fragile' Funder/grantee relationships (power (-))	Young organization New relationships with local agencies New funding partner	New funding partner Existing funder/grantee relationships (power), Grantee competition New people
	Trust	Low/mixed	Low	Mutual understanding low	Low/mixed
	Diversity	Moderate (cross level, cross function; geo bounded)	High (cross sector, cross level, cross function; geo bounded)	High (cross sector, urban/rural, cross function)	Very high (cross sector / culture, function, level)
	Interdependency	High, structural	High, structural	Moderate - low	High, structural
Uncertainty	Program	New, no model	New, no direct model	New, no model	Established, changing
	Expectations for program impact	Near term, direct, high stakes	Near term, direct, high stakes	No specific targets	Long term, indirect, stakes
	Investment	High risk, high stakes	High risk, high stakes	Low stakes, pilot	Risk, stakes
Evaluation	Evaluator(s)	New to org	New to org; from 'away'	New to org, not from sector	New/mixed, not from sector
	Opinions about Evaluation	Mixed, Wariness Needed++; Accountability, Oversight Preconceptions: evaluation is done a certain way, for some	Mixed, Fear (punitive) Needed++; Accountability Preconceptions: evaluation is done a certain way, for some	Wariness (can alienate, inappropriate) DE might be helpful Preconceptions: eval a certain thing by nature; Most/All	Mixed, Wariness (burden, negative past experiences) Needed +; Learning and Accountability Preconceptions: eval has to be done a certain way, for some
	DE	New, unfamiliar	New, unfamiliar	New, unfamiliar	New, unfamiliar
	Sector norms re. learning	Mixed. Action, doing Learning a luxury	Mixed. 'Under siege', "compliance mode"	Mixed Exploration, reflection	Mixed. Compliance, performance
Norms	Internal learning (re. reflective practice)	Mixed / Anti-norm, learning is a luxury History of evaluation use	Mixed / Anti-norm, putting out fires Openness to innovation History of evidence use	Norm / Complexity thinking No past evaluations	Mixed / Anti-norm, compliance, performance Openness to innovation History of eval but limited use
	Complexity thinking	New	New	Norm	New

Members of the collaborations came from different operational functions (e.g., project funding and project delivery), organizational levels (Cases 1,4), sectors (Case 2) and cultures (Case 3, 4). The diversity contributed to tensions. These affected, for example, members' ability to reach agreements on objectives (Case 1, 4) and make decisions about appropriate actions for the initiative (Cases 2, 3, 4). In all cases, the evaluator(s) were also new to the initiatives. In three cases, they were outsiders at some level in a way that affected trust (2, 3, 4).

Uncertainty and anxiety. None of the initiatives had an established, proven model to follow that matched what actors envisioned. Three were entirely new interventions. The fourth was an older program that had suddenly expanded in scope and was shifting its approach. This contributed to differences in understanding of the intervention among actors. Participants also described risk in the decision to innovate: the scale of investment was another key contextual factor in three cases (1, 2, 4). One program participant described their initiative as: “a huge risk... a huge investment beyond anything the organization had done before” (c21). These were noted sources of uncertainty, at times anxiety.

Mistrust of evaluation. In all cases, at least some actors were skeptical of evaluation. Some saw evaluation as potentially “alienating” and counter to a creative process (Case 3) or as an undue burden, a perspective influenced by unsatisfactory past experiences with evaluation (Case 4). In two cases, some actors were said to have experienced evaluation as ‘punitive’ in the past or were concerned that the funder would be ‘scrutinizing’ them (Cases 1, 2). In at least two cases there were actors who felt that evaluation might harm the interventions (Cases 1, 3).

This history was described by participants in these cases as a reason to choose an approach that emphasized participation and learning. Rather than making evaluation easier, however, DE proved demanding. It unsettled some program actors. Particularly in three cases, there were reports of tensions arising from the DE process because participants were confused about expectations, the process was intrusive (Case 2), resource intensive (Case 1), or pushing people out of their comfort zones (Case 1, 2) or there were concerns about credibility (Case 4).

A Case 1 participant explained that the importance of establishing a shared understanding of the evaluation approach is not unique to DE: “Evaluation can mean 200 different things. People often talk past each other, you’re thinking performance measurement, I’m thinking

adaptation in an innovation sense, and then something happens and one of is not happy with what happens” (c11). Yet establishing understanding was compounded because DE was new and unfamiliar. For some actors, DE ran counter to preconceptions of what evaluation is or should be, for example some stakeholders were said to be “conditioned to think evaluation has to be a particular way” (Case 2, v12). The evaluators, with help from others acting as champions, needed to explain and protect the approach and its underlying assumptions (e.g., complexity) to avoid shifting to formative or summative evaluation, which were more familiar, while the program needs were still developmental. A Case 2 participant reported that it took stakeholders about a year to “really understand what this approach is all about” (c23).

Although learning and program development were the primary objectives for all four evaluations, in three cases (1, 2, 4) accountability for the use of resources was also a driver. For these three, there was a need to be accountable to organizational leadership, donors and/or other funding partners, and evaluation was described as necessary for this purpose. Case 3 was an exception. The funder approached this intervention as a pilot. Evaluation was described as potentially helpful, but not necessary, and accountability was not identified as a driver. This was reported to lower the stakes in a way that was important for “creating space for experimentation and adaptation” (3d0) but may have also reduced commitment to evaluating the intervention. This will be revisited below.

Norms and climate related to learning. Although accounts describe all four funders as committed to internal learning and innovation, they also describe variation in perspective among members inside and across organizations in this respect, that is, how learning should be undertaken, as well as differences in orientation toward or capacity for inquiry. For example, participants described efforts through the DE to “help people to be curious about their work” (Case 1, c12), to ask questions they would not otherwise have asked (Case 2), or to develop a general stance of inquiry (Case 4). A participant explained the need to develop skills for collaborative analysis and for using results to inform decisions, because this capacity is undeveloped in many organizations (Case 1, c11, see also Cases 2 and 4).

Differences in norms around learning were striking in at least three cases (Cases 1, 2, 4). In one case, a participant described the two principle partners as having: “two entirely different organizational cultures trying to come together” (c21). One partner (the funder) was “privileged,

for all sorts of reasons” which enabled “an adaptive culture” and a secure position from which to embrace evaluation. In contrast, the climate of the partner organization was described as highly political and turbulent, with a background adding to mistrust that was unhelpful to a joint learning process (Case 2). In other cases, an ‘action-oriented’ culture, that is, one that focused on ‘doing, not learning’ (Case 1, 4), actors conditioned to operate in ‘compliance mode’ that is, focused on ‘checking boxes’ on promised outputs (Case 2, 4), or limited by a focus on short-term financial considerations (Case 3) were said to inhibit in-depth engagement.

Strategies for making space. The case accounts describe a variety of activities in response to the above challenges in order to ‘make space’ for productive, collaborative inquiry. Within an overall emphasis on quality facilitation and relationship-building, these included activities under four broad themes:

1. ‘cooling’ strategies to reduce pressure on participants and improve mutual trust
2. efforts to balance and fit the DE to diverse stakeholder needs
3. normalizing evaluation, and
4. developing opportunities for actors to experience the process and results of the evaluation work as a way to build capacity and support.

Champions also played a key role (see Table 6.2, below).

Cooling. For three cases, and to a lesser extent the fourth (Case 3), actions were taken to lower the heat, so to speak, to reduce stress, allow engagement to be more productive and help improve trust. Cooling efforts included ‘lowering the stakes’, using objects to mediate interactions, and modeling or “walking the talk” of collaboration.

Table 6.2. Strategies for making space.

		Case 1	Case 2	Case 3	Case 4
Cooling	Lowering Stakes	Resetting expectations for impact, coaching.	High quality communications Transparency in-group; Resetting expectations for impact (near term)	Stakes already low Communications: using the 'rhetoric of research'	Reframing success as adaptive capacity, not compliance
	Epistemic objects	TOC as a thinking tool	System map (medium for interaction; widening frame)	TOC, but for developing org identity & purpose (cooling tool (-))	TOC to negotiate, identify assumptions, guide action, give traction; widening frame
	Walking the Talk: Co-creation, ownership, reciprocity	Joint learning agenda Participatory data analysis Data sharing Sharing results back (voice) Participation at project level, also org staff, advisors (-) Face-to-face++ at multi-level	Eval design Joint learning agenda Participatory data analysis Sharing results back (raising voice – some groups) Participation all levels Face-to-face++ at multi-level	Eval design Data collection Interpretation Participation primarily project level, also org staff Face-to-face++ project level	Joint learning agenda Joint data collection, analysis Sharing results both up and down Participation all levels Face-to-face++ multi level
Fit and Balance	Deliberate Experimentation	What works for agencies	Methods responsive	+++Find fit with artistic initiatives	+++Balance heterogeneity and structure
	Language, Communications	TOC as boundary object Documentation (tracking, discipline)	System map as boundary object Face to face meetings; informal debriefings; structuring communications	TOC as boundary object Methods as communication (story writing, video) Attention to language	TOC as boundary object Close attention to language
Normalizing evaluation	Evaluating quietly	Simplicity (one question); opportunistic data collection TOC via questioning	Habit-building, routine	Co-creation with artists to integrate Creativity with methods	Integrating without calling attention
	Runways, pacing	TOC via questioning	Runway (concepts) Pace of information flow to capacity, lower stakes		Runway (analysis, interpretation) Laying the ground
Champions	Champions	External experts Project lead, anchor, mediator, counsellor	Trusted advisor, on the ground, interpreter	Advisory group members	External experts Director, foundation VPs. Anchors, mediators. ** higher in hierarch.

Lowering the stakes. In two cases, important stakeholders held very high expectations for impact from the intervention in the near term, namely within 4-6 years (Case 1, 2). In these cases, actors described coaching efforts to reframe stakeholder expectations at multiple levels. A Case 1 program participant noted that: “through DE we were able to have very honest conversations about ‘how doable is this?’”. They were able to gain support for the view that:

the high school graduation goal - that should be in another room. That’s still our end game... but we need to be much clearer about what’s more proximal to us, and ... how that is an important contribution to shifting the system (c12).

In a third case, although the principal funder expected impact to accrue over a longer term and in a less direct way, participants reported a need to reset some actors' understanding of success so that they would focus on building their adaptive capacity rather than demonstrating compliance with planned activities and outputs (Case 4).

In Case 2, where the climate for a partner was especially political, a process was put in place to ensure that communications about evaluation findings could be reviewed, questions asked, and the findings understood, before communicating findings outside the collaboration. Regular meetings helped reassure stakeholders by increasing the transparency of the evaluation. Even in Case 3, where stakes were intentionally held low, the evaluation work was framed by some actors for their collaborators by "using the rhetoric of research" (c31). This framing was said to avoid the impression that the work would be used to judge their work as 'good' or 'bad'. It was also thought to reinforce the message that the inquiry was to generate learning of mutual interest.

Using objects as 'mediums for interaction'. An 'epistemic object' is something that can be used to help people engage in joint discussion and interpretation. A systems map can be used in this way, for example. When approached as a living or working draft that is open to iteration, it can be a lower-stakes way to participate in joint meaning-making. It can help to depersonalize issues, open up questions, and reduce attachment to ideas before they have been tested (e.g., c12, see also Engeström & Middleton, 1996). The evaluation teams used theory of change in two cases in this way (Case 1, 4). The TOCs were described as a collective "thinking tool" in Case 1, that "[helped us] hold our ideas about the initiative lightly" (c12) for example. In the one case that did not use a TOC (Case 2), systems mapping served a similar purpose. In the words of one participant, the map: "really helped to get folks, especially the folks putting out fires every day, to kind of get their head up and see how the work, what we're trying to pull off here... to really recast the players for everyone" (c21).

Modeling principles, walking the talk. Co-creation, joint ownership, and reciprocity were voiced as important commitments in the evaluations in all cases, and these commitments were made visible and explicit. They were described as important for building trust, a sense of ownership and assurance through shared control, while also increasing the relevance of the

questions being asked and the overall salience of the evaluation to stakeholders. They were made explicit through, for example:

- jointly developing a formal learning agenda, following it, and updating it collectively over time (Case 1, 2, 4)
- joint design of the evaluation in at least two cases (Case 2, 3)
- participatory data collection (Case 3, 4) and participatory analysis and interpretation (all cases)
- supported by an overall appreciative approach that emphasized learning for development.

Participants in Case 2 remarked on how a partner's perspective on the evaluation changed after an opportunity to co-develop a redesign of the evaluation. In this case, the evaluation team and other members of the collaboration were taken by surprise when the partner rejected the initial design after months of preparatory work and consultation. Although the new design was not dramatically different, the redesign *process* was a turning point. A participant remarked: “[t]hey needed ownership and buy in. In order to achieve that, they really needed to be part of the sausage-making” (c21). The redesign process strengthened the relationship and established a way of working; it: “helped us gel more as a team and understand each other's perspectives better and how we're approaching this initiative” (c21). Other examples of modeling collaboration include structuring feedback processes to ensure that partner agencies (who contributed data and who sometimes competed with each other for funding) had joint access to and use of the data generated (Case 1) and designing creative methods to enable meaningful participation by members of the community (Case 3). A Case 1 participant explained:

...the nature of how people engage with [evaluation] is totally different [depending on] how you frame and position it... So instead of saying, ‘I’m going to collect some info’ and then disappear, [say] ‘we’re going to collect some info together and we’re going to use it here now for something that’s probably relevant to you (c11).

Reframing evaluative inquiry as learning-oriented rather than judgement-focused was considered key for shifting stakeholders' focus from “did we make the mark?” to ‘how can we learn in order to adapt?’ (Case 4, d34).

There were limits in place, both formal and informal. For example, in Case 4, regional and project level diversity and autonomy were points of sensitivity; the team needed to respect this while promoting alignment. In Case 3, the evaluation was explicitly scoped to exclude partnering funders and quantitative methods. In Case 2, the DE's scope excluded a primary partners' internal operations. This was described as very challenging because the work of the partner and the work of the initiative were so interconnected; as described by a participant: "we had to be very careful, we were always walking a really fine line" (c22). The teams' ability to 'walk this line' and respect the limits in the collaboration helped to hold the space for the DE.

Efforts of the evaluators to model co-creation and reciprocity were built on and enabled by commitments from within the collaborations in all four cases, expressed as values commitments or principles, particularly of the funding organizations. However, a stated commitment by a funding organization is not sufficient to make principles live in the program on the ground. The examples above illustrate ways that the teams sought to operationalize principles in practice and demonstrate that commitment to program actors in concrete ways.

Seeking fit and balance among different needs. The teams experimented during implementation of the DEs as part of efforts to negotiate diverse needs and support engagement. For example, evaluation tools and approaches were piloted at small scale and extended where a good 'fit' was found. For one of the cases, the 'fit' they were seeking was with the use of evaluation methods for highly creative and exploratory work (Case 3). In the other cases, experimentation was aimed at ensuring participation of a broad range of stakeholders, for example, to 'figure out what works' in terms of methods for participants from front-line agencies (Case 1, also Case 4) or for clients of the intervention that were more difficult to reach (Case 2).

The case data highlight that attention was paid to tailoring communications as well. For example, in Case 3, the conventional visual format for a ToC did not resonate and an extended narrative was written instead. This was described as highly unusual, but effective for this organization's culture. In Case 2, regular verbal debriefings with key actors were added to a reporting cycle to provide reassurance. The debriefings allowed members of this stakeholder group some time to process the information before a written report was prepared, and to flag to the evaluators if findings touched on a potentially 'hot button' issue so that it could be approached with sensitivity.

Language affected how different groups responded to the DE work. In Case 4, a priority was placed on the careful use of language after specialized and unfamiliar terminology resulted in pushback. Sometime later, when developing a monitoring tool, the team invested significant time to identify and incorporate terms and concepts familiar to program actors, that is, “the way we talk about it in [our] world”. The result: “really helped bring people around to the idea [of using the tool]... everybody kind of liked it” (c41). Other examples are noted above, for example, where DE was referred to as research to make it seem less threatening (Case 3).

Normalizing evaluative practice. Some actors held preconceptions about evaluation that contributed to skepticism about it. In some cases, learning-focused activities appeared to run counter to organizational or sector-based norms. The teams encountered active opposition as they were implementing the DE in at least three cases (1, 2 and 4) and they anticipated resistance in the other (Case 3). Making the DE process ‘normal’ was an important part of efforts to make space for it, and they applied various strategies to encourage actors to participate, and develop familiarity, comfort and capacity. These included evaluating quietly, that is, in ways that aligned with and did not disrupt the intervention; introducing intensive activities in stages by using ‘runways’; and creating structures that helped to protect time for and/or anchor aspects of the evaluation in the initiatives. Some hoped that, if successful, establishing a habit of practice would lead to greater valuing and further engagement with evaluative inquiry.

Evaluating quietly. In three cases, accounts describe introducing evaluation activities without calling attention to them, for example collecting data in very simple ways through existing activities (e.g., events for purposes other than evaluation) (Case 1, 3, 4). In Case 1, simplicity and opportunism was an intentional aspect of design. This was not an effort to hide the evaluation (c11), but to integrate it into regular events and activities of the initiative to normalize it. A team member explained the purpose: “to make sure that evaluation was the work and the work was the evaluation” (c12). A participant described, as an example, incorporating a ‘mini evaluation’ at every event: “where we just had people write on a sticky note the answer to one question” (c12). Another participant in this case noted: “the more [evaluation is] just an extra thing the greater the likelihood that it gets dropped”; and, “[Evaluation] should be part of their work more than this add-on thing, the more that’s a mantra you take, the more you have something useful and sustainable” (c11). Similarly, in Case 3, integrating the evaluation so that it

was seamless with the projects was intended to avoid “too much evaluation that didn’t feel like part of [the intervention]” (c32). In one Case 3 project for example, the evaluator collaborated with artists to structure the event so that attendees could easily share experiences while moving between sites (3d0, 3d15), in order to “layer evaluation into what was happening” (3d0).

Evaluating quietly also involved using of familiar or ‘friendly’ language as noted above (Case 3, 4). In Case 4, after a proposed approach for the DE was rejected by stakeholders, evaluation team members wrote: “we learned it is not important to name things; it is sometimes enough simply to integrate the concepts into the work, without calling attention to them” (4d0). These are examples of efforts to make the DE unexceptional and normal.

Using runways. More intensive activities – such as participating as a group in data analysis or co-constructing a TOC – were sometimes broken down into a series of steps. Accounts in three cases (Case 1, 2, 4) speak of introducing intensive activities slowly, and laying “a runway” (in the words of Participant c23). A runway was described as a way to warm people up or prepare the ground for a more intensive or more abstract activity, particularly a collective one. Participants talked about how this allowed for more in-depth interpretation and ‘ah hah’ moments for the program teams. For example, the Case 4 evaluators had found that simply giving raw data to a group to analyze “really didn’t work well” (c42) and for a subsequent event, they set up a phased approach. In advance of a meeting, small teams were asked to do some data analysis on their own. Later, each group presented their findings to a larger group. Related data were introduced. This was followed by a facilitated whole-group discussion. The last step took them ‘deeper into what the data meant for the initiative’ (c42). A Case 2 participant described a similar process for co-developing a systems map. A runway was also thought to add time for individuals to interpret and reflect, important to a process of eliciting and integrating multiple perspectives in analysis.

Structures. In each case, the teams used more formal structures, such as regular reflective practice meetings, as a way to reserve space for DE activities. Regular face-to-face meetings following a consistent structure and with stakeholder groups at multiple levels were a feature of at least three cases (1, 2, 4). At these meetings, program actors would formulate questions, discuss and interpret data, and consider next steps for the initiatives. For example, in Case 4 the addition of project-level meetings was described as a way to: “structure things to allow change to

happen” at that level (c41), that is, to create a platform for project teams to be more deliberative in their project development. In these meetings, project teams were encouraged to use a consistent question framework (‘what?, so what?, now what?’) in their discussions and to develop, review and revise project-level theories of change, “as a springboard for conversations” about planning (4d07, also 4d22, 4d26). The team introduced the same question sequence in leadership meetings, and over time the framework became habitual (4d0). The use of structured question frameworks was noted in other cases as well. For example, in Case 1, the regular probing questioning used in the DE at regular intervals was said to encourage the team to pause and make sense of what was happening, to “avoid running headlong into implementation of the initiative” (1d-er/ c12). In Case 2, describing a recurrent meeting that followed a set structure, a participant noted how opportunities to practice made engagement with the DE easier for participants: “after the first year especially... [stakeholders said] ‘ok here is the meeting we come to where these guys talk about some findings and we talk about implications’” (c23). A team member expressed surprise at how productive the meetings became, and they attributed the success to structure, repetition and practice.

The opportunity to practice was considered important because actors believed that appreciation for evaluation would follow from experience with it. As noted above, a Case 4 participant described how often people “hate the evaluation” until they have a chance to see the relevance of the data it provides (c41). Cooling, evaluating quietly, staging the introduction of intensive activities and adapting or introducing supportive structures - these are examples of strategies used to encourage participation, navigate around resistance, and give people a chance to practice doing evaluative work. Importantly, priority was placed on developing actors’ capacity to question, and to receive and use feedback, over and above training on methods-related or technical skills (Cases 1, 2, 4).

Champions. The efforts described above took time to bear fruit. In the meantime, actors needed to sustain the work. As noted in a Case 4 document: “[i]mplementing an evaluation that is very different from what people are used to can provoke doubt, resistance, pushback and even (occasionally) hostility. We experienced all of these at different points” (4d0). In addition to efforts such as those noted above, participants in Case 1, 2, and 4 described points in time when the active support of internal or external champions helped them turn a corner or get unstuck.

Participants reported intervention by champions within the funding agencies. In one case, an internal staff person in a leadership position acted as a broker between levels of the organization, reporting and providing assurance upward about the value that DE was bringing to the initiative. They also acted as a counsellor for staff who struggled to adapt to the approach. A participant described how this person “held the space [for DE] and normalized it” among staff in the organization, until eventually, “the team saw value and benefit and it just became part of their practice” (c11). In another case, an internal staff person acted as a trusted advisor who interpreted local issues for the evaluators (based elsewhere) which supported their sensitivity to changing conditions. As well, in all four cases, external experts or advisors supported the DEs. As stated by a Case 1 participant, external advisors: “brought a lot of currency, they saw this as a legitimate approach and people then within the organization said, ‘if they’re on board, then we are too’” (c11). In two cases, outside advisors provided specific interventions linked to turning points in program development by adding their weight to the recommendations of the evaluation team (Case 1, 4). In at least one case, their influence was linked to their status, which allowed them to say things that the evaluator could not say (Case 4).

Closing note. In the cases described in this paper, the data indicate a need to actively establish or build space for collaborative inquiry. Evaluators invested effort in “helping people to be curious about their work” (c12) and to “see the bigger picture” of the intervention (Case 2, 4). A case document stated that the evaluation team “identified ways to build capacity among... stakeholders to *receive* feedback [and also] to use information to drive decision making” (Case 2 d20). A participant described ‘norm-setting’ around evaluative inquiry as part of the intent of DE and a reason for DE to be long term, because establishing such norms is a ‘longer term endeavor’ (c23). Progress was uneven. There were fits and starts, steps forward and steps backward. Some groups were able to move faster than others could.

Case 3 stood apart in some respects. The perceived need for the evaluation was less, the stakes were lower and there was less interdependency in the relationships in the collaboration. Multilevel engagement, which characterized the other three cases, was less evident. Although tensions surfaced among partners, a solution was found that involved reshaping how the organization formed partnerships. Where there was need to create space for DE to function in the

intervention, the challenge centered more on space to experiment with evaluation as a legitimate and useful mechanism for learning.

In all cases, the consultant evaluators were not working alone. This work was shared with other actors (e.g., program teammates and champions) and supported by tools. Although accounts in the four cases agree on positive outcomes of these evaluations (see below), in two cases participants voiced disappointment that they were not able to go further or move more quickly. This will be elaborated in the next section, “Learning Forward”.

Learning Forward

Action (informed and viable action) is an important component of the definition I have used in this study for adaptive learning. Action is also one of the three main phases of the sensemaking cycle depicted in the theoretical framework (see Chapter 3). The section above dealt primarily with establishing conditions for and carrying out collaborative inquiry. In this section, I report findings about applications and actions: about ways in which evaluators’ activities enabled forward movement to help develop the interventions. I have organized the findings under two broad themes: traction and outcomes. I close this section by noting differences in the directions taken for evaluation by the collaborations later.

Traction. In all four cases, the evaluations were multi-faceted in terms of their foci for inquiry and all involved exploration of both foreseen and unforeseen influences on the programs. In at least two cases (2, 4), the actors reported producing so much information that at certain points the teams were struggling to absorb the data. The evaluations were also unfolding alongside rapidly changing initiatives that were themselves embedded in dynamic contexts. For these reasons, clearly documenting the decisions, actions and progress made by collaborators was important for actors.

The documentation was valuable for internal reflection as well as for external reporting. In addition, it provided a platform for the DEs to promote prospective thinking and forward movement; to gain traction and avoid repetition. The data suggest efforts to gain traction in multiple ways, including clarifying ideas, prototyping and testing, navigating and nudging, and using anchors (see Table 6.3).

Table 6.3. Efforts to gain traction.

	Case 1	Case 2	Case 3	Case 4
Clarifying ideas	-Clarifying the TOC; surfacing need for changes <i>and</i> alternate pathways forward (e.g., systems change via org learning to systems change via lateral connections) -Use of questions that “help people see better”	-Using data to “crystallize underlying issues” -Data for more certainty (instead of working on assumptions) -Clarifying fuzzy constructs (e.g. “college culture”) <i>and</i> implications for implementation; surface need for change (e.g. a ‘critical mass’)	- Documenting the participant experience. - Documenting the collaborative process to understand its features; illuminate the “what”. - Leading to options for approaches.	- Clarifying internal connections via the TOC; clarifying assumptions and “miracles”; - supporting agreement on scope; “bounding creative energy” of the team (4d0).
Prototyping, testing	-Systematic testing TOC to deepen understanding of the initiative -Prototyping, e.g., approach for building a network (1d0); ideas for engaging front-line staff	Not noted explicitly	Experimenting with approaches to collaboration, feedback on efficacy (c32, d33), indicating new direction (alternate modes).	Testing theories about the program against measured results (d29); testing and refining the TOC
Navigation, nudging	-Use of data to indicate solutions to challenges -Structured questioning and reflective practice to “hone strategy, course correct” (c12) -Structured “strategy sessions” with core agencies -Using data to say “are we on the right track?; pulling people back” (c11) -Not losing sight of impact -Use of principles (inclusiveness), explicit learning mission as guidance (e.g., re. advocacy role).	-Data that “point to solutions” -Productive, data-driven conversations (2d0); Eval as a mechanism E.g., meetings “where the rubber met the road” (c23) -Data as rationale to funders. -Timing of data: upstream of implementation, e.g., equity issue -On-track/off-track feedback at multiple levels of implementation -Support to prioritization based on understanding of change processes -Keeping heads up; “remember where we’re going” (c22)	Developing the foundation’s TOC (principles as guidance)	-Data indicating possible solutions -TOC and collaborative development of TOC giving traction; Project level TOCs guiding alignment between intentions and actions (d25) -Outside champions (leverage to move on key recommendations)
Anchors	-Question set for events used 2-3x / year -Structured doc to record reflective practice -TOC documents learning.	-Regular meetings -Communications structures -Making progress visible; Data to discern trends and see early outcomes in rapid change.	-Documentation -Repeat staff interviews	-Tools to anchor systems thinking (e.g., project quality checklist); TOCs; project selection ‘action plan’ in TOC -Operational principles embedding M&E.

Clarifying ideas important for implementation. Participants in all cases provided examples of identifying ideas that were core to the implementation of the initiatives but that were either not well defined or were understood in different ways by different program actors. For example, in Case 2, the initiative intended to promote a “college culture” among local young people, but early evaluation findings pointed to differences in understanding of what this meant. The team worked to help actors articulate and develop a shared understanding because while the idea remained “fuzzy” it was too difficult to implement. It could even be difficult for actors to recognize. In the words of a participant: “it’s just really hard to see it, right?” (c22). They explained how this enabled them to move forward: “when we started getting a more common understanding, then we started thinking about: what are the supports that schools should have in place? Or, what are the external relationships that schools need in order to ... create that culture?” (c22). Examples in other cases include identifying assumptions in theories of change (Case 1, 4) and illuminating features of the collaborative process (Case 3). Notably, each of the cases describe clarification of concepts leading to forward movement, for example, identifying a range of options for action, and then testing these with further data collection.

Prototyping, testing. In three cases, the teams tested ideas by collecting data against proposed paths in a TOC and/or experimenting with alternative approaches to implementation (Cases 1, 2, 4). For example, in Case 1 the team identified underrepresentation of front-line staff in the initiative, facilitated the generation of ideas for engaging them and tested these ideas on a small scale. This led to changes in the initiative to facilitate their participation.

Navigating. A substantial amount of the case data in three cases (1, 2, 4) concerned explicit attention to course correction and adjusting strategy. Participants and documents recount the use of feedback as “on-track/off-track” indicators (2d0) and “watching for deviations, reminding of the rationale, questioning shifts, pulling people back” (c11); “using data to say, are we on the right track?” (c12). “Strategy sessions” were organized with core agency members in Case 1 (c12), and stakeholder meetings in Case 2 were events where “the rubber met the road” for strategy for the coming year (c23). The evaluators’ involvement in development of strategy is revisited below (see Roles) and ranged from encouraging the creation of structures to enable strategic work to be collaborative, to contributing their own understanding of change processes to inform actors’ priority-setting (see e.g., Case 2).

The teams in three cases reported using the initiatives' goals and established principles as signposts to help actors find their way (1, 2, 3). In Case 1, intended impact on graduation rates was described as the "goalpost" to be kept in sight even if understood to be a long-term aim (c12). A second participant noted: "there's some confusion that people think [DE is] just about process, and to me you can never lose sight of the impact, the intended impact. Even though your understanding of what that is or what that looks like evolves over time" (c11). This was echoed by a participant in Case 2: "sometimes clients [are] caught in the day to day operation of the initiative... what we're trying to do is to continuously say: 'hey, yes, you're having this problem right now, but remember that this where we're going... are we getting to where we said we were going to?'" (c22). Along with goals, principles were also used to inform direction, particularly when issues were contentious, or a clear path was difficult to see, such as in Case 1 when the principle of inclusiveness was used in the decision to continue growing membership.

Anchors. Making progress visible through tracking and documenting, for example in TOC updates (Case 1, 4), appears to have been an important basis for forward action. It offered a thread of continuity in interventions where change and disruption were a norm. A participant explained: "when you are innovating or building something new, sometimes it's really hard [to see] what are the early wins, the early outcomes of what you're doing" (c22). Other tools used for tracking, such as a regular interview with key staff repeated at intervals over years (Case 3) helped add clarity. A Case 1 participant said about a tracking tool that they 'really stuck with': "At least 2 or 3 times a year that would get updated... what that revealed to us was the arc of the whole experience, how the thinking evolved over time, what we figured out, how they responded" (c11). In at least one case (Case 4), tools were used to anchor strategy in the operations of the initiative, for example by formalizing a new framework for prioritizing projects by embedding it within the TOC (4d05) which helped translate the framework "into an action plan" (4d05) or by embedding systems thinking into a project quality monitoring tool.

Outcomes. In all four cases there is agreement across accounts (program staff, evaluator interviewees and documentation) that the DE work contributed positively to the interventions. Reported outcomes included instrumental use of evaluation findings to make program changes that helped beneficiaries (e.g., to redefine the boundaries of the intervention to address a gap in inclusion in Case 2). They also included conceptual use, such as rethinking program approaches,

rethinking how the organization learns and/or shifting culture around learning and evaluation as a group (e.g., normalizing reflective practice in Case 4). Other noted outcomes of the DEs included improved clarity of theories of change and establishing guiding principles or guiding values for the work (for the latter, see e.g., developing guiding principles for the foundation in Case 3 and operational principles in Case 4). For some, the evaluations were credited with helping to secure greater support for the initiative (i.e., from collaborating actors, from the public, other funders) which represents a form of evaluation use sometimes called tactical or persuasive (e.g., demonstrating commitment to accountability) (see D. Fleischer & Christie, 2009). Non-use is evident in two cases (see below), but there were no reports of misuse¹⁴. These data are summarized in Table 6.4, below.

Producing results that actors value. In all these cases, accounts concur that engagement with the evaluation increased as participants experienced results that they valued. Examples of such valued results include renewed or expanded funding for the initiative (Case 1, 2, 4). A program participant underlined the importance of this, saying: “we were able to point to the evaluation findings to say ‘look’ there’s data to support this... If you want to talk about a win, that was huge, huge” (c21). Other ‘successes’ included the evaluation’s ability to provide rapid feedback about emerging and unanticipated problems that actors saw as a threat (Case 2, 3), or surfacing issues that actors cared about and then supporting the decision-making that resulted in the issue being addressed (e.g., Case 2). Accounts stress that the use of evaluative data to make ‘win-win’ changes to the program encouraged active engagement with the DE. Program changes based on evaluation findings helped to establish a visible link between learning and doing for actors who valued action, and they helped reposition actors’ perceptions about the evaluation’s purpose (e.g., from compliance to development) and increase their confidence in the ability of evaluation to achieve that purpose.

Non-use. There were reports of non-use in three cases (1, 2, 4). Note that non-use or decisions not to act on feedback from an evaluation can sometimes be warranted (see e.g., Cousins, 2004) and a suggestive example of this was noted in Case 1 (see Table 6.4). In Case 2, three participants expressed some regret that the evaluation was not able to motivate progress more quickly. Some recommended changes were more readily actioned than others, in part

¹⁴ The coding for this section on use outcomes follows a framework described in Cousins (2004).

depending on where the change was indicated (which organization, at what level), capacity to make changes, and dynamics in the context (c23, c22, c21). One participant explained: “I think developmental evaluators are like, ‘on it, let’s make changes, let’s do this’ – and it takes longer for people to implement. They have to deal with structures, they have to deal with budgets” which are sometimes outside their direct control (c22).

DE is characterized by participation in the evaluation process, frequent feedback and regular reporting. This makes it difficult to clearly distinguish between use that arises from the process of participating in the evaluation (process use) and the use of findings. One program participant attributed many of the outcomes of their DE to process use, saying that although findings were also impactful and “have absolutely shifted how we’re doing our work” (c21), other outcomes, especially the changes in the quality of the collaborative work, came about through process use. This participant noted that process use, while important, is harder to articulate and easy to overlook.

In this section, I have been reporting on how the inquiry was used to enable development. This was not limited to (and sometimes did not involve) immediate, concrete action; it sometimes involved less tangible outcomes such as conceptual clarity or strengthened commitment to collaboration. This suggests that greater nuance is warranted in the third “phase” of this study’s conceptual framework which focused on action. This will be discussed further in Chapter 9.

Transitioning. These four collaborations stayed with DE for multiple years. In two cases, they later shifted into formative evaluation (Case 1 in year 5; Case 2 in year 4), while also retaining DE elements. Accounts in Cases 1 and 2 attribute the shift to formative evaluation to a greater comfort with and stability in the initiatives, accompanied by increased focus on improvement and refinement than development. In Case 3, the funder transitioned to action research after three years with DE. Participants reported that research questions were of greater interest than questions aimed at internal operational development, where evaluation usually focuses. Moreover, findings from the DE indicated a need for deeper engagement with participants, for a longer duration. The program actors took this up through a new program of action research.

Table 6.4. Reported outcomes.

Use	Case 1	Case 2	Case 3	Case 4
Instrumental	-Expanded use of evaluation in agencies; earlier use in program cycles (d-er) -Strategy changes; changes to plans; shifts in direction (c12, 1d0) -Changes to implementation -Decisions <i>not</i> to change path (e.g., to continue welcoming new people; not to pursue an advocacy role) -Year 2 and Year 4 funding extensions	-Changes in strategy (c22) -New activities / actions to respond to implementation issues raised by the eval (e.g., parent awareness; equity concerns; relationships with universities; targeted support for specific students) (c23; c22; c21) -Introduction of sister initiatives to support the existing e.g., low income housing support (c22); parent network.	-Changes in approach to collaboration; less centralized decision making (c31) -Contributing to guiding principles for the foundation -Changes in plans (e.g., to create a pedagogical toolkit) (d30) -Changes in methods for learning and programming (action research)	-Development of action plans -Changes to activities (c41) -Expansion of the use of eval framework to other areas (d05) -Operational principles include M&E -New 5-year grant from partner based on proposal written by eval team; partner recognition of innovative work in eval.
Conceptual	-Key assumptions about the initiative changed (d-er; c11) e.g., pathway to systems level change -Deeper understanding of intended outcomes (c12) -Realization “they had no idea how to get to” some of the intended outcomes (doubt) (c12) -Resetting expectations for outcomes; shift to contribution perspective -Technical eval skills, ECB -Eval theory ECB -It “shifted our culture” (c12); Culture and practice of inquiry (d/c12, c12, c11)	-Helping to “open our eyes” to the complexity of the problems faced; how problems could be approached differently (2d0/c21, d23, 2d0), understanding from multiple perspectives -Changes in conceptualization of the intervention influencing strategic thinking (2d0, c22) -Norms for group interaction; productive conversations (c23); strengthened working relationships (c21, v12); increased trust (c22, c23) -++ capacity to use eval (2d0, c21, v12) -More positive attitudes to eval (c21)	-“Prompting a way of thinking” (c31), “Rethinking [the org’s] way of knowing” (c32) -New ideas for programming (c31) -Developing understanding of organizational role -Developing / clarified TOC -Realizing how challenging the intended outcomes are to achieve, and why -Realization that fundamental program change necessary	-Thinking more systemically -Changed perspective on scope of research quality -Changed perspective on evaluation (improved attitude; more committed to eval) (c41) -More coherent, complete, shared understanding of initiative (via TOCs at multiple levels).
Tactical / Persuasive	-Enhancing the credibility of the intervention; to demonstrate accountability (1d0, c11) -Providing time to develop and demonstrate worth of the intervention which the DE helped give.	-Data use to persuade donors to support program changes -Data used to inform the public; increase credibility of intervention; demonstrate accountability (e.g., d3) -Meetings to ensure transparency and ownership to reassure stakeholders.	None reported	None reported
Non-use	- Some feedback not actioned following deliberation (e.g., to not adopt advocacy role), legitimate/warranted.	-Some findings not actioned, related to sphere of operational control, capacity to act, ‘dynamics of situation’ (c23, also c22, c12), Mixed data on warrant for non-use.	Not reported	-Partial uptake on project level theories of change -Insufficient data re. warrant for non-use.

The fourth case continues to use DE as its primary approach. With Case 4, constant change both inside and around the intervention supported maintaining a developmental purpose. A Case 4 participant described making decisions about evaluation approach based on assessment of stability:

If things are really stable, you can probably have an evaluation process that is much more stable. You set out a bunch of indicators, you have your conceptual framework, you do your formative evaluation to see if it all fits and then you move forward and collect data and do the work. Because developmental evaluation is really hard. And resource intensive. So why do it in an environment [that is stable]? (c41)

This participant qualified their statement, however, as follows: “I have to say, I don’t think I have a single environment I work in that isn’t really dynamic, I don’t know if any of us do when we’re working with human systems (c41).

Performing Roles

The above two sections of this chapter focused on activities: how they were carried out, and how processes unfolded. I also analyzed the data from the perspective of roles, and in particular on functions and positioning of the consultant evaluators on the development teams. I present these results below in comparison with two frameworks drawn from the mediation literature, following a brief note on the concept of ‘role’ in the context of evaluation.

I agree with Greene (2000b) that the concept of “role” is limited in its ability to capture how evaluators do their work because their relationships are multifaceted and dynamic (see also Segerholm, 2002). However, the concept of role remains useful here because of this study’s interest in functions and activities, and so I have used it as the primary frame for the analysis reported below. Unpacking and clarifying key roles is important for reflection and self-awareness among practitioners (D. Fleischer & Christie, 2009; J. King & Stevahn, 2002), particularly if supported by empirical study of practice. It can complement research on evaluator competencies, because some competencies relate to particular functions and activities (Stevahn, King, Ghore, & Minnema, 2005) and can thereby benefit ongoing efforts to advance

professionalization of the field. I revisit alternatives to ‘role’ suggested by Greene (2000b) near the end of this chapter by analyzing the data through the lens of the evaluator’s relationship with others, and the evaluator’s location or positioning relative to the intervention.

As noted in Chapter 2, the role of sensemaking facilitator is sometimes identified for evaluators (Dozois et al., 2010; Mark et al., 2000; Patton, 2011; Patton et al., 2016). The term “facilitator” fits well with the idea of someone enabling a productive process among collaborating groups; it also implies a nondirective neutrality in positioning (Kemmis & McTaggart, 2008b) and a role that abstains from participation in idea generation or solutioning (J. Fleischer & Zumeta, n.d.). The results presented above suggest that “facilitator” does not completely describe an evaluator who works as an active member of a program development team. For this study, I have been using “mediator” as the reference term. Mediation in the context of this study is understood as an active intervention that influences a social learning system, including the system’s actors, its structures, and the relationships between them, in order to promote adaptive learning (see Chapter 3).

Because of the variety and ambiguity of labels for roles in the evaluation literature (see Chapter 2), I chose not to code the case data by these labels. Instead, as described in Chapter 4, I coded the data inductively, foregrounding descriptions associated with turning points or trajectory in case accounts and discernable in at least three of the cases. Although evaluation functions in all four cases were shared among various actors, and this sometimes blurred the line between evaluator and program team members, I focused the analysis on the consultant evaluators. I grouped resulting codes in increasing levels of abstraction, ending at three broad categories, working iteratively and recursively with concepts from the evaluation and mediation literature. The three categories are: “technical expert” (i.e., technical support to sourcing, interpreting and reporting data); “capacity builder” (encompassing both evaluation capacity and collaborative capacity); and “guide” (actively promoting forward movement) (see Figure 6.1, below). These were not the only roles played by the evaluators but were the most consistent and visible in the cross-case data.

Mediating Social Change

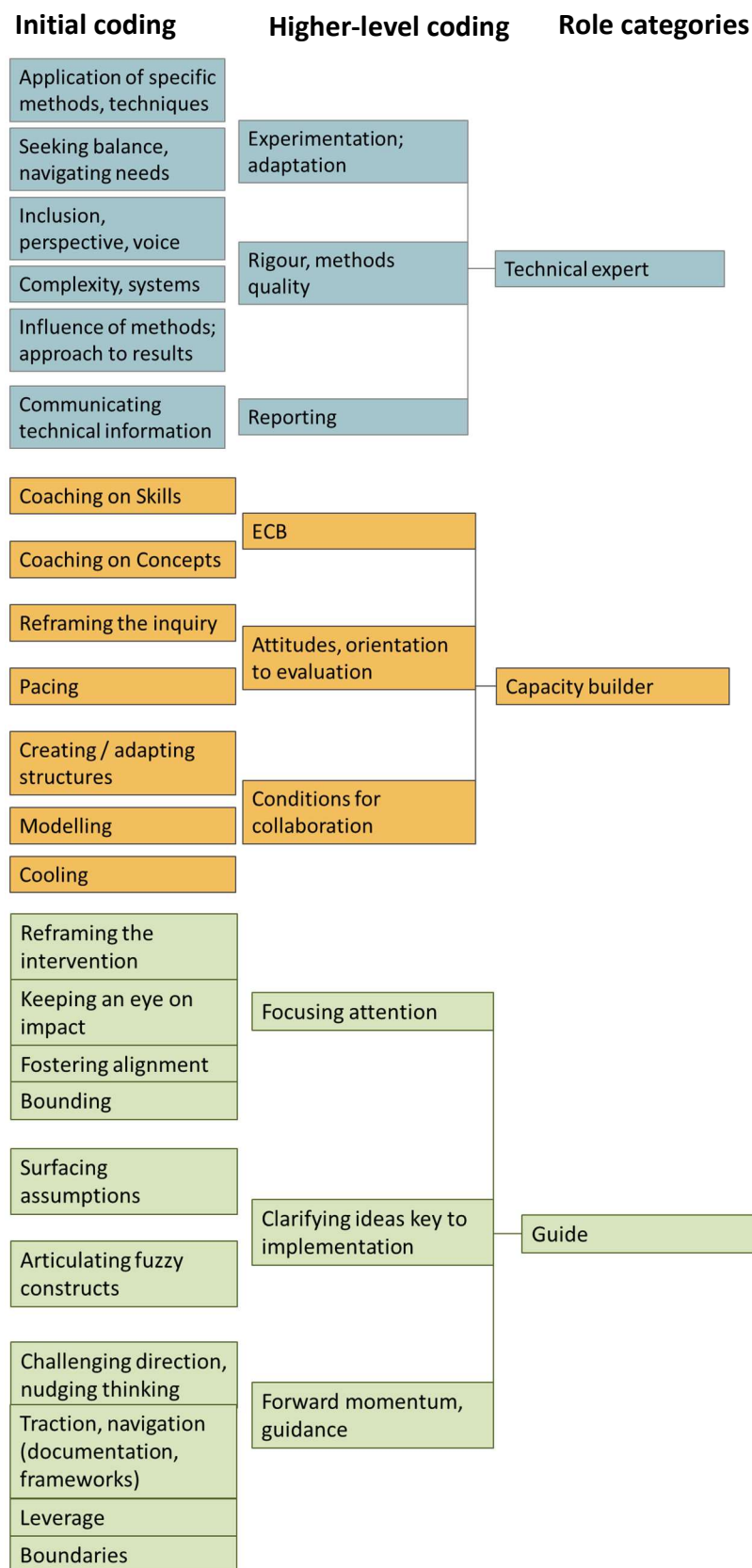


Figure 6.1. Primary roles.

I then drew on two complementary frameworks from the mediation literature for further analysis. I was not able to locate a relevant framework in the DE literature, and although frameworks for considering evaluator roles exist outside of DE (e.g., three offered by King & Stevahn, 2002), I found the mediation frameworks resonated more closely with these data. They also prompted analysis based on relational engagement, a view suggested in Greene's (2000b) critique of the concept of role. The first framework describes three broad categories of strategies that mediators employ and arrays them on a continuum. The second identifies three distinct relational targets for mediator activities. I begin by presenting the results on roles using the lens of the first framework, then the second.

Framework 1: A three-part classification of mediator strategies. Mediation strategies¹⁵ have been located on a continuum from passive to directive (for a review of related research, see Druckman & Wall, 2017). Kleiboer (1996), for example, describes three categories of strategies emerging from accumulated mediation research (Figure 6.2, see also Wall, Stark, & Standifer, 2001).

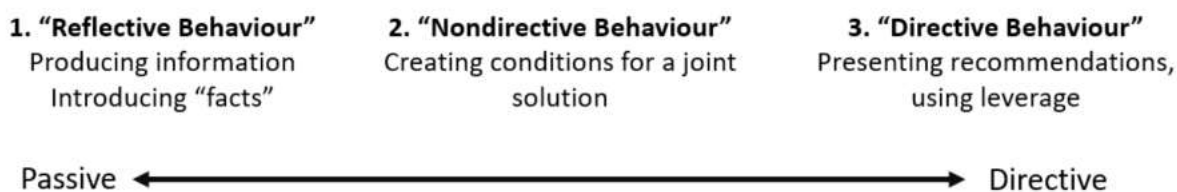


Figure 6.2. Continuum of mediation strategies

In Figure 6.2, the leftmost category, called "reflective", is at the passive end of the continuum. This category includes sourcing and providing information to the parties involved, including fact-finding and information gathering for the purpose of reducing uncertainty about the facts of the situation, and thereby correcting "distortion, ignorance, misperception or unrealistic intentions" among actors (Bercovitch, 1984, cited in Kleiboer, 1996; see also Kressel, 1972, Wall et al., 2001). The middle category involves activities to create conditions for a positive outcome by increasing the capacity of parties to work toward solutions. Strategies in this

¹⁵ The mediation literature includes studies on mediating collaborative initiatives in the public sphere (e.g., cross-sectoral environmental initiatives) and in this way touches on the domain of evaluation.

category are considered more proactive, but nondirective, for example, reducing the exposure of the parties to public pressure or influencing the structures and processes around the actors in a way that supports constructive engagement. The third category is described as “directive”. In this category, the mediator’s activities include presenting recommendations for action (i.e., in a certain direction or as a solution), and sometimes promoting the uptake of these recommendations. The mediator may use forms of leverage at their disposal to bring pressure on the parties to accept a recommendation. Finally, the research indicates that the status of a mediator determines which strategies they are able to successfully employ (Kleiboer, 1996; Wall et al., 2001). This will be discussed further below. First, I expand on the three roles identified in the case data and draw comparisons to this first framework.

Role 1: Technical expert. In all four of the cases in this study, the consultant evaluators’ formal role was to provide technical expertise. The purpose was to ensure high quality information was available to stakeholders to inform their decision-making. It corresponds to the first category in Figure 6.1, above. The role had three primary functions in these cases: to support overall methodological quality of the evaluation; to provide a rich ‘toolbox’ of methods to enable tailoring of the DE to diverse and changing needs; and to act as advocates for the data at the decision-making table. In three of the cases the evaluators also supported actors’ ability to meet commitments for outward accountability through performance of the technical expert role.

Quality. Across the cases, participants raised the importance of quality or rigour in interviews (unprompted), and this featured prominently in the content of case documents as well. Participants stressed that while DE is flexible and collaborative, it is still evaluation, especially in terms of quality of the inquiry¹⁶. There were several ideas in common across the cases with respect to quality. At a fundamental level, these included the importance of using empirical data to test assumptions and an overall stress on disciplined inquiry and critical thinking. In all cases quality was linked to the importance of aligning methods with the nature of the phenomena under study and the questions to be answered. For example, in Case 3, quality was tied in part to the meaningfulness of some forms of data versus others as indicators for understanding citizen engagement (c31). More specifically, high quality (well-constructed and focused) questions,

¹⁶ With one possible exception: a program participant in one case commented that DE is “not really evaluation” because it does not focus on summative judgements (in other words, is not summative evaluation).

well-designed instruments, regular data collection, systematic use of data, context-sensitivity and transparency of the inquiry were important aspects of quality in all cases.

In all cases, the emphasis on complexity thinking led the evaluators to encourage the inclusion and consideration of data from multiple perspectives - to the extent of helping to create structures to advance inclusion, for example the parent network in Case 2. Complexity thinking also made it imperative to attend to unanticipated issues, events or pathways of change as an aspect of methodological quality. As examples, the Case 1 and 3 teams invested in systematically identifying and then exploring alternate (unanticipated) pathways to impact.

Experimentation and adaptation: the technician's toolbox. The evaluators in these cases demonstrated significant range in their ability to adapt the evaluation to user needs, evident in the extensive tailoring of the evaluations over time. For example, methods in Case 3 were varied to fit with multiple individual projects and levels of inquiry. A participant in Case 2 described their efforts to align methods as the initiative evolved: “given that the initiative is changing, the kind of DNA of evaluation has to fit the DNA of the initiative” (c23), which included, “iterating the [evaluation] design as you go along and being flexible with it” (c23, also v12, d0). Their methods included (for example) systems mapping, surveys, analysis of quantitative administrative data, interviews, case studies and focus groups, and over time transitioned from developmental to formative (c23; c22). Where additional capacity was needed, outside resources were sometimes used, for example, to carry out time-bound summative studies in Case 4 or case studies in Case 1.

Advocating for the data. As members of development teams, the consultant evaluators were often present as decision-making was carried out. This gave them an opportunity to speak on behalf of the data, that is, to: “represent the data in the larger conversation” around the table (c32). In Cases 1, 2 and 4 accounts describe the consultants helping program actors to course correct based on evaluation feedback. In these three cases, program participants described the consultants’ ongoing advocacy for the evaluation process as well. For example, in Case 1, the consultant was credited with providing ongoing ‘reminders of the rationale’ to staff about “why it’s important for us to work this way” (i.e., to develop the program systematically) (c12). It is important to also note that there were occasions when the evaluators were *not* included in decision-making sessions. This will be revisited below.

Accountability. The consultants' voice on behalf of the data was said to support internal accountability; in the words of a program participant, to ensure for themselves the "responsible stewardship" of the resources involved (c12). With the exception of Case 3, the program teams also used the DE to demonstrate accountability outward to other stakeholders and the evaluators' work was sometimes highlighted in public documents, such as funders' annual reports, through presentations by the evaluators to community members or indirectly through the media.

The active role of the technical expert. The framework in Figure 6.2, above, labels activities like information gathering and fact-finding as 'passive' strategies. However, these case data clearly describe an active role quite unlike that implied by passive provision of 'found facts'. In case accounts, actors stress how choices of methods influence the direction of the inquiry and its findings; a high level of awareness of this is clearly evident. In all four cases, participants especially emphasized that the types of questions asked and *how* they were asked will affect the results obtained. The consultants used their technical knowledge to direct (and flex) the evaluation based on their understanding of the evaluation's purpose, the needs of stakeholders, and the complex nature of the initiative.

The evaluators also used their technical knowledge to support greater diversity and balance in the range of perspectives included in the inquiry, and to support integration of multiple perspectives at depth (see e.g., the use of runways or empowering stakeholders through question construction in Case 2). In the mediation literature, a term 'equidistance' is used to describe "the active process by which partiality is used [by a mediator] to create symmetry among the parties" (Rifkin, Millen, & Cobb, 1991, p. 153, citing Kolb, 1985). This refers to a mediator's ability to help all parties to voice their concerns or positions, particularly if they hold unequal power or resources. This idea can be extended to efforts to ensure data of range and balance in an evaluation.

Designing the evaluation to enable attention to both anticipated *and* unanticipated data or pathways to change can also be considered as an effort at creating symmetry or balance by widening the inquiry frame. This is potentially very important. In complex situations where patterns are difficult to discern, or where there is active competition among interpretations, such balance in the inquiry may be critical. One participant described such efforts as driven by a sense of humility, realizing that data would always be partial (c22). Although I focus here on

methodological rationales for the activities above, it is important to note that the case data also included ethical and values-based rationales (see Chapter 4).

Conferring status. Lastly, it is important to note that the expert role conferred status. It gave the consultants at least entry-level credibility among program actors. The consultants' status likely contributed to their flexibility to work across boundaries, giving them a place at the table with groups at different levels, from front-line to leadership and across organizations. There were limits, however, in all four cases. The consultants needed to develop their own personal credibility as well as the legitimacy of the evaluations in order to implement and sustain their work (see "Making Space", above). This is consistent with the idea in the mediation literature that multiple aspects of the relationship with stakeholders (including perceived status and trust) will define a mediator's available strategy set (Wall et al., 2001). Mediators, and evaluators, seldom have authority to effect change directly but rely on their ability to exert influence on others.

In sum, the technical role is also an active role, and in terms of the *inquiry* a directive one. Also, arguably, the performance of this role can blur a distinction between technical functions and relational aspects of evaluation practice (see e.g., Sullins, 2000, cited in Greene, 2000b).

Role 2: Capacity Builder. This is a frequently mentioned role for evaluators, usually with respect to teaching evaluation-related skills. It is often put forward as a rationale for participatory forms of evaluation (Cousins & Chouinard, 2012; Stevahn & King, 2005b) but generally considered subordinate to the evaluator's primary data-gathering and reporting responsibilities. In these four cases, however, this was the role that was emphasized most by both the evaluator and program participants. It involved efforts to increase the capacity of the actors to engage productively in the DE. It also involved capacity for fruitful collaboration more generally. It aligns roughly with the middle category in Figure 6.2, in which a mediator creates conditions that enable engagement with a process. In that framework, it is described as an active role, because it involves establishing or modifying structures and processes of interaction for the parties involved, but a nondirective one. I report details about activities related to this role under the theme "Making Space", above, however a brief outline of key points follows.

ECB. While accounts mention coaching or training on specific skills, for example developing data collection instruments (Cases 1, 4), the stress was overwhelmingly on capacity at the front end of the inquiry process (questioning) and the tail end (data interpretation and use) (all cases). In addition to skills, the consultants coached program actors on concepts related to evaluation, such as distinctions between contribution and attribution; summative, formative and developmental purposes, and to rethink their program in terms of systems (apart from Case 3, where a complex systems perspective was already present). This involved ‘unlearning’ for some actors who had specific conceptions of evaluation. The consultants actively reframed the evaluations for these actors (what it looks like, what it is for), for example from being about ‘judging what the players were doing’ to seeing the evaluation as ‘in service to them’; and ‘about stuff they are curious about’ (Case 2, c23) (all cases)¹⁷.

Conditions for collaboration: Cooling. The evaluation teams reduced pressure on program actors to create an environment more conducive to learning and development (Case 1, 2, 4). These activities correspond to “cooling” strategies described in mediation literature to help create conditions for improved trust and more productive interaction among the parties involved (Wall et al., 2001). Details are described above and included resetting unreasonable expectations for program impact (Case 1, 2), addressing fears about exposure to public scrutiny (Case 2), and facilitating ‘depersonalized’ problem-solving and solutioning (all cases).

Modelling. Participants operationalized commitments to co-creation, joint ownership and reciprocity in concrete and visible ways. Accounts attribute turning points in the cases to specific examples, for example, a process of co-developing a monitoring tool in Case 4 which contributed to a breakthrough on a long-stalemated and contentious program issue. Arguably, the teams were supporting the development of norms for collaborative engagement, also identified as a key strategy for mediators (Wall et al., 2001). Preskill and Torres (1999) include acting as a role model as part of a role of “coach” for evaluators (cited in Ensminger et al., 2015).

¹⁷ Notably, reports of a need to unlearn preconceptions about evaluation extended also to evaluators in Case 4, where it was found to be easier to train researchers to learn to think evaluatively and provide DE support than to train evaluators with experience in conventional approaches to approach evaluation differently.

Creating / adapting structures. The evaluators established or adapted structures and processes to support actors' participation. Examples include regular face-to-face meetings for collective data analysis (Case 1, 2, 4), and tailoring communications to fit organizational decision cycles (e.g., Case 1) and to make it easier to receive feedback (e.g., Case 2).

Attitudes and orientation. Part of the capacity building role involved supporting a shift in attitudes toward evaluation, including reframing the purpose of evaluation toward development and mutual learning (see "Making Space", above). Accounts suggest that pacing change, having patience, learning when *not* to push were also important (c23, v12), along with preparing the ground. A participant described working toward change slowly until conditions enable movement:

you can do little things that kind of plant things... It's not like you just hold the idea and then you jump. You're suggesting things and bringing things up and doing things and then when there's an opportunity to make a shift... everybody's sort of ready (c41).

Deliberately pacing change is also suggested by another participant who explained that they needed to be cautious about: "innovating too fast" for program actors' capacity (c42).

The data suggest capacity building was an active role of at least of equal importance to the technical expert for turning points and breakthrough moments in these cases.

Role 3: Guide. A guide "shows you which direction to walk or travel in" or "helps you to decide what to do" (Guide, n.d.). Of the three roles described in this section, this is the closest to the "Directive Behaviour" category in the mediation framework in Figure 6.2. While the other two roles focused primarily on supporting collaborative inquiry, this one extended toward the intervention itself. It was a role that involved encouraging movement in a direction (i.e., in accordance with identified program needs). It was aimed at focusing actors' attention, helping to clarify undeveloped ideas important for implementation, and giving momentum and guidance to actions. Importantly, consultants in all four cases articulated self-imposed boundaries on this role (see Framework 2, below).

Focusing actors' attention. While much of the DE effort involved 'widening the frame' of inquiry, it also involved focusing actors' attention. For example, participants described the

evaluators helping program actors to keep an eye on intended long-term impact while specific activities were being implemented (Cases 1, 2, 4 especially). The evaluation teams helped actors to reach agreement on the intervention's scope and improve coherence and alignment among components of the initiative (especially Cases 2 and 4). In Case 4, the program TOC was used as a centring tool to "bound the creative energy" of leadership in a way that improved focus (4d0). The evaluations sometimes surfaced priority issues that were already known to stakeholders, but their path forward was unclear. Participants described the use of data to point toward a solution (all cases).

Challenging actors to clarify ideas (together). In all cases, accounts describe the consultants challenging actors to clarify their thinking and the direction they were taking. This is described above in "Learning Forward", but I will note a few examples to illustrate. A "conceptual rethinking of the initiative" was attributed to the evaluations in Cases 1 and 3 which changed strategic directions. In Case 2, clarifying a key concept ("college culture") was reported as a factor leading to a reconfiguration of the program the following year (c21). The joint development and use of common tools, such as theories of change (Case 1, 3, 4) and monitoring instruments (Case 4) were highlighted as integral to the process of eliciting, documenting and then anchoring ideas in a way that gave traction and aided forward navigation of the interventions (Case 1, 2, 4). In at least one case, the evaluation team participated in delineating priorities for action based on their understanding of change processes (Case 2).

Using leverage. When the teams encountered resistance they often negotiated a way forward collaboratively by drawing on the relationships and trust that they developed with program actors, particularly later in the evaluations. Their ability to be flexible and responsive in the DE also helped them find ways forward. On occasion, outside experts were invited to provide input, and they reinforced the DE process or lent weight to specific recommendations (e.g., Cases 1, 4). In other situations, internal program champions provided assistance (Cases 1, 2, 4). In one case, the funder was asked to mandate compliance with a recommendation to strengthen program quality when other efforts were exhausted (Case 4). These are examples of the use of leverage by the teams, akin to strategies described in the mediation literature.

In sum, the three primary roles drawn from these data - technical expert, capacity builder and guide – share some similarity to the framework on mediation strategies (Figure 6.2).

However, unlike this framework, all three of the roles are active ones. The ‘guide’ role is closest to descriptions of more “directive” mediation, and touches on the program domain. The roles described above are drawn from data focused on evaluators’ activities and functions on the teams. Mark (2002) and Greene (2000b) suggest the evaluator’s relationship with others is a missing dimension in how roles have been described in the literature, and I take this up as an analytical frame next.

Framework 2: focal targets for a mediator/evaluator. The case data reflect three distinct relational foci for activities of the evaluators: (1) the relationship between themselves (as evaluators) and the program development team, (2) the relationship of the members of program development team with each other, and (3) relationships with other stakeholders not directly involved in the DE (see Figure 6.3). This is similar to a framework in the mediation literature which identifies three distinct sets of relationships of interest for the mediator: the relationship between the mediator and the parties involved, the relationship among the parties, and the relationships between the parties and their more distant stakeholders (Wall et al., 2001) ¹⁸.

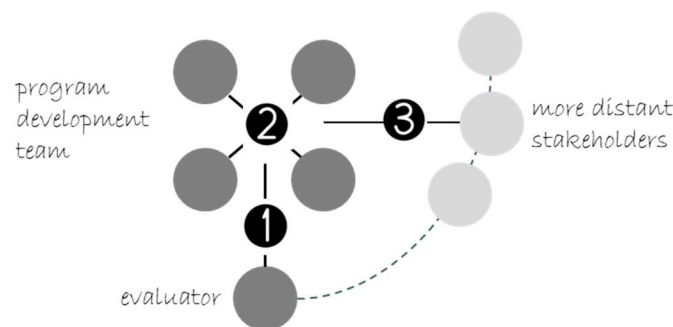


Figure 6.3. Three relational targets.

Relationship 1: between the evaluators and program actors. Case accounts describe the first relationship in terms of explicit or official positioning and also in terms of how it played out in practice (see Table 6.5). While the two overlapped for the most part, case accounts describe multidimensionality and fluidity.

¹⁸ Mark offers a model for exploring evaluator roles based on interpersonal theory in his 2002 article. This and other notable models of evaluator relationship with stakeholders, such as Cousins & Whitmore’s (1998) model of participatory evaluation, delve more deeply into the first of the three relational foci illustrated in Figure 6.3.

Table 6.5. Evaluator relationship to program actors

	Case 1	Case 2	Case 3	Case 4
Explicit structure	Internal / external hybrid	External evaluators as collaborators.	Internal / external hybrid.	Internal / external hybrid. Evaluation team part of leadership team.
Structural linkages	Three tier: internal team; primary team; external advisory panel. Multi-level ties.	Three tier: internal team; primary stakeholder team; community advisory group. Multi-level ties.	Minimal structure. Minimal formal ties to higher levels or outside of projects.	Three+ tier structure: internal team, regional teams, leadership team. Multi-level ties.
Limits on scope	Not noted.	Evaluation formally scoped to exclude internal org operations.	Evaluation scoped to exclude other relationships. Informally bounded at project/staff level.	Not noted.
Described as:	- Coach - Facilitator (of sensemaking, analysis). - A different perspective, but not really a 3rd party	- At the table - A neutral advocate - A neutral broker - A third party - 'we' and 'them'; eating tacos, 'not just a guy from outside'	- Embedded, 'inside the project' - Facilitator (design, sensemaking), Coach - Researcher, Observer - 'Not enough understanding of our world'	- Facilitator "walking alongside our program colleagues" - Involved in the shaping, trying to make it work - Not from our sector – how can they understand?
Unfolded as:	- Some staff jazzed, some staff stressed / outside comfort zone - Agency concerns about judgement - Pushback	- Evaluation a 'sensitive issue' - People traveling different distances; some outside comfort zone. - Outsider status affects credibility but also useful - Pushback ++	- Actors skeptical of evaluation, "alienating" - Outsider status (weak impact on credibility) - No explicit pushback	- Some actors skeptical of evaluation, but also leadership level champions. - Outsider status affects credibility - Pushback ++
Informal limits of role on team	- not the evaluator's role to resolve tensions, but can help to identify / surface them	- provide information, help brainstorm solutions, put people in the conversation who can assist. Client decides and implements. Wanting the intervention to succeed, but commitment to rigour and balanced feedback	- no data	- impartiality; distance; 'there are some things I wouldn't do, because I'm too involved' (e.g. summative) - helping to make meaning but not making it for program

Two of the evaluations were explicitly called internal/external hybrids (Case 1, 3). An internal/external hybrid structure refers to evaluation activities (e.g., data collection) shared between internal actors (e.g., program staff) and external consultant(s). It is described as providing dual benefits to the evaluation of deep internal context knowledge and external perspective/impartiality (see e.g., Togni et al., 2016). All four DEs were structured to include an external evaluator or evaluators. ‘External’ was understood as ‘not staff’ (e.g., c41); the consultants were described as either embedded members of the development team (Case 3, 4), or collaborators/coaches (Case 1, 2) alongside staff from the participating organizations.

In all four cases the evaluators actively worked to develop strong relationships between program actors and themselves as evaluators, for example, through extensive face-to-face interaction (all cases) and by modelling principles that supported trust and collaborative norms (all cases). In other respects, actors’ perspectives varied about their position relative to the team. For example, in one case, their ‘external perspective’ was said to be valuable and needed by the collaboration (c12), and in another case, the ‘external status’ of the evaluators helped actors to ask difficult questions and to table contentious issues in a way that staff could not (c21). External status gave these evaluators a unique position and status. Yet, the consultants were also understood to be internal to the initiative because of their active participation in its development. In at least two cases, their semi-internal position raised questions or uncertainty about impartiality (1, 4). This was twofold, as a participant noted: “There’s the reality of [impartiality], and then there’s the perception of that, and you have to take both into consideration” (c41). The desire to be assured of an impartial perspective was important enough that input from other parties (not involved in the DE) was brought to the evaluation at specific time points (Case 1, 4). For example, in one, an expert panel provided a review for the funder’s board of directors (Case 1). In another, outside evaluation consultants were brought in for summative work (Case 4).

Some accounts emphasize the integration of the evaluators on the team (e.g., as embedded, as ‘not really a third party’) and other accounts of the same case place more emphasis on distance (e.g., as a ‘neutral broker’ or ‘observer’). The idea of being both an advocate (wanting the initiative to succeed) and a critical voice was raised explicitly and unprompted by consultants in two cases (Case 2 and 4). For example, in Case 2, the evaluators were based outside the community and were described in interviews as ‘a third party’ (c21). Yet they saw

their role as members of the team, travelling regularly to the community and working “side-by-side” with stakeholders (2d20). A participant explained the evaluator ‘is at the table with the client’ (c23), and elaborated that: “you are ‘on their side’, right? [The evaluation is] not to put anybody down, it’s not to point fingers, it’s not to say ‘gotcha’. It’s about for everybody to say: ‘how do we make this better?’” (c23). The evaluation was structured to enable close communication with core team members and deep understanding of the context over a long period of time. Their dual positioning is reflected in some of the data such as when evaluator participants referred variably to the initiative’s team as ‘we’ or ‘them’; to the initiative as ‘our initiative’ and ‘their initiative’ or to youth clients of the program as ‘our kids’.

In three cases (1, 2 and 4), evaluator participants volunteered comments about bounding their role. These included, for example, surfacing implicit assumptions about the program that are causing tensions, but not resolving those tensions (Case 1), taking part in brainstorming solutions, but not implementing the solutions (Case 2), helping program actors to make meaning, but not making meaning for them (Case 4), and referring summative questions to external evaluators not involved in the DE when they felt they had been too closely involved (Case 4). These boundaries were informally drawn.

Program actors’ different perspectives about evaluation likely contributed to some of the variation in how the evaluators were positioned and how their activities were scoped. For example, the evaluator in Case 3 was described as a researcher because some saw evaluation as incompatible with their work. In Case 2, a periodic meeting that excluded the evaluators was said to be more comfortable for one of the parties who had had negative past experiences with evaluation. The potential for attitudes to evaluation to affect relationships between program actors and an evaluator is supported by literature elsewhere (e.g., Fam, 2017; Wall et al., 2001) and a reason for the efforts of the evaluators to reframe evaluation for program actors in the context of DE (see “Making Space”, above).

Relationship 2: among program actors. A focus on strengthening relationships among program actors is clear in three cases (1, 2, 4). In two of these, the need to develop trust within the collaboration was named as a factor in the decision to use DE over other evaluation approaches (1, 2) and the need for alignment among the efforts of program actors was a factor in choice of approach in the other (Case 4). In addition to efforts to strengthen relationships

generally, the DEs in these three cases were applied as a mechanism for addressing controversial issues and navigating a contentious or uncomfortable decision-making process, in a way that reduced tensions among decision-makers (see Table 6.6, below).

As noted above, the work of the evaluators and the process of DE were described by both program and evaluator participants as important for the functioning of the collaboratives. For example, the DE was said to have been aimed at creating more productive conversations among program actors in two cases (1, 2), that is, a “more egalitarian and inquisitive conversation” about learning and development instead of one that centred on lack of funds or other barriers (c12). Both the processes and the findings of the DE were said to help improve understanding each others’ perspectives (Case 2, 4), and to change “how people connect with one another” (c12). In two cases, the work was explicitly linked to development of collective identity (Case 1, 2), for example, in Case 2, a program participant reported that the DE helped the team to see themselves as: “about learning together, and that we’re about co-creating and that we’re about being transparent, and we’re about being adaptive” (c21).

The functioning of the collaboratives was also said to be enabled by improving their processes to critically assess and select among options for program direction, in a way that offered transparency (i.e, based on data and established principles, e.g., in decisions to continue recruitment in Case 1). When issues were controversial, being able to ground the process in the data helped actors to navigate uncomfortable decisions: “going back to the facts, going back to the data” (c23); and by the evaluators’ use of careful questions, efforts to clarify the rationale, and make the decision-making process transparent (Case 4). In Case 2, the evaluators were said to be able to bring sensitive issues forward because of their status as “a third party”. This served to ‘give cover’ to actors to address these issues without damaging relationships (c21). In this case, the evaluators were credited with an ability to “have difficult conversations and negotiate them and help move [the team] through when there’s discomfort” (c21).

Table 6.6. Relationships among program actors.

	Case 1	Case 2	Case 3	Case 4
Relationship development	- Creating a 'different kind of conversation', about learning. More 'egalitarian and inquisitive conversation'. (C12) - "it shifted our culture... how people connect with one another" (c12).	- 'helping them have productive conversations' (c22); 'carved out space' to have conversations in a "non-defensive way" (c23). - demonstrated a genuine commitment to learning together - DE process a 'glue' that advanced the partnership. - 'understand each others' perspectives better" (c21); help stakeholders see multiple perspectives. (c22)	- No / limited data. (Use of research language to gain acceptance; use of the DE to assess collaboration strategy and strategy /partnerships changed)	- helped project actors reach a collective understanding of the program - built coherence among interpretations (d29)
Mechanism for critical debate	A mechanism to critically assess competing (contentious) options for strategy and resolve them with: -data -established principles of engagement for the collaboration (d-er).	- Got issues on the table a vehicle for those kinds of things to be put forward by a third party (c21). 'gave cover' (c21). - Being able to ground the work in the data when issues are controversial (c23) - being able to "have difficult conversations and negotiate them and help move through when there's discomfort" (c21).	- Limited data (Tensions among collaborators observed, but explicitly scoped outside of evaluation)	A mechanism to help actors navigate difficult, uncomfortable decisions (e.g., ending funding to long-term grantees).

It is important to note that evaluation can *add* to tensions in a program team, (see e.g., Fleischer & Christie, 2009). Case accounts point to this, for example Case 4 actors questioning why the evaluation was “stirring the pot” (4d0). As an interventionist approach, DE may be especially likely to surface tensions - sometimes intentionally – making attention to relationships among program actors more important. In general, mediation research suggests that a mediator is more effective when they focus on improving the relationships among the parties, “rather than pushing for settlement or focusing on facts” (Wall et al., 2001, p. 383).

Relationship 3: between program actors and their stakeholders. Important relationships also existed outside the program development team in all four cases. Case 4, for example, included a partnering funder whose support doubled the program’s scope and increased its need to be “data rich”. Case 2 included an extensive array of secondary relationships in the community, described as “a web of public-private partnerships that goes in every direction” (2d35, also 2v8), to which the foundation saw itself as accountable.

In three cases (1, 2, 4), the evaluation was structured to include participation by groups of stakeholders at multiple levels (at least 3 levels) giving the consultants direct connections to or interaction with actors at multiple levels in the interventions. For example, in Case 1, the consultant coached the internal staff team, participated in activities with partner agencies, and chaired the expert panel which provided advice to the funder’s Board of Directors¹⁹.

The presence of a technical specialist or expert on the program development team provided assurance of due diligence to more distant stakeholders in these three cases (1, 2, 4) and communications developed through the evaluation helped increase understanding of the intervention by actors not directly involved in its development (Cases 1, 2). In some cases, communication pieces were used as boundary objects, that is, their design made them useful for communicating to more distant stakeholders elsewhere in the ‘system’. For example, visual models of the TOC in Case 1 were shared across functional groups and “helped people talk with others about [the initiative]” (1d11). In this way, communications played a role to help stakeholders understand what was happening as the intervention developed, and to hold space for DE among those who had influence over it (i.e., in leadership) but were not actively involved at

¹⁹ The evaluation in Case 3 did not have a formal multi-level structure and had fewer direct interactions between the evaluator and program actors outside the core team.

that level. In two cases, participants noted that recommendations could be difficult to implement when they affected structures or budgets outside the direct control of the development teams (Case 2, 4). The evaluations were credited in three cases (1, 2, 4) with the teams' ability to secure increased or extended funding from outside groups. These examples highlight how attention to more distant stakeholders may influence progress. This can in turn affect how the evaluation is valued by the program actors expected to engage with it.

Case accounts also suggest attention was paid to less powerful stakeholders outside the program development team, particularly clients or intended beneficiaries of the interventions in the local communities (Cases 2, 3, 4). This was manifest through, for example, how the evaluations were designed, the emphasis on developing capacity for systems thinking (Cases 1, 2, 4), and commitments to inclusion and reciprocity for both methodological and ethical reasons (see above "Making Space").

Closing note. Overall, the data suggest a flexible set of complementary roles played by evaluators in these cases, at times formally and explicitly, and at times informally and responsively. Although I approached the analysis to consider the performance of roles, in some respects the data indicate that evaluators both positioned themselves and *were positioned* variably depending on need and situation. The roles therefore involved navigation and attempts at balance and compromise.

Chapter 7 Group Concept Mapping

Overview of Strand 2

In the second strand of this study, I used Group Concept Mapping (GCM) (Trochim, 1989, 2017) to explore DE practice from another perspective. The GCM process was conducted in parallel with the multi-case study.

Details of the design and the procedures followed are available in Chapter 4. In brief, participants provided a list of statements about important DE activities. I supplemented this list through content analysis of other DE case studies (published case studies that were not part of Strand 1). A second set of participants then sorted the statements into conceptual groups and rated them for priority to DE practice. The sorting results were analyzed using multidimensional scaling (MDS) and hierarchical cluster analysis (HCA) to produce a concept map. Lastly, a third group of participants provided their interpretations of the map. I returned a summary of this interpretation to the sorting participants for feedback. In total, 21 individuals contributed to the GCM process.

In this chapter, I first outline the results of the statistical analyses, and then follow with the findings from the interpretation steps. The chapter closes with a summary discussion.

Results

Point map. Figure 7.1 displays the point map produced from MDS analysis of the sort data. In the point map, each statement is displayed as a numbered point in 2-dimensional space. The point map uses the distance between the points to describe how statements were sorted. Statements sorted frequently into the same group are located closer together; the distance between points therefore carries information about how participants perceived statements to be related. A full statement list is in Appendix G.

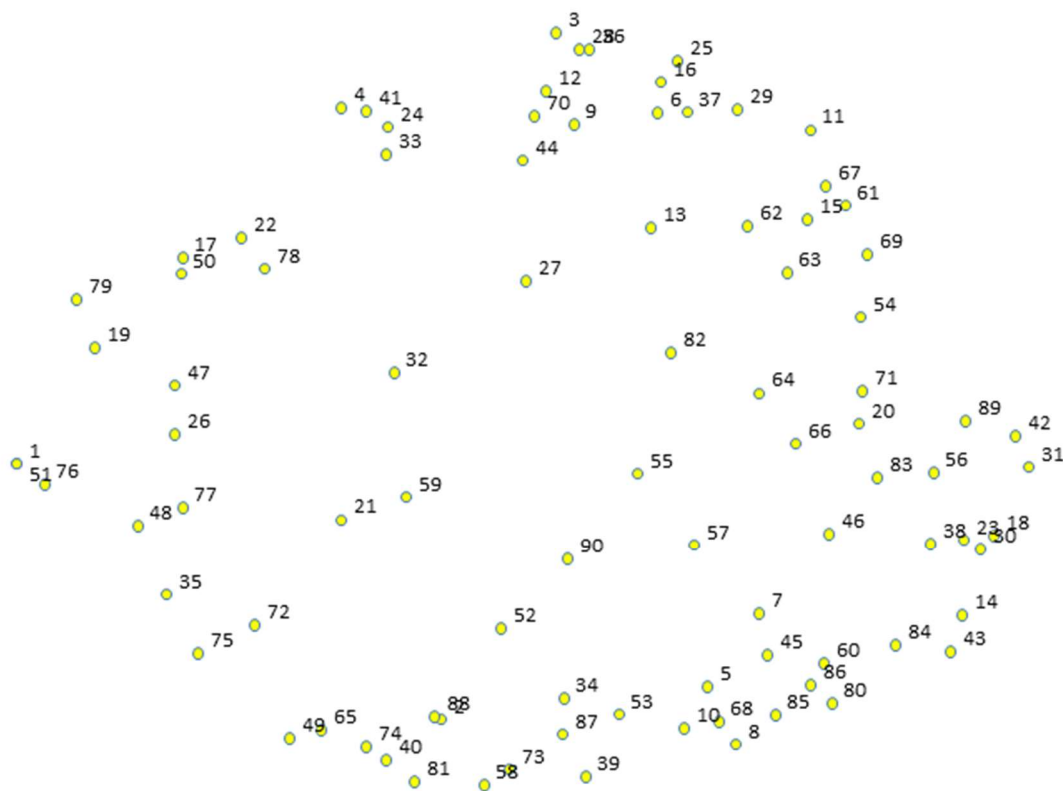


Figure 7.1. Point map output of the MDS analysis.

A goodness-of-fit diagnostic statistic, called the stress value, is used in GCM to indicate how well the distance matrix (which underlies the point map) corresponds to the aggregated similarity matrix of the participant sort data from which it is drawn. A lower stress value indicates better ‘fit’ between the plot and the original sort data and is an indicator of higher agreement among individual sorts²⁰. There is no agreed cut-off score for stress. What is acceptable depends on the purpose of the study. Values between 0.16 and 0.35 are considered typical and reasonable for studies of social phenomena (i.e., that are not highly controlled) (Kane & Rosas, 2017; Kane & Trochim, 2007; Rosas & Kane, 2012; Trochim, 1993)²¹. Final stress for these data was 0.31. This is within the reasonable range,

²⁰ Less interpolation is needed to plot the points based on the similarity matrix when there is high agreement in sorts. Agreement in sorting can be affected by the complexity of the concepts and/or how participant interpretations of statements vary (Kane & Trochim, 2007; Petrucci & Quinlan, 2007).

²¹ Rosas and Kane (2012) note that when stress value is below 0.39 there is a less than 1% probability that the map has no structure, or the structure is random.

especially given the content of the statements, which includes a number of concepts and metaphorical language that are open to different interpretations (e.g., “be a credible instrument” and “illuminate insight”). The stress value indicates acceptable agreement between the individual participant sorts, that is, that the individual participants sorted the statement list with sufficient consistency to provide a basis for further analysis.

Concept map. A concept map was produced through HCA and is shown in Figure 7.2. I followed a procedure described by Kane and Trochim (2007) to produce a map of 7 clusters (see Chapter 4). The numbers inside each cluster correspond to individual statements as displayed in the point map above.

Each cluster contains statements that were sorted frequently together indicating conceptual similarity. Adjacent clusters are likely to be conceptually more closely related than clusters that are further away in the map. For example, clusters 3 and 4 should be more similar in their content than 3 and 1. The labels assigned to each cluster are a product of the interpretation step, described below.

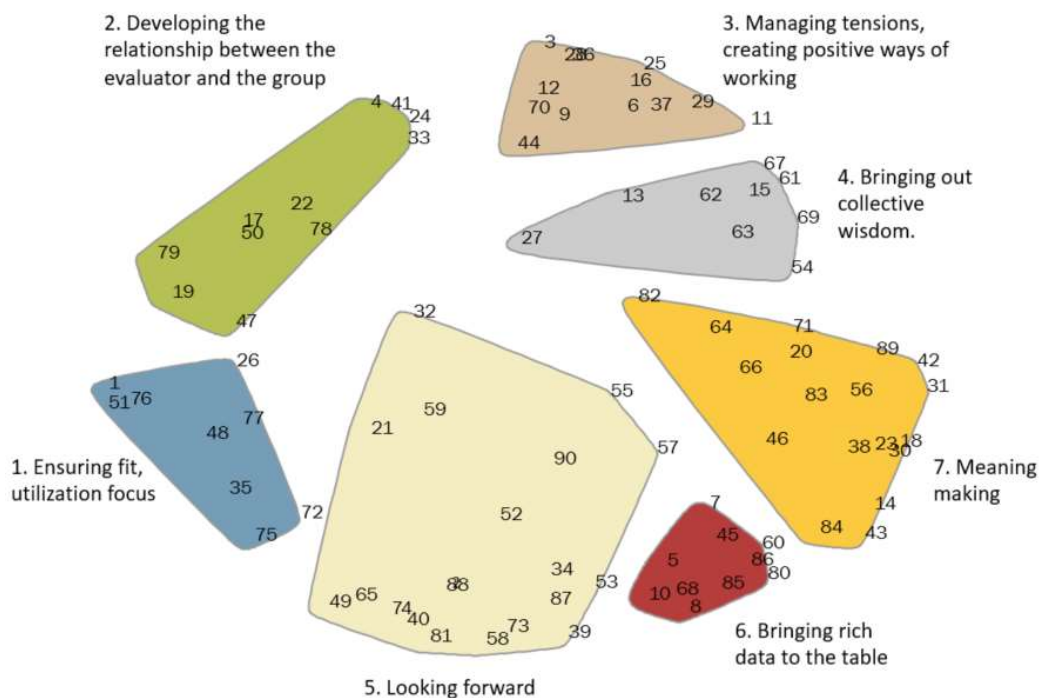


Figure 7.2. Concept map of DE activities.

Bridging and spanning analysis. Statements with high ‘bridging values’ were sorted variably by different participants together with statements in other areas of the map. Six statements have bridging values above 0.80, circled in Figure 7.3, and identified in Appendix G. These statements potentially ‘bridge’ or link different clusters. They may also represent ambiguous concepts or points of disagreement among sorting participants. As an example, the relationships between Statement #26, “Be a good strategy coach”, and other statements are illustrated in Figure 7.4.

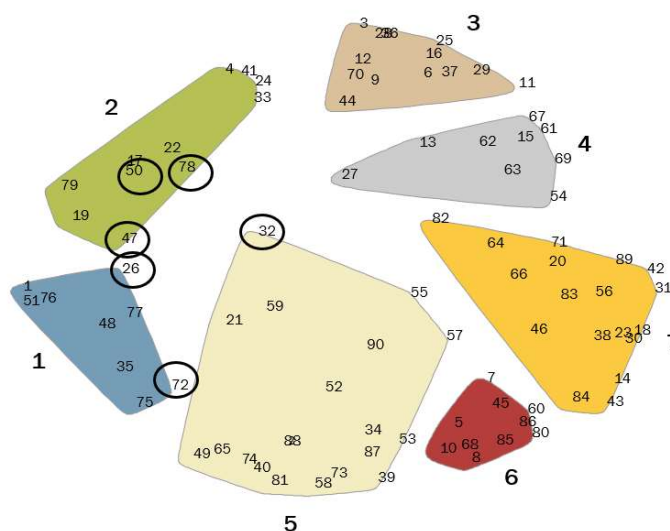


Figure 7.3. Bridging statement locations.

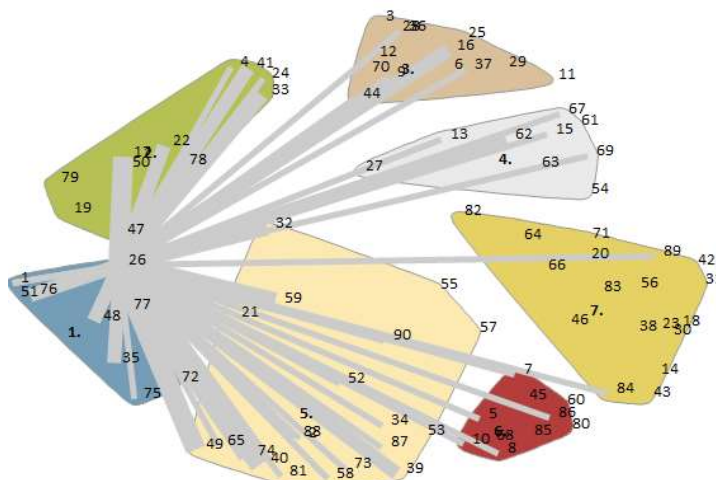


Figure 7.4. Spanning diagram illustrating relationships of Statement #26 “Be a good strategy coach”.

Overall, Clusters 1 and 2 have the highest average bridging values in the map, although the value for both clusters, 0.63, is near the middle of the possible range of 0-1 (Dave et al., 2017). These values are not high enough to warrant revisiting the cluster analysis but were sufficient to bring forward to participants for their interpretation of meaning. This is reported below. Cluster 3 has the lowest bridging value (0.23). This suggests that Cluster 3 is the most internally cohesive of the 7 clusters of activities. In other words, the sorting participants interpreted the statements in Cluster 3 with the most consistency.

Statement ratings. The sorting participants rated each statement on a 5-point Likert-type scale²² in response to the question: “In your experience, how often is this a priority?”. Only one statement was rated below 3.0 (sometimes)²³. Most of the statements (78%) had an average rating of 4 (often) or 5 (always). The high ratings by this second group of participants affirmed the relevance of the statement set, which had been generated by the first participant group and content analysis.

Figure 7.5 illustrates mean rating by cluster. This is the mean or average rating across all participants for all statements in the cluster. Clusters with higher rating are shown with more layers. The range is limited, from a mean rating of 3.99 (Clusters 1 and 3) to 4.34 (Cluster 2). I therefore used a paired samples t-test to better understand whether participants were making a distinction in their ratings at a cluster level. The difference in mean participant rating of the statements in Clusters 1 and 2 is significant, as is the difference in rating of statements in Clusters 2 and 3 ($t=-3.54$, $p<0.01$, 99% CI and $t=2.59$, $p<0.05$, 95% CI, respectively), suggesting that participants were discriminating between the content in these clusters. Although the small size of the ratings differences limits further interpretation, participants in the interpretation phase were interested in the high ratings of Clusters 2 and 4 relative to Cluster 3, and this is discussed below.

²² 1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Always

²³ This statement was “Be clear about what kind of evaluation they are doing. Not drifting back to formative or summative without intending to”. The average rating was 2.88 (Cluster 1).

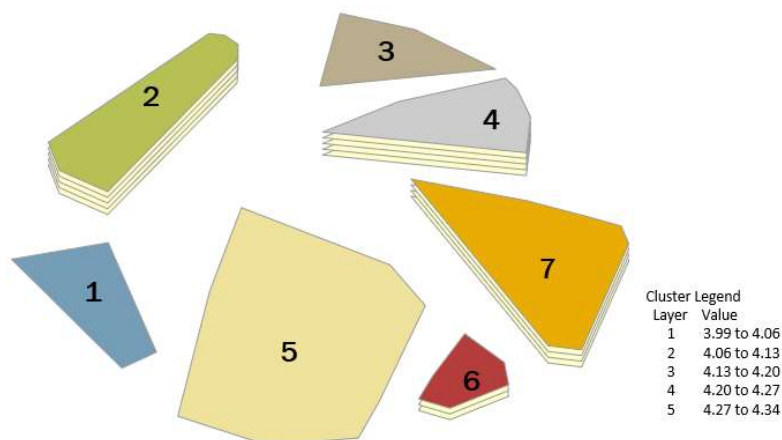


Figure 7.4. Cluster-level ratings.

Interpretation of the Results

The map was presented to a third set of experienced DE practitioners for interpretation. Four participants discussed the results as a group in a 1-hour online panel. A fifth participant gave input via a 1-hour individual interview. Participants were asked about the resonance of each part of the map with their experience. They were also asked to consider relationships among clusters, to identify aspects of practice missing from the map and about implications of the results for DE practice.

Participants in both sessions easily connected the map's contents with their practice and arrived quickly at similar names and descriptions for the clusters. As mentioned, Figure 7.2, above, shows the cluster names; Table 7.1 below, presents cluster names and descriptions based on these data, with example statements from each cluster.

Apart from the content of individual clusters, most of the interest and 'energy' in both interpretation sessions centred on the interrelationships that participants identified in the map. Themes emerging from these data are described below. I begin by describing proposed macro-level "regions" in the map, followed by suggested temporal relationships, and the implications participants raised for practice.

Table 7.1. Names and description of the clusters.

Cluster		Description	Example Statements (Statement Number)
1	Ensuring fit, utilization focus	Ensuring fit of DE with the purpose and goals of stakeholders. Establishing stakeholder understanding and expectations about the DE process. Aligning with initiative and decision-making processes.	Ensure DE will be a good fit for the purpose of the evaluation (1). Ensure the people involved are ready / really wanting to do DE (51). Create some kind of framework to start with, pulled out of what the project says they're doing (75). Iterate the evaluation design (35).
2	Developing the relationship between the evaluator and the group	Building trust, establishing the parameters of the evaluator's role (e.g., coach, facilitator, co-creator), the evaluator's relational position, evaluator skillset.	Develop good trusting relationships with people around the table (4). Have cultural competence to bring to the data collection (22). Be a credible instrument in the evaluation process (17). Collaborate with clients on an ongoing basis (79).
3	Managing tensions, creating positive ways of working	Creating an enabling environment and appropriate ways of working together by navigating and facilitating group dynamics.	Figure out the sources of tensions in the group (3). Mediate power dynamics (36). Offer actors in a collaboration a new way to operate with one another (70). Have honest and challenging conversations (37). Help move the group through when there's discomfort (25).
4	Bringing out collective wisdom	Questioning, listening, focusing attention and mirroring back. Inquiry to illuminate insights, how individuals understand things, "what's our common space?", holding collective space.	Listen and pay attention to what is going on (13). Help to illuminate insight of people around the table (67). Know the right questions to ask at the right time (63). Facilitate conversation through questions (69). Help actors to see their work differently (54).
5	Looking forward	Exploring pathways, grounding decisions in data, tying to existing strategy, helping to hone the strategy.	Frame up emerging working models of the intervention (65). Help to course correct (74). Point out where decisions are being made without (or despite) considering key data (2). Help to navigate the path forward (49).
6	Bringing rich data to the table	Using evaluation tools; formal data and feedback.	Provide rich feedback from the field (5). Use evaluation tools to bring more certainty to stakeholder hypotheses; walk on more solid ground (68). Bring data to the table to support conversations (80). Gather data from multiple sources on an ongoing basis (85). Encourage openness to new sources of data and perspective (60).
7	Meaning making	Enabling collective analysis, reflection and learning; highlighting patterns; pulling together 'what have we learned'?	Help the team to spot new patterns (30). Facilitate participation in analysis of data (18). Support collaborative meaning making or sensemaking (38). Provide a systematic structure for reflection (83). Develop evaluative thinking (66).

Two hemispheres. At a very broad level, participants described two ‘hemispheres’ in the map: one focused on relationship-building, the other on ‘grounding decisions in data’.

Grounding decisions in data. Clusters 5, 6 and 7 in the lower right of the map were described as “in conversation with each other” (CMP-S)²⁴. This conversation is about: “helping to ground decisions in data” (CMP-S), with Cluster 6 at the centre described as “the hard stuff” (CMP-W). Clusters 5 and 7 were both described as places where learning through the evaluation is accumulated or “pulled together” (CMP-W). Yet the clusters also differ: Cluster 7 is oriented to reflective meaning-making and Cluster 5 to prospective testing of ideas and navigating a path forward.

Building relationships. The clusters in the upper half of the map (2, 3, and 4) were also described as ‘in conversation’ (CMP-M). In the words of one participant: “the top half of this whole thing, is [about] building relationships, facilitation, mediation” (CMP-S). The highest frequency-of-priority rated clusters are in this area of the map. That such a sizeable proportion of the map is oriented to relationship-building was said to demonstrate that these activities are “just as essential” in DE as the more “hard core” activities, such as data collection and data analysis in the lower half of the map (CMP-S).

Three regions. Although relationship-building connects Clusters 2, 3 and 4, participants elaborated by describing a distinction among them. Cluster 2 was described in terms of relationships *between* the evaluator and the program development team, including defining the evaluator’s role or position relative to the team. Clusters 3 and 4 are about relationships *within* the program team, oriented to intra-group facilitation and mediation. Notably, a focus in Cluster 2 on roles and “relationships between” is consistent with its location next to Cluster 1. Cluster 1’s activities centre on the ‘fit’ between DE (as an approach) and the intervention. This suggests three ‘regions’ of focus for activities in the map, illustrated in Figure 5.5.

²⁴ Participant identities have been withheld for confidentiality; names are cross-referenced with these codes in my records.

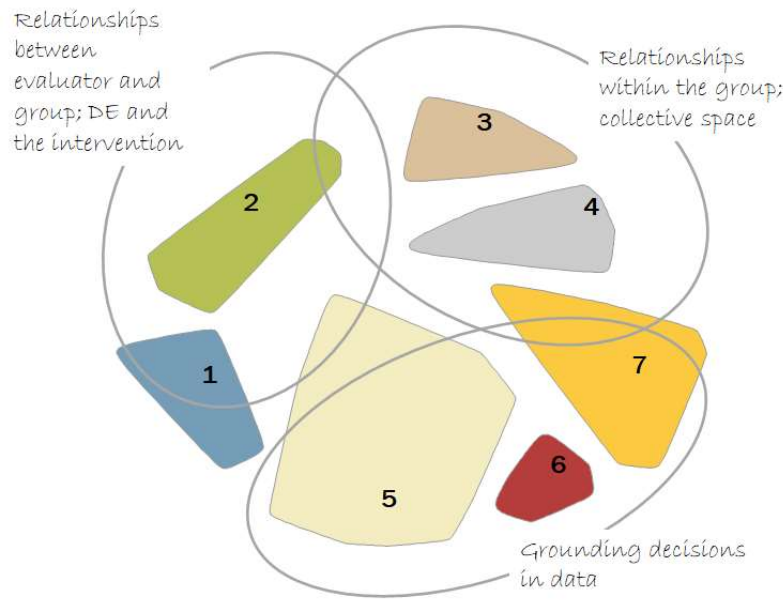


Figure 7.5. Clusters in conversation: 3 ‘regions’ of activity.

The metaphor of ‘space’ was raised in the panel discussion (e.g., “holding collective space” (CMP-M)) and I used this as a lens for interpreting the map. From this perspective, activities relate to space in three ways: scoping a space for inquiry (i.e., defining and aligning boundaries, roles) in Clusters 1 and 2; developing and holding collaborative space in Clusters 3 and 4; and using or leveraging that space to develop the initiative in Clusters 5, 6 and 7. This is illustrated in Figure 7.7, below. Compared with Figure 7.6, above, this view shifts the emphasis slightly from people to situation. Cluster 1, which had been left partly outside the three regions in Figure 7.6, is now in the centre of the ‘scoping space’ region, because of the focus in Cluster 1 on fitting DE to the context. Crossover in the lines drawn around each region are intentional, to emphasize that these are open and dynamic boundaries; the sets of activities overlap and interrelate.

Temporal relationships. Participants in both sessions described how a developmental evaluation unfolds over time through an iterative back and forth conversation between these clusters of activities.

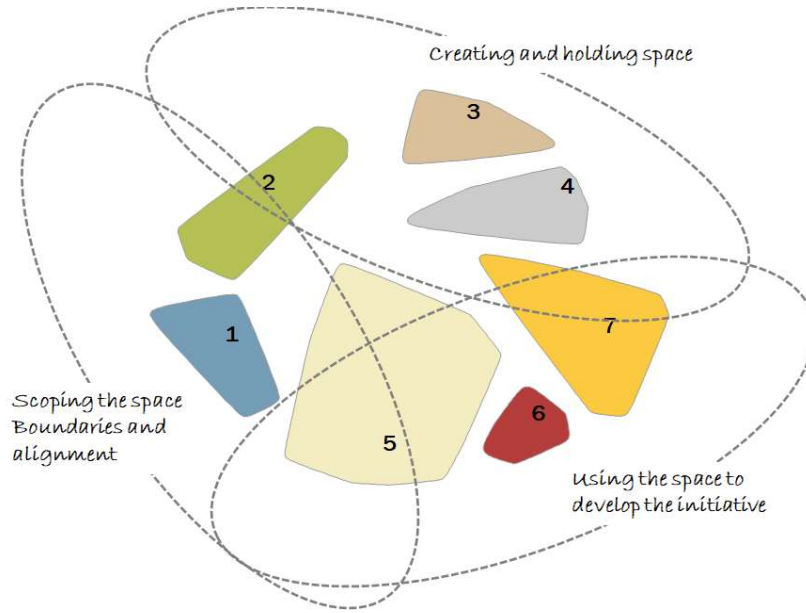


Figure 7.6. Clusters in conversation, metaphor of ‘space’.

Foundational work. Foundational work was said to reside in Clusters 1, 2 and 4. Participants in both sessions emphasized the importance of Cluster 1 (Prioritizing fit) early in the evaluation process. These include ensuring fit of DE to the purpose of the evaluation, that people really want to use DE, and that there is good understanding and clear expectations for a DE process. Establishing an understanding about purpose, goals and the DE process was said to be especially important if the group’s experience has been with other forms of evaluation: “otherwise you’re opening yourself up to all sorts of problems” (CMP-K). Stakeholders might not have had experience with DE, and DE will “look and feel and be implemented in a different way” which can run counter to expectations (CMP-K). Spending time here is also important when groups have had negative past experiences with evaluation that may affect their willingness to engage; it is “important/empowering to say this is a little different, stick with us” (CMP-N). Cluster 1 also comes into play again later in the evaluation process, as noted below.

Cluster 2 is where the evaluator “establishes the parameters of [their] role” (CMP-W). This also needs to be done early and done well because it is essential to trust: “If you don’t establish those roles very clearly you can violate trust” (CMP-W). However, it is important not to assume that, once established, the work of Cluster 2 is over. As with Cluster 1, Cluster 2’s activities may need revisiting as the evaluation unfolds.

In both sessions, participants placed Cluster 4 (Bringing Out Collective Wisdom) next in a temporal sense. Cluster 4 builds on the work of Cluster 2, because it establishes “a way of working... that helps to define and clarify the relationship... it provides a foundation for the way that you’re going to work, as the whole initiative unfolds” (CMP-K). Cluster 4 activities need early attention: “if [collective space] is nailed in the beginning, in the first month or two... everything else becomes easier... I hate to use the word easy, because our work is not easy – but easier to execute” (CMP-M).

The inquiry space. Cluster 4 is also the point at which the group moves into “the inquiry space” (CMP-W). Cluster 4 was described as about “asking, listening, questioning, exploring” (CMP-W), as well as “framing” (CMP-W), and “focusing attention and reflecting back and allowing the group to see themselves in a different way” (CMP-S); this last participant used metaphors of a mirror and magnifying glass to describe the work in Cluster 4. This type of work can surface tensions:

Generally, it’s in the Cluster 4 activities that differences of perspective, and differences of goals, and differences of needs – in other words, the tensions – start to emerge, so even though you’ve done this work establishing your role in the beginning... in your inquiry it becomes transparent that different stakeholders have different perspectives, different power, different needs, and you have to mediate those (CMP-W).

Cluster 4 was described as a “good conversation starter” with respect to ways of working collectively, but these are tested when tensions surface (CMP-N). Cluster 3 (Managing Tensions), was said to “emerge through Cluster 4” (CMP-W).

Participants’ views were mixed about how often Cluster 3 may become a priority. The likelihood and degree of conflict were felt to vary among projects and depend on the diversity of the stakeholder mix (CMP-W)²⁵. When successful, the work of Cluster 3 can help to sustain and deepen the work of Cluster 4. However, sometimes the group may need to “back out [of the inquiry process] and re-establish roles” by returning to Cluster 2 (CMP-W). In a sense, Cluster 2

²⁵ Discussing the lower mean rating for Cluster 3 in terms of how often it is a priority, a participant wondered if sorting participants may have self-selected into this study because they’ve had especially good (low conflict) experiences with DE, introducing a bias. Cluster 3 also has the lowest average bridging value; its activities may be more ‘specialized’.

is at a different level: “it’s a meta level, because you have to establish it in the beginning and you have to nurture it through basically [Clusters] 3, 4, 5, 6 and 7” (CMP-W). Like Cluster 1, Cluster 2 is positioned ‘outside’ the inquiry and helps to define and to bound the inquiry. This may explain why Clusters 1 and 2 have the highest bridging values in the map (see above).

Developing the initiative. A relationship or dialogue between Clusters 4 and 5 was described in both sessions. For example, ‘rapid experimentation’ is an activity in Cluster 5 that begins in Cluster 4, where “how to even think about that” is established (CMP-K). This relationship between 4 and 5 was described as a “reciprocity” in the words of one participant: “between inquiry, decisions, stepping forward, feedback, inquiry, decisions, stepping forward. So in itself it’s an integrative cycle” (CMP-W). Feeding this cycle are the data gathered in Cluster 6 and the meaning making from Cluster 7 (CMP-W).

Maintaining alignment. Lastly, from a temporal point of view, the neighbouring clusters 1 and 5 also interact. Cluster 1 includes iterating on the evaluation design and maintaining a utilization focus. As the evaluation unfolds, alignment with decision-making processes (Cluster 5) is key for the evaluation to be useful (CMP-K).

Implications. ‘Idea streams’ and distributed practice. As noted above, the ideas expressed in some individual statements may act as ‘bridges’ between clusters. When discussing the bridging statements, a participant suggested a metaphor of an ecosystem of ideas, or of ‘idea streams’ that may flow through the map (CMP-M). This led to discussion about distributed practice.

An activity described in a statement may, in practice, be distributed across functions and time in the way that it is carried out, and hence be relevant to multiple areas of the map. To illustrate, a panel participant suggested how a developmental evaluator helps a team to prototype their ideas for an intervention. The panel members agreed that the *process* of this resides mostly in Cluster 4, but its *purpose* is in Cluster 5 where actors would see its concrete manifestation. Cluster 4 includes helping actors think differently about the intervention, and Cluster 5 includes experimentation and framing up working models. This participant described why prototyping ideas spans both clusters:

[It is a] challenge for some organizations, how they're thinking about the problem, how they're thinking about their proposed solutions and how you can actually move towards an approach where they can actually learn from them... it's being open to the idea, and *then* it's moving to how to put that into practice and saying ok so how would we test that? (CMP-K).

Another example is the bridging statement: *“Take time to get really clear on the focus of the program and how it all fits together; how do the actors think change happens?”* (#32). A participant described how clarity on theories about change would develop over the course of the evaluation. It would begin in early discussions between the evaluator and program actors in Cluster 1 but be developed in Cluster 4 where ideas and assumptions are explored. Cluster 4 operates in ‘conversation’ with Cluster 5 where ideas about theories of change are tested (CMP-W). Clusters 6 and 7 ‘feed’ this iterative conversation between Clusters 4 and 5 (CMP-W). This is illustrated in Figure 7.8.

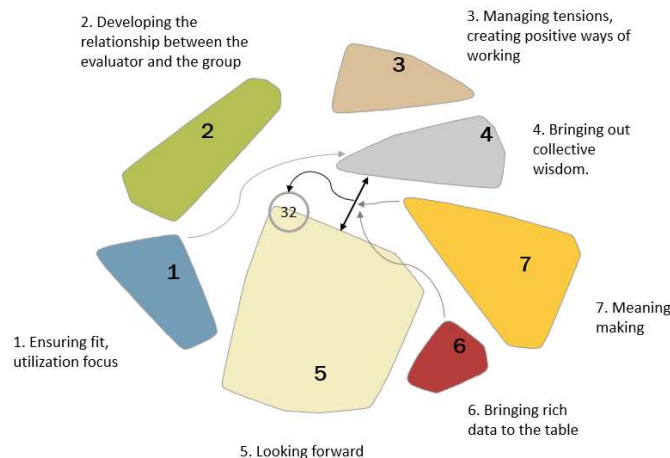


Figure 7.7. A proposed process for statement #32.

The invisible DE. Importantly, participants noted that we may under-appreciate the kinds of activities like the one reflected in Figure 7.8 in real life. Where activities or functions are distributed in a practice, the breadth of what they involve may be more difficult for stakeholders to recognize. This can contribute to some aspects of DE practice being ‘invisible’ (CMP-S; CMP-M). This proposed characteristic of DE may make it more difficult to understand than some other approaches:

When you're doing traditional evaluation, it is a much more linear process – you look at the program, you either collected baseline data or you don't have baseline data, you collect your process data or your outcome data and you do your report. Developmental evaluation is so much more ambiguous and complex” (CMP-W).

A whole cluster of activities in DE might be hard to see as well. For example, a participant suggested the activities in Cluster 3 may be under-appreciated by stakeholders because a developmental evaluator may be diffusing tensions before they become visible, in other words, “untying knots before they happen” (CMP-M). Roles (Cluster 2) may be obscured by DE's ambiguity. Although often considered a member of a collaborating team, developmental evaluators are “separated somewhat” from the team (CMP-W), in part because of the technical skills they bring and their simultaneous positioning as coach or facilitator (CMP-W). Negotiating the role can be more difficult if the evaluator has both internal and external commitments, representing a dual role (e.g., as a member of an internal program development team and for reporting out to a nonparticipating funder) (CMP-W).

Participants suggested that these features can make DE challenging for program stakeholders to understand, and even for other evaluators to understand. It also makes it difficult for DE practitioners to articulate and teach.

Locking down. Participants also noted that ‘design thinking’ was not explicitly mentioned, which was a surprise, nor was ‘locking down’ what has been learned for use in the longer term. In the course of this discussion, aspects of both were noted by panel participants in Cluster 5. With respect to ‘locking down’, the interview participant described both Cluster 5 and Cluster 7 addressing aspects of ‘accumulating learning’ for use in the future. With respect to design thinking, this has recently come to be associated with innovative approaches in social programming (Liedka, Salzman, & Azar, 2017), and social innovations are considered a niche focus area for DE. Its location in Cluster 5 (Looking forward) aligns with the action orientation of design thinking.

A need for guiding principles. The panel participants noted that none of the statements in the map explicitly refer to guiding principles or guiding values, and this was raised as a potential gap. Writing on DE, Patton (2016a) defines a principle as “a fundamental proposition that serves

as the foundation for a system of belief or behavior or a chain of reasoning” (p. 22). A participant suggested that the lack of explicit reference to principles may be because the question used to generate the idea statements for the map was worded to focus on actions (CMP-K). Although the process to build the map was indeed aimed at the level of concrete activities, the importance of principles to these activities was stressed by panel participants, because they help navigate the challenges that arise in DE over time (CMP-N). For example, Cluster 4 activities such as: “enable conversations that explore ambiguities and push for insights in a space separate from focus on decisions or closure” (Statement 15), were described as incomplete without reference to guiding principles: “for me I would want to have guiding principles to deal with that complexity” (CMP-M). The ‘fluidity’ of DE is part of the need for principles; they offer guidance as situations change, and in “times of tension or times of confusion or times of pressure” (CMP-N). This extends beyond principles of the evaluator to principles of the collaborating group. Principles for the group were described as “the principles that would guide you together” (CMP-M), and were linked back to Cluster 4’s activities and establishing a ‘collective space’ and ‘ways of working’, which can include asking “what are the principles that we agree on?” (CMP-M).

Summary

The GCM process produced a concept map of important activities for DE. These activities are organized into 7 interrelated clusters in the map. Two clusters provide a space apart from the active inquiry of the evaluation (Clusters 1 and 2); efforts here are important for positioning and aligning DE with the initiative and defining the parameters of the evaluator’s role. A space for collective inquiry is established and sustained through relationship-building within the group, which includes mediating tensions as they surface (Clusters 3 and 4). This collective space is leveraged to develop the initiative, through rich data collection and collaborative meaning making; these are applied to navigate a path forward that is ‘grounded in data’ (Clusters 5, 6, 7). In this respect, the map’s regions can be viewed as three layers of activities: a foundational layer where the inquiry is scoped, bounded and aligned; a relational layer where capacity for collaborative inquiry is created and sustained; and a third layer which contains most of the activities usually associated with evaluation by stakeholders. This last layer produces the tangible outputs, including raw data, results of analysis, and recommendations for program decision-making.

The activity clusters are nonetheless mutually reinforcing. Pathways through them are described as nonlinear, and iterative. The map acts as a lens to illuminate types of activities in DE practice that have essential functions for the whole yet can be difficult to ‘see’ in real life.

The map compartmentalizes activities that in practice are executed and experienced in a deeply interconnected way. This may be seen as a limitation. However, like a foil, the contrast serves effectively to highlight this quality of DE and offers a basis to describe the ‘flow’ of the practice. This ‘fluidity’ of DE presents challenges for stakeholders to understand DE and for students to learn DE. It also challenges DE evaluators, reinforcing a need for positive and productive ‘ways of working’ to help the group, and the evaluator, to see their way through.

Part IV. Integration

In the final stage of this study, I systematically integrated the two strands to check, develop and extend my findings. I followed my research questions to conduct the integration analysis in three steps. Two of these steps focused on my first research question. I report these results in Chapter 8. The third step centred on my second research question, which asks about implications for adaptive learning. I present these results in Chapter 9. Finally, Chapter 10 concludes the thesis with a summary of the study's most important findings, notable limitations of the study, and an agenda for future research.

Chapter 8 Integration of the Strands: Roles and Activities

In this chapter, I report the results of the first two steps of the integration analysis. These findings further develop and extend my findings on how a mediation function is carried out and how it unfolds, with a focus on primary roles and activities. The approach I followed, and its rationale, are described in Chapter 4.

Step 1. Roles

For the first step of the integration analysis, I began with the three primary roles identified through inductive coding of the case data: technical expert, capacity builder and guide (see Figure 6.1). I placed the 90 statements from the concept map in random order, and then sorted them by these three role categories. As I sorted, I made reference to the types of activities associated with each as shown in Figure 6.1.

Most of the statements (80) fit one of the three categories. The remaining ten statements did not. After review of these 10, I grouped 6 into a new category called “Learner”; this new category is described further below. I coded the remaining 4 as “cross-cutting”.

Overall, 42% of the GCM statements described activities related to the technical role as defined in the cross-case results. This makes the technical expert role the largest by number of statements (see Figure 8.1, below) – not surprising considering this is usually an evaluator's formal and most explicit role. Guide is next (32% of statements), followed by capacity builder

(14%). The new role, learner, is small by comparison (1%). However, of note, this category's statements have the highest average rating by GCM participants for frequency of priority (4.24 on a scale of 5)²⁶. Figure 8.1 depicts relative size of the role categories with number of statements and average rating in brackets.

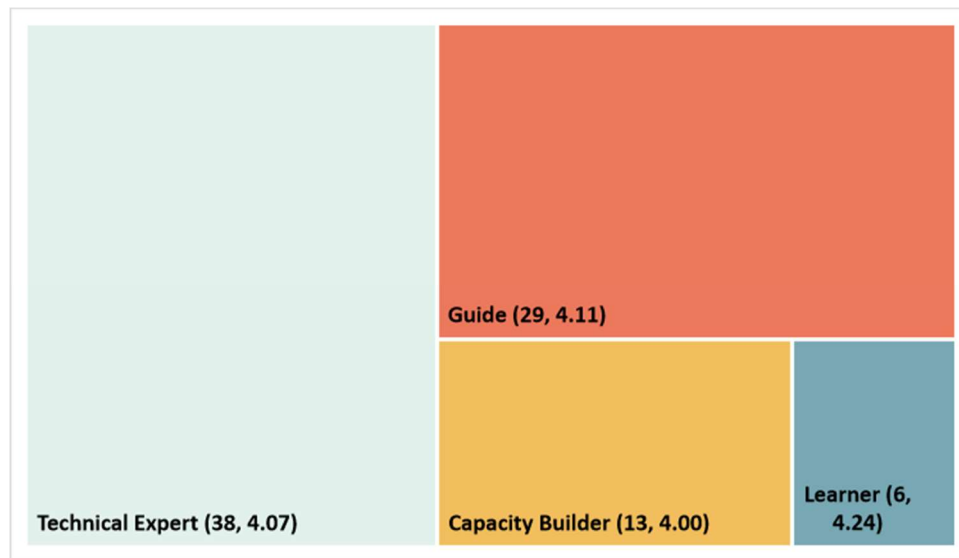


Figure 8.1. Size and average rating of cross-case role categories based on GCM statement coding.

I then sub-coded the statements in the two largest categories by theme. The statements in the two smallest categories, capacity builder and learner, did not easily suggest subthemes and were left whole. Next, I considered the content of each category with reference to the framework of mediator strategies (Figure 6.2), and lastly for how the statements in each category are distributed on the concept map. I outline these results next.

Technical expert. Overall, the statements in this category have a facilitative and nondirective tone. They describe “facilitating”, “enabling”, and “supporting” the work of other actors through the application of specialized skills and by ensuring quality of data and analysis. This is consistent with the findings from the cross-case analysis that the role’s purpose is to support others’ decision making, and that this role sits at the “nondirective” end of the framework of mediation strategies. It is directive only in terms of shaping the evaluation.

²⁶ The sorting and rating participants in the GCM (Strand 2) were asked to rate each statement following a 5-point Likert type scale in response to the question “In your experience, how often is this a priority?”.

Within the category, the GCM statements reflect two primary themes. The first theme is about “opening up” the frame of inquiry and includes statements about helping actors “see the unexpected”, “be open minded”, “spot new patterns”, and to bring out actors’ creativity and insights. The second, larger, theme includes statements about applying frameworks and using tools, rigour and design. I called this second theme “grounding”.

As a next step, I considered the location of the two themes’ statements on the concept map. I found their distribution lends weight to grouping into these two categories. The statements in the “opening up” theme are located in Clusters 3, 4 and 7, which are neighbours in the map. The “grounding” theme’s statements are located across clusters except 3, 4 and 7²⁷ (Figure 8.2). Overall, statements relating to the technical role span all clusters of the map and yet appear to have two distinct facets. These two facets will be revisited in the next chapter on implications for adaptive learning.

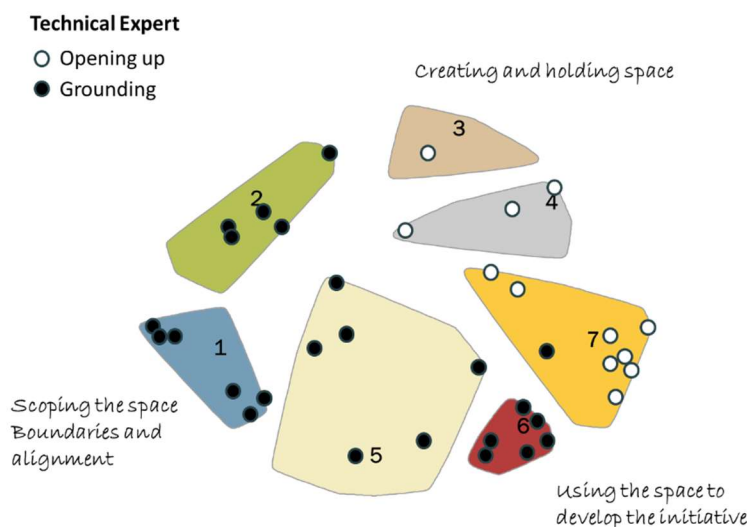


Figure 8.2. Distribution of statements relating to subthemes of the technical role.

Guide. The statements in the guide category imply forward movement and sometimes link to the initiative itself in addition to the evaluation. Overall, they describe efforts to create momentum, to stay on course or course-correct, and to clarify, “nudge” and “evolve” actors’

²⁷ One of the 24 statements under this theme is located in Cluster 7 “Provide a systematic framework for collective learning”.

thinking. This accords with a position closer to the directive end of the mediation strategy continuum in which recommendations for action are given and forms of leverage may be used to move the process forward.

The guide statements also appeared to group into two themes. The first theme concentrates on momentum and provoking actors' thinking. They describe “using evaluation as a vehicle”, “nudging and challenging stakeholders thinking”, and “pushing for new insights”. The second theme is about staying on track, for example “helping to course correct” and “framing up models of the intervention”. Unlike the technical role above, the statements in the guide category do not distribute over the entire map; they concentrate on the right side, away from Clusters 1 and 2²⁸. The two themes in the guide category overlap in the map, although the “staying on track” theme is concentrated in the lower clusters, as illustrated in Figure 8.3.

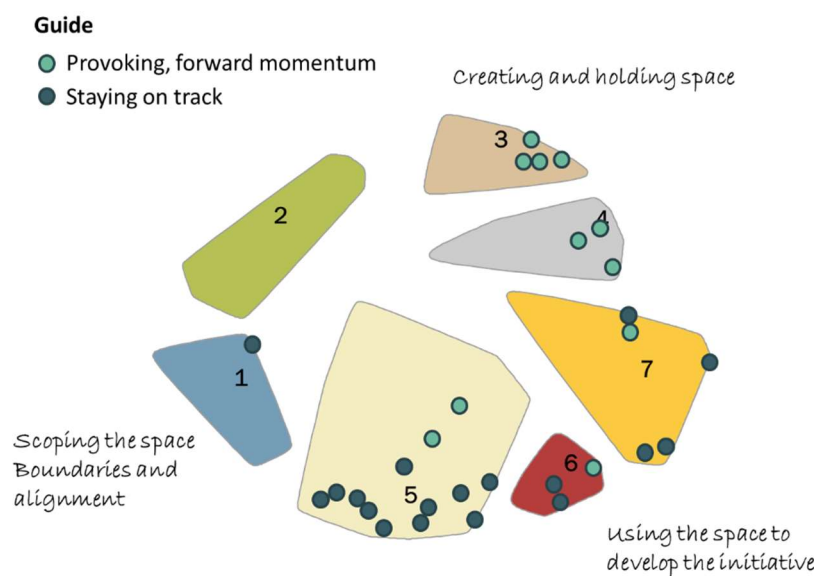


Figure 8.3. Distribution of statements by subtheme in the guide role.

The near-complete omission of Clusters 1 and 2 suggests that the activities in the guide category are embedded in “the inquiry space” (i.e., the rest of the map as described by participants in Strand 2). This is consistent with most GCM statements in this category; they are activities that would take place during the active part of the process rather than the planning of it, such as

²⁸ With the exception of one statement in Cluster 1, “Be a good strategy coach”

“Point out where decisions are being made without (or despite) considering key data.” The guide role may also be a less conventional understanding of an evaluator’s function compared with a technical or capacity building role. For this reason, it may be less formalized and feature less in the explicit scoping and positioning activities of Clusters 1 and 2. Yet, accounts of self-imposed boundaries are reported in the cross-case findings with respect to this role (see Chapter 6) and echoed in a few of the GCM statements in this category, for example, “know when not to intervene too much with their ideas, and when to intervene because people are getting stuck”. The bounding of the role suggests there should be a place for it in Cluster 2 where the evaluator’s role is defined. I will take up the importance of defining role boundaries again in the next chapter under implications for practice.

Capacity Builder. This category is much smaller than the ones above. If quantity of statements is an indicator of importance, this finding would conflict with the cross-case findings. The cross-case analysis found that capacity builder was the role most emphasized in the case data and that it was associated with several “breakthrough” moments in the cases. The difference may relate the multi-case study’s emphasis on turning points and trajectory during data collection and analysis (see Chapter 4), which foregrounded a capacity builder role. The concept mapping methods, in contrast, did not focus on turning points but on important activities in general.

The statements in the capacity builder category do correspond to the cross-case findings in that they suggest a location in the middle of the mediation strategy continuum, linked to “creating conditions for a joint solution”. The statements emphasize relationship development and the development of evaluative thinking, for example “shifting how people connect with one another”, “mediating power dynamics”, “building a culture of curiosity and inquiry” and “helping actors make evaluation part of their work”. More than a third of the statements are from Cluster 3 (5 of 13 statements), although there are statements from every cluster in this category. This role serves the development of collaborative capacity and has central importance from a process perspective. Potential implications for adaptive learning are discussed in the next chapter.

Learner. This category was created for a small subset of 6 statements that did not easily fit the pre-existing categories. The category includes statements such as “listen and pay attention

to what is going on”, “be really aware of power dynamics” and “have a deep understanding of the work by being integrated in the work”. I did not identify this as a distinct role during cross-case analysis. However, activities like these can be found in the case data (see Chapter 5). In Case 2, for example, accounts describe the evaluation team monitoring local media and participating in local community events to increase their sensitivity to context. A chapter 3 participant describes the importance of understanding the perspectives of actors within organizations in order to be able to adapt to “an organization’s culture and how it thinks and what it likes” (c32). An awareness of diverse needs and stakeholders is evident in the way that the evaluators successfully adapted communications, tools and methods to fit the context in all cases (see Table 6.2, cross-case, strategies for making space). The learner role would be located with capacity builder in the middle of the mediation strategy continuum. They both represent efforts aimed at developing underlying conditions or capacity to advance the process. However, this role is distinct in that, rather than focusing on program actors, it is meant to advance the evaluators’ own learning or sensitivity to the context by being actively present and by collaborating and co-creating with others. The statements are located in the upper half of the map, roughly opposite the concentration in the guide category, as shown in Figure 8.4.

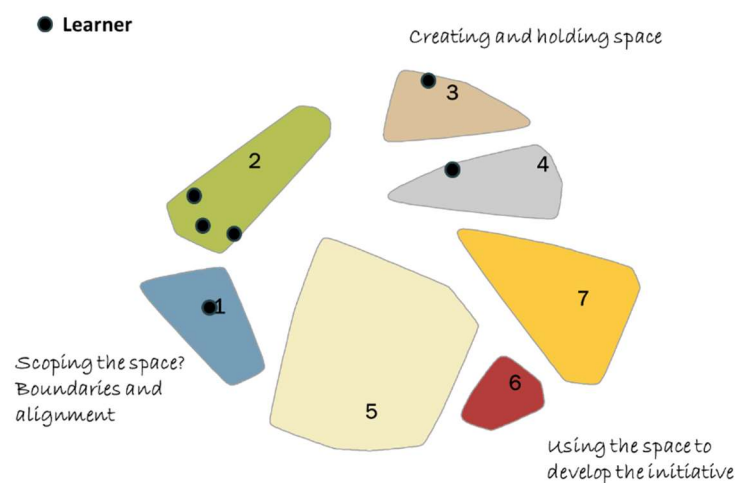


Figure 8.4. Distribution of statements in the learner role.

In Table 8.1, below, I reconfigure the information on roles from Chapter 6 and include the refinements based on this analysis.

Table 8.1. Primary role categories

Role	Areas of activity
Technical expert	- Supporting rigour and methodological quality - Providing a 'rich' toolbox for experimentation and tailoring of methods - Clear communication and advocacy for the data at the table
Guide	- Focusing attention (eye on impact, reframing the intervention) - Clarifying ideas key to implementation of the intervention (surfacing assumptions, articulating fuzzy constructs) - Forward momentum, guidance (nudging thinking, traction, leverage)
Capacity builder	- ECB (skills, concepts) - Attitudes, orientation to evaluation (building a culture of inquiry, pacing, reframing inquiry) - Creating conditions for collaborative inquiry (structures, modelling, cooling, facilitating)
Learner*	- Being actively present, listening - Deep understanding through integration in the intervention - Creating conditions for collaborative inquiry (context sensitivity)

*added at integration step.

In Figure 8.5, below, I adapt the mediation strategy framework from Chapter 6 to illustrate the relative positioning of the four role categories in terms of “directiveness” and also domain (the inquiry or evaluation and the intervention itself), as suggested by this study’s findings.

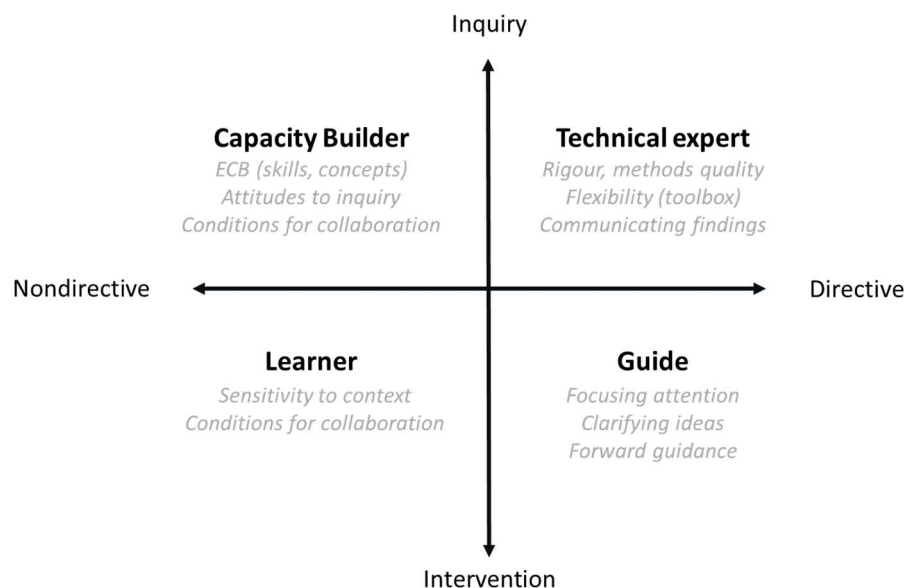


Figure 8.5. Location of four roles in terms of directiveness and domain focus.

It should be noted that although the guide role is positioned as “directive” in terms of the intervention, this does not mean that the evaluators are making program decisions. Rather, it uses the term “directive” in the sense used in the original framework from the mediation literature, and as noted above in Table 8.1.

Also noted above, the technical role is most prominent in terms of number of activities – this is the role that is front of mind when the structure of the practice is considered through the GCM – and the capacity builder role takes precedence when the process of the practice is emphasized. Notably, these sit beside each other in Figure 8.5, in the upper half of the grid focused on the inquiry. The foregrounding of these two roles may signal an overall emphasis in focus by the evaluator on the inquiry relative to the intervention.

The mediation literature suggests that a mediators’ strategic range may vary and is influenced by multiple factors such as actual and perceived expertise, status and trust (see Chapter 6). The cross-case findings support the idea that flexibility to act across the quadrants in Figure 8.5 may be mediated by similar factors. In terms of expertise, the range of skills required in DE has prompted suggestions that they may surpass the capabilities of a single evaluator, and DE may be best carried out in teams (see e.g., Cabaj, 2011). The case data describe evaluators extending their own strategic capacity by harnessing support of other actors. This included contracted support for activities in the upper half of Figure 8.5 (inquiry), for example, contract evaluators and professional facilitators. It included drawing on support of champions in the lower half of Figure 8.5 (intervention) (e.g., to further their own sensitivity to context and lend weight to recommendations). It is also important to note that one of the cases (Case 2), involved a team of evaluators and in two other cases (1, 4) the evaluation consultant worked closely with internal program staff with substantial evaluation training and capacity. It would seem to be difficult to achieve this range or to accomplish what is included in Table 8.1 (above) as a single evaluator working alone. The implications of this for a practice of AL mediation is discussed in the next chapter.

In summary, the findings identify four primary roles, each with a discernable function. Each role spans multiple clusters of the concept map, indicating that each is multifaceted. The findings are consistent with observations by Patton (2007) and Greene (Greene, 2000a, 2000b) among others, that evaluators perform multiple roles and that their activities and functions are

dynamic over time and context. In the next step of the integration analysis, I revisit my first research question by using the concept map's 3 domains of activities as the starting point.

Step 2. Domains of Activities

In this step, I continued with my first research question by responding to its sub-questions. These are intended to shed more light on the role of mediation by exploring important activities and how they manifest in practice:

- What are the most important activities or points of intervention for a mediator? Why?
- How is mediation enacted (e.g., activities, tools, timing)? How does it unfold? Why?

For this step, I used findings from the concept mapping component as an organizing framework. I used the three domains of activity that were identified in the map as a framework to code the findings on the cross-case themes: “Making Space” and “Learning Forward” (see Chapter 6).

The three domains. The GCM findings suggest that the most important activities for a developmental evaluator fall in three domains that correspond to regions of the map (Table 8.2).

Table 8.2. Domains of activity.

Activity Domain	Description	Clusters
1. Scoping, boundaries and alignment	Ensuring fit between DE and the intervention; defining relationships between evaluator and program development team	1, 2
2. Creating and holding space	Developing ways of working within the development team that enable adaptive inquiry; questioning, eliciting insight, exploring, framing, addressing tensions.	3, 4
3. Using the space to develop the initiative	Enabling collective reflection, analysis and learning; exploring pathways, grounding decisions in data.	5, 6, 7

The cross-case theme “Making Space” coded across all three of these domains, indicating that this theme cuts across the whole of DE practice as mapped in the GCM. The largest proportion of references fell in the domain “Scoping, Boundaries and Alignment”, which foregrounds the importance of boundaries and alignment to the findings in “Making Space”. The

cross-case theme “Learning Forward” coded primarily under Domain 3: “Using the Space”: this theme and this domain align closely. Coding results are illustrated in Figure 8.6.

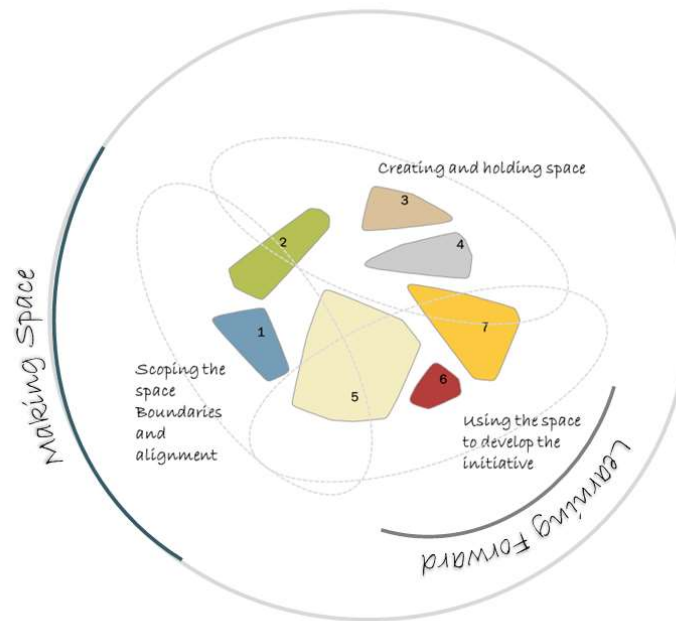


Figure 8.6. Overview of NVivo coding against the GCM framework.

When I analyzed the results in greater depth, I found strong correspondence between the GCM activity domains and the cross-case findings, and the cross-case findings helped to further contextualize and extend the GCM. I also found some points of divergence. These results are reported below.

Domain 1. Scoping, boundaries and alignment. As noted in the GCM findings (Chapter 7), the activities of Domain 1 help to define the inquiry process both at the beginning and as it unfolds. Cluster 1’s activities centre on fitting the *evaluation* with the intervention’s context. Cluster 2’s activities focus on the *evaluator*, and the evaluator’s relationship to the program development team. The cross-case findings under this domain help to develop the concept map by expanding on aspects of a process of alignment and raising questions about navigating boundaries.

Four components of alignment. The GCM findings identify two important components of aligning the evaluation to the intervention: timing and focus (Table 8.3). Timing includes assessment up-front and also decisions about design and maintenance of fit over the course of the evaluation. The other component concerns the focus of alignment: aligning the evaluation to the nature of the initiative, the purpose of the evaluation and the needs and interests of actors.

The cross-case findings suggest two additional components to consider: what aspects of the evaluation are being tailored for alignment, and the purposes for alignment (why). The four components are interrelated, and together they form a tentative framework for a fuller understanding of alignment. Table 8.3 provides a summary of the four components. I elaborate briefly on each of the four below, beginning with the two identified in the GCM (timing and focus) followed by the two identified through the integration analysis (aligning what and why).

Table 8.3. Components of alignment.

Categories	Elements
When? (Timing)	• Up front assessment • Iteration over time
Aligning to? (Focus)	Characteristics of the initiative and its context • Initiative structure and processes (e.g. decision-making processes) • Stakeholders: their purposes, goals, cultures, capacity (expressed, emergent, diverse) • Scope and boundaries of the evaluation (implicit, explicit) • Discovered or emergent issues in context
Aligning what?	Evaluation approach, design, methods, activities, including: • Questions • Framing, language • Pace, intensity of activities, pressure • Feedback (quantity, timing, priorities, format) The evaluator's role
Why?	• Compatibility, appropriateness (methodological quality) • Relevance, salience (engagement, use) • Allowing the inquiry to go deeper

When? GCM participants described up-front assessment of fit as essential foundational work for a DE. This foundational work informs a go/no go decision on the use of DE and choices for initial evaluation design. They also noted that fit may need to be revisited; actors may need to

be able to adjust the design or the parameters of the evaluator's role over time. The cross-case findings correspond well. They provide examples of up-front assessment (e.g., via extensive interviewing of program actors before making design decisions in Case 2 and 4) and iterative re-alignment. However, the cross-case findings also strongly suggest that up-front assessment is challenging and likely to be imperfect (see below).

Aligning to what? The GCM results refer to ensuring that DE fits the evaluation's purpose (i.e., there is a developmental purpose) and readiness of stakeholders to engage in an approach like DE. The cross-case findings agree. Similar to the GCM, the case data describe attention to the overall aims for the evaluation, the initiative characteristics (e.g., its maturity, complexity) and stakeholder needs and characteristics to inform design. The data especially reinforce the importance of alignment to stakeholder needs and characteristics. At least two teams were surprised by setbacks related to misalignment as the evaluations unfolded; these were attributed to needs, characteristics and expectations of actors in the collaboration, not misalignment with assessment of the initiative or the purpose of the evaluation (e.g., in Case 4, methods lacked sufficient credibility with some stakeholders). The study findings align with assessing readiness for DE in three steps, as suggested by Cabaj (2014), who provides a checklist for readiness assessment in DE. This checklist breaks readiness into (1) Situation (does the intervention need development and an adaptive approach?), (2) Adaptive Capacity (does the group have capacity to work in an adaptive way?) and (3) Readiness for Learning and Evaluation (are they "ready to embrace evaluative feedback and data" to inform decisions?) (p. 1). DE is a good fit when the group scores high on all three of these categories. When scores are low in the latter two categories, Cabaj advises actors to address capacity limitations before beginning the DE or else to "intentionally approach the work moving forward as an opportunity to strengthen your capacity for developmental evaluation" (p. 4).

The case data suggest that up-front assessment of stakeholder characteristics may be difficult for at least three reasons, however:

1. program actors may be enthusiastic to try DE, but this may not correlate with 'readiness' because their understanding of DE may be limited
2. there can be a diversity of needs and interests, and DE may align better for some stakeholders than others, and

3. stakeholder needs may only surface after the DE is well underway.

Up-front assessment is also likely to be challenging to conduct and imperfect, although it may provide useful information and help to raise stakeholders' awareness of the need for capacity.

A complementary starting point may be to consider whether features are in place that can support longer-term stakeholder engagement with the DE while mutual understanding and capacity develop. The cross-case findings point to features such as structural interdependency, which appeared to help keep actors at the table in spite of differences among them; the active participation of representatives from the funder, which helped support engagement in the process; funder commitment to principles of collaboration; and availability of champions, who helped out at times when teams got 'stuck'. It may also help to consider activities that deepen an evaluator's understanding of stakeholder needs and context, for instance those echoed in the "learner" role, above.

Part of the work of Clusters 1 and 2 is to ensure that stakeholders have appropriate expectations for DE, especially if actors have experience with other forms of evaluation, because DE "looks, feels and is implemented differently" (CMP-K). The cross-case findings highlight that participation in the inquiry process (i.e., through the activities in Clusters 3-7) may be key to developing such expectations because participation advances actors' understanding, capacity for and valuing of the approach. This suggests a two-way process of alignment over time. Establishing the "fit" of DE to the initiative may also (in part) involve developing actor alignment with DE. "Establishing fit" is a process of "fitting" that takes place over the life of the DE and may be bi-directional.

One element in Table 8.3 (above) is a point of divergence between the cross-case and the GCM findings. This relates to negotiating boundaries set for the evaluation. The case data describe the evaluation teams adjusting activities to suit predetermined constraints on the evaluation, but this is not identified in the GCM. For example, the evaluation in Case 3 was bounded to exclude other funding partners and quantitative methods. Boundary setting on an evaluation affects the process in multiple ways evident in the cross-case findings, including questions asked, what activities were undertaken and how, and the delivery of feedback. It may also influence how the evaluation is used or not used (see 'non-use' outcomes, Chapter 6). While

the evaluator may have little control over where these boundaries are drawn, especially early in a process, evaluators may be able to inform actors' awareness of implicit boundaries, the effects of boundaries, the intentionality and transparency of boundary-setting, and the negotiation of boundaries over time.

Aligning what? The cross-case findings suggest two additional components to alignment that were not raised explicitly in the GCM. These are included in Table 8.3, above, under headers "Aligning What" and "Why". In the case data, the aspects of the evaluation ('what') that were being tailored ranged from specific activities (e.g., timing of feedback) to the overall approach, for example by hybridizing to incorporate formative and summative elements (Case 1, 2, 4) and research (Case 1, 3). The evaluator's role is also subject to alignment, and as role is of central interest to this study, this warrants brief elaboration.

Cluster 2 includes activities related to aligning the evaluator role with stakeholder expectations. The importance of establishing clear parameters for the role and managing expectations early in the process was noted by participants. A least three challenges exist for the evaluator in establishing and then meeting expectations for their role:

1. expectations are linked to stakeholder understanding of DE, which may be incomplete at the beginning (see above)
2. it was noted that the evaluators may hold a 'dual' relational commitment as a member of the development team and also with responsibility to report out to non-participants, such as a funder, making their role more difficult to navigate, and
3. developmental evaluators appear to be performing multiple roles, some of which are more difficult to articulate than others

The evaluator's relationship with the development team is the first of three relationship foci denoted in Figure 6.3. The cross-case findings illustrate how confusion about the evaluator's role can result in tension and how low levels of trust or credibility can hinder or stall the evaluation process.

Here, the case studies echo GCM participants' caution about navigating dual relational commitments. A GCM participant described challenges related to the evaluator working with program implementers and also reporting outside the team to a remote funder. In the cases,

activities are described that reflect tensions due to external reporting (i.e., outside the development team) as part of the evaluator's function, even though in these cases representatives of the funder were at the table. In the cases, the 'duality' extended to various groups, for example boards of directors, donors and other funding partners, and/or public opinion via the media. For example, in Case 2, a process was established to ensure stakeholders had opportunity to review, discuss and understand findings before they were released to the public. The relationship between the development team with more distant stakeholders is the third of three relationship foci denoted in Figure 6.3,

Why? The case data also include at least three reported benefits ('why') of fitting the evaluation. One of these is increased methodological quality of the inquiry because methodological choices should align with the nature of the object of study, for example the initiative's complexity or rate of change. Another is that fitting or tailoring increases relevance or salience of the evaluation to stakeholder interests and needs, which in turn fosters engagement with the process. The third is that fitting allows the evaluation to probe more deeply over time as the team better understands the nuances of the context and as unexpected questions surface.

Overall, these findings shed light on ways that the activities of Domain 1 function as integrated components of the overall process. I will return to this in the next chapter on implications for AL.

Domain 2. Creating and holding space. Domain 2 is about opening up the inquiry, "in a space separate from focus on decisions or closure" (Statement 15). Participants interpreting the GCM described this as the site of activities to support an enabling environment, that is, one that enables inquiry and productive ways of working. Activities in Cluster 4 include questioning, exploring ambiguity and helping actors to see their work differently. The activities here are likely to surface differences among stakeholders (e.g., differing theories of change). Cluster 3's activities include attending to group dynamics and facilitating working relationships to overcome tensions, for example mediating power dynamics and shifting how people connect with one another. GCM participants described how Cluster 4's activities drive the need for those of Cluster 3, and how the activities of Cluster 3 may help to sustain and deepen the work of Cluster 4. Cluster 3 has the lowest average bridging value in the map, indicating the statements in this cluster were sorted together with highest frequency and rarely with statements outside the

cluster. This suggests it is the most internally cohesive and possibly a specialized cluster of activities.

As with Domain 1, the case findings served to extend and elaborate the GCM. For example, the case data describe a need to “carve out space” for productive engagement in the DE (see Chapter 6). In all the cases, the evaluation brought diverse needs and understandings to light, sometimes intentionally (as in Case 4 when leadership first attempted to co-construct a TOC), and sometimes not (as in Case 1 when some staff struggled to adapt to ongoing development as an approach while others embraced it). The case findings also provide examples of strategies related to the opening, exploring and (re)framing purpose of Cluster 4 and how these strategies were used, for example, with TOCs to surface assumptions. Conceptual use as outcomes of the evaluations is reported in all four cases, for example rethinking program approaches and/or shifting culture around learning and evaluation as a group. The cross-case findings also include activities and strategies related to Cluster 3, for example, cooling, pacing, and using objects to mediate interactions.

Participants in the GCM panel identified a gap during their discussion of this domain: they asserted the need for guiding principles to help navigate this aspect of the work, particularly in “times of tension or times of confusion or times of pressure” (CMP-N). They recommended that a program development team establish a common set of principles as part of creating and holding space for inquiry, that is, “what are the principles that we agree on?” (CMP-M). Although the GCM statements do not include activities related to establishing principles, arguably some statements imply the existence and use of principles, especially principles relating to collaboration. To test this, I compared a subset of the GCM statements with the recently published Principles for Collaborative Approaches to Evaluation (CAE) (Shulha, Whitmore, Cousins, Gilbert, & al Hudib, 2016). Table 8.4 lists examples for illustrative purposes.

Table 8.4. Examples of alignment of GCM items with CAE principles

CAE Principle	Example GCM Statements
Foster meaningful relationships	• Develop good trusting relationships with people around the table • Engage with collaborators in multiple ways.
Promote appropriate participatory relationships	• Create the right environment for different groups (e.g. members of the community) to contribute to the evaluation. • Facilitate participation in analysis of data
Promote evaluative thinking	• Detail the initiative's evolving theory of change for critical review • Help to build a culture of curiosity and inquiry about their work • Develop evaluative thinking.

Principles also feature vividly in the case findings. In particular, co-creation, joint ownership and reciprocity were expressed commitments in common across the cases. The application of these principles to help guide the evaluations is evident. For example, co-creation was adopted and applied as a “mantra” (Case 2, c21, 2d0, v12); integrating evaluation into everyday work of the intervention also as a “mantra” (Case 1, c11); using the principle of inclusion to inform collaborative decision making (Case 1); ensuring shared access to data (Case 1, 4); and giving and raising voice (Case 2). The GCM participants suggested identifying and naming common principles, and the case data describe how principles were actively modeled. Overall the cross-case findings provide support for the GCM participants’ view that guiding principles are important to the activities of a developmental evaluator, how the role is performed and how the process unfolds²⁹. In the next chapter, I discuss the implications of this finding for a definition of adaptive learning.

Domain 3. Using the space to develop the initiative. Clusters 5, 6 and 7 work in concert to help “ground decisions in data” as described by GCM participants (CMP-S). This is where the “hard stuff” of data collection and analysis take place (CMP-W), and where learning through the evaluation is pulled together to be applied in decision making. The GCM findings identify five categories of activities in this domain. These are provided in Table 8.5 in relation to findings from the cross-case on how they were expressed in the cases.

²⁹ In addition to the use of principles to guide the specific DEs, the development of principles for the implementing organization (e.g., guiding principles for the foundation in Case 3 and operational principles in Case 4) were reported as outcomes of these DEs.

Table 8.5. GCM activities and cross-case extensions for Domain 3

GCM	Cross-Case (how)
Data collection and feedback with use of evaluation tools	• Multiple data collection methods and sources (all cases) ○ Efforts to ensure input from relevant groups (e.g., front-line workers Case 1; community members Case 3). • Integration of data collection in the day-to-day of the intervention: ○ Tailoring, pacing to limit burden or for synergy with structures (all); design of instruments to fit context, e.g., language, formats (1, 3, 4). • “Data overload” in two cases (2, 4); Adjusting delivery to avoid pushback (format, pace, planting seeds) (Cases 1, 2, 4).
Collective analysis, reflection and learning; highlighting patterns; pulling together ‘what have we learned’?	• Facilitation to improve depth, quality of collective analysis (e.g., using runways) • Establishing structures (e.g., regular meetings, analysis frameworks) to protect time and space • Contributes to cooling, engagement, capacity (Domain 2) via practice, transparency of process (Cases 1, 2, 4)
Exploring pathways	Exploring both foreseen and unforeseen • Testing preconceived ideas by collecting and analyzing data against a TOC (e.g., Case 1, Case 4) • Attention to emerging events or evidence of ‘theory failure’ (e.g., systems change pathway Case 1; collaboration issues Case 3) • Identifying and tracing alternate pathways (e.g., evidence of laterally-driven systems change in Case 1)
Grounding decisions in data, course correcting	Efforts to ground decisions in: • Data. Emphasis on use of data for ongoing navigation (Cases 1, 2, 4). Attention to course correction, e.g., “using data to say, are we on the right track?” (c12). Explicit attention to using the data to inform future strategy (all cases) With reference to: • Goals/principles: helping actors navigate with reference to goals and/or principles (all cases). • Theory of change. TOC clarity reported as an outcome; used to guide decision-making (1, 4). “watching for deviations, reminding of the rationale, questioning shifts, pulling people back” (c11) Aided by: • Use of pause. Using structured processes to encourage the team to pause and make sense of what was happening, “avoid running headlong into implementation” (c12). • Gaining traction. Using anchors (e.g., monitoring tools, Case 4), making progress visible (Case 1, 4); help from champions. • Support for resourcing. Evaluations credited with helping to secure support from donors and other funders for course corrections and changes to the initiative (e.g. Case 1, 2, 4). Overall - instrumental use outcomes reported in all cases.
Honing strategy / framing up models	• Clarifying ideas core to implementation of the initiatives • Solutioning / testing solutions: support to ideas generation and testing proposed solutions, e.g., new collaboration approach (C3); solutions to underrepresentation of front-line staff (C1). • Documenting. decisions made, actions taken, iteration to TOC (1, 2, 4); helping actors see learning and progress, avoid repetition, move forward.

The correspondence between the GCM and the cross-case findings is evident in Table 8.5. The table also highlights a few extensions to the GCM content suggested by the cross-case. For example, the GCM findings point to the importance of using data to help actors to stay on track or to course correct. The cross-case findings elaborate on this by exemplifying how this can be put in practice, for example by linking data to stated goals, principles or a developing TOC, aided by processes to insert pause or gain traction. Overall, Table 8.5 serves to illustrate how case teams employed diverse aspects of their role(s) to use the inquiry to serve the developmental purpose: to bridge a wide gap between the provision of data and data-informed program development.

Interdependencies among the domains. GCM participants described how success in Domain 3 can depend on activities in Domain 2 (see e.g., the relationship described between Cluster 5 and Cluster 4 in Chapter 7). This is because Domain 3 requires an openness to iterative inquiry, experimentation and feedback, and “how to even think about that” (CMP-K), which can be a challenge for some organizations. This ‘opening up’ and establishing new ways of working is part of Domain 2. The cross-case findings further reinforce this idea of a need to build capacity for the activities of Domain 3.

The cross-case findings also suggest feedback in the reverse direction, that is, that Domain 3 feeds into Domain 2. For example, experiences in collaborative analysis and interpretation, and seeing concrete decisions and positive developments to their programs (Domain 3) were reported to help build trust and aid with cooling efforts (Domain 2) because they increased transparency of the evaluation, enabled shared access to data and/or developed a sense of ownership of the process.

Interdependency among the domains of activity is supported by both strands of the study. Although there is a temporal aspect to the clusters in the map (see Chapter 7), their interdependency reminds us not to assume the DE process unfolds as a series of discrete phases over time, that is, first Domain 1, then Domain 2, followed by Domain 3. This corresponds to observations of others who describe activities for developmental evaluators which, although they can be placed logically in temporal sequence, are described as ongoing and interconnected in DE (Dozois et al., 2010). The interdependency of the CM’s domain lends support to the view of DE as a multifaceted and distributed practice.

Closing Notes. *Role fluidity and ambiguity.* Like a mediator, an evaluator has no explicit or direct authority with respect to the intervention. She relies on influence to go beyond providing information to enabling its use in practice. The cross-case findings suggest that the evaluators' ability to navigate a constellation of roles was likely critical to moving the DE work forward for use. Reinforcing this, the integration analysis indicates that performance of each role crosses multiple regions of the GCM; each role carries across multiple aspects of DE practice. Roles in the cases were sometimes fluid and ambiguous, which allowed them to be interpreted and labelled in different ways. Roles become even more ambiguous as functions are shared across actors and time and integrated into day-to-day program activities as illustrated by the case findings.

Lines may be blurred between functions within a single role, particularly between technical and relational functions. The identification of at least three relational targets for DE activities provides context to help explain the apparent shifting between 'insider' and 'outsider' positioning, as needs and situation present (see Chapter 6, and the next chapter). Moreover, different frames of analysis used in this study served to foreground different roles. Above, I noted that attention to turning points casts light on capacity building; while sheer number of discrete activities establishes the importance of the technical role.

The idea that the work of developmental evaluators can sometimes be 'invisible' was described by participants in both strands of this study. Creative and improvised combinations of roles likely contribute to this. As well, while concrete outputs are a feature of many of the technical role activities, and some of the guide activities, a sizeable number of the GCM statements do not point to visible 'products'. Case study participants attributed invisibility of DE to its emphasis on process over reporting, that is, that DE is about "the thinking that people do along the way. That's often very invisible" (c11), that the building of trust is not visible (c22, c23), that the way that DE "creates a container or space for regular reflection, for relationship building, for learning together" is invisible, never seen in reports (c12). Another participant said:

...the shifting, the slow gradual shifting of the... team to sort of really stop being a, moving from being really sort of fearful and apprehensive about this process to embracing it – that, again, was because of the process, that wasn't because of the findings... [That is] absolutely not captured in that binder full of findings. It's not there.” (c21).

These observations are discussed further in the next chapter.

The cross-case and GCM findings complement and reinforce each other, and sometimes challenge and extend each other. Together, the two strands of the study have brought forward insights in response to the first research question for this study: on the types of roles performed, the functions and activities of these roles, and various constraints and enablers that help define how they are performed. In the next chapter, I compare the study findings back to the core components of this study's conceptual framework and discuss implications for an adaptive learning process.

Chapter 9 Implications for Adaptive Learning

So far, responses to my second major research question about implications for adaptive learning (AL) have been implicit in the findings from my two strands. In this chapter, I respond directly to this last question by discussing implications for an AL process. In response to this question, I revisit the definition for AL used in this study in light of my findings, and then briefly review key components of my conceptual framework. I then propose a tentative model for thinking about AL as an interplay of divergent and convergent thinking supported by collaborative capacity. I close with a discussion of three important implications for practice.

Implications for an Adaptive Learning Process

Revisiting the definition of AL. I defined AL as a process which unfolds within a social learning system. Drawing on existing literature, I proposed that the process involves: (1) generating data of range and balance, (2) debate and critical interpretation, and (3) informed, viable action. These three components are clearly evident in my findings (see Table 8.5). However, the findings also suggest that the definition I used was deficient in three important ways: where the process begins; where it ends; and how it may be navigated.

Where the process begins. The study has identified three broad domains of activity (see Table 8.2). My working definition for AL was limited to only one of these: Domain 3. This is where the “hard stuff” of data collection, analysis and interpretation occur, and where they are ‘pulled together’ to support action. The other two domains are about (1) scoping, bounding and aligning the process and (2) creating and holding space. The study findings strongly suggest that the three domains are interdependent and that all three together are needed for a collaborating group to learn adaptively.

The tasks of Domain 2 support opening up the inquiry to new perspectives and ideas, for example, developing skills, norms, habits and trust in a collaboration that are supportive to exploring a diversity of ideas. Domain 2 is where ideas and assumptions are brought forward in earnest, potentially bringing conflict with it. The prospect of conflict to be generative (as well as obstructive) also emerged in these findings.

The findings also highlight the importance of decisions about the boundaries of the inquiry (Domain 1). Scholars have argued for attention to boundaries, particularly in an inquiry for which diversity of perspectives is important, including DE, for both ethical and practical

reasons (Patton, 2016a). Examples include attending to which perspectives are at the table, and which are not (but should be) (Greene, 2000b; Midgley, 2016). The findings from this study locate boundary decisions in Domain 1 but also throughout the process and indicate that they need to be revisited and negotiated over time as learning progresses via periodic return to Domain 1 (Chapter 8, Table 8.3, see also Midgley & Lindhult, 2017). In short, the first two domains are not precursors to AL or about getting ready for AL, but an integrated part of an AL process.

How the process ends. The second deficiency relates to outcomes of the process. The definition stated that an AL process results in informed, viable action. However, outcomes described in this study did not all involve actions. Although instrumental outcomes were identified, such as implementing program changes, other outcomes were equally important, such as shifting culture around learning and strengthening commitment to collaboration. Such developments arose largely through participation in the process. They accord with findings by others, such as Amo and Cousins (2007) who describe non-instrumental process use outcomes related to learning (e.g. ‘learning to learn’) and attitude/affect (e.g., confidence in ability to influence change). These other types of outcomes are important influencers of action, but the path is indirect (see e.g., Amo & Cousins, 2007; Cousins & Bourgeois, 2014; Cousins et al., 2014).

Even when findings point to an instrumental application, action may be delayed. This was illustrated in the case studies (e.g., Case 4 in which recommendations were made but action had to wait for opportunity). It aligns with other empirical research on interventions that involve systems change, describing periods of stagnation followed by breakthrough when levels in a system align (Berta et al., 2014), and research on how organizational adaptations tend to occur in a “lumpy” way over time (Tyre & Orlikowski, 1994, p. 115).

A tendency to focus too much on products or outcomes in the definitions of processes is critiqued in cognate areas of research (e.g. on innovation, creativity and KMb) for at least two reasons. First, because it can lead to too much emphasis on producing actionable results and securing their immediate implementation, which can result in unanticipated negative consequences (Van de Ven & Johnson, 2006). This is because we tend to assume that the products of systematic inquiry or innovation must be positive or beneficial, but they may not be (see e.g., Dahler-Larsen, 2016). Second, we might overlook or ignore ideas that, in the end, did

not warrant action, but can be important *inputs to* what is essentially a long-term learning process (Drazin, Glynn, & Kazanjian, 1999). Drazin, Glynn and Kazanjian (1999) reject defining a creative process by specific outcomes (as do others, see e.g., Smith & Smith, 2017). Instead they define creativity as “a process of engagement in creative acts, regardless of whether the resultant outcomes are novel, useful or creative” (p. 287). These authors assert that the process is *a necessary but not sufficient condition for desired outcomes*³⁰.

In the context of adaptation by organizations, Pelling et al. (2008) define learning as “a transformation in the *potential* for behaviour” (p. 9, citing Ison et al., 2000, italics original) and they link adaptive learning to changes to agency that in turn draw on changes in cognition, affect and relationships, all of which are intangible outcomes. This accords with B. Davis and Sumara (2006)’s definition of learning as a process that involves “becoming capable of more flexible, more creative activity that enables the unity to maintain its fit to ever evolving context” (p 92). Here, the emphasis is on a process of developing adaptive capacity (see also Muro & Jeffrey, 2008). I return to the importance of both actionable and nonactionable outcomes to an AL process below (Figure 9.1).

Principles for navigation. Third, the study’s findings highlighted challenges for navigating the process, and the use of data, goals and principles as signposts and to inform decision-making (see Table 8.5). In particular, the importance of principles for the navigation of an AL process is missing from my original definition. This study’s findings suggest that principles can help members of a collaborating group find their way through a complex process of learning and development, particularly in times of uncertainty, conflict and stress. The cross-case findings point to the importance of principles relating to collaboration (e.g., co-creation, reciprocity), the role of the funding organization to support these, and the role of the evaluators to operationalize and demonstrate them (see Chapter 6). The findings do not speak to how principles might first be negotiated and established with a development team, although findings suggest that the Collaborative Principles for Evaluation are likely to be helpful (Shulha et al., 2016) (see Table 8.4). Others have described explicit processes for developing and agreeing on

³⁰ Smith and Smith (2017) offer the example of Leonardo da Vinci’s plan for a helicopter. The plan was not viable and not actioned by da Vinci. Does this mean that the process was not creative? Or maybe only counts as creative retroactively?

shared principles for collaborative work in social interventions (e.g., Murphy, 2016; Patton, 2018).

In sum, my definition of adaptive learning was too narrow for these findings. I propose a revised version: in a collective sense, adaptive learning is a process through which a group increases its capacity to make effective decisions under ambiguous or changing conditions. It includes enhancing its ability to seek out diverse and balanced data, to critically interpret the significance of these data, and to use the results to advance understanding. The process involves actively developing ways of working that strengthen the collaborative process and align it over time to the needs of the context. The process is a necessary but not sufficient condition for viable action. AL both relies on and further develops adaptive capacity, defined as: “an ability to experiment and foster innovative solutions in complex social and ecological circumstances” (Armitage, 2005, p. 703).

Revisiting the conceptual framework. At the start of this study, I also drew from sensemaking theory for a more detailed conceptual framework. Sensemaking theory offers flesh for the bones of the AL process, that is, conditions and sub-processes that influence what sense is made and by whom and how this influences trajectory. In the original framework, I noted transition points in collective sensemaking. Existing research suggests these are important but not well understood. The framework also located the sensemaking process in a multi-layered social learning system. I touch very briefly on these elements in the context of my findings.

Layers of the learning system. The sensemaking cycle in my framework (Figure 3.3) is embedded in an active context. The mediator is a member of the active context along with the program development team, their tools and their tasks. The active context is embedded in two other more distant contextual layers, a latent context (organizational cultures and structures surrounding the team), and the environmental context (funding policies, distant stakeholders, e.g. the public).

The findings clearly reinforce the view that surrounding, and/or intersecting systems (the latent and environmental contexts) are important influencers of learning by those in the active context (Argote & Miron-Spektor, 2011). In addition, this study describes how actors responded to alter, mitigate and/or leverage the influence of these other systems. For example, mediators

can help to effect changes in structures in the “latent” context to facilitate learning (e.g., efforts in Case 4, “to structure things to allow change to happen” (c41)). They can help to establish buffers against influences thought to inhibit AL, for example, by helping to establish norms or identity at team level that are more conducive to collaborative learning than those in the latent context (e.g., in Case 2, developing team identity, “that we’re about co-creating... and we’re about being adaptive” (c21)). Mediators and their teammates may attempt to alleviate pressure of other systems by negotiating communications processes or role parameters to support transparency and trust. These are discussed in the findings on responding to multiple relational targets (Figure 6.3) and “cooling” activities (Figure 6.1). Finally, the mediator may interact directly with stakeholders in the “environment”, either through direct reporting to a distant funder (see the GCM results), by facilitating and/or chairing advisory groups of stakeholders at multiple levels (see Case 1, 2 and 4) and/or by drawing on the assistance of outside champions (Case 1, 4). The above are examples of a mediator performing a bridging or brokering function across systems.

I adopted the active, latent and environmental labels in my framework from existing literature (Argote & Miron-Spektor, 2011) but the study’s findings reinforce that “latent” and “environmental” should not be taken to mean given, fixed, or out of scope of agents in the active context. This is consistent with observations by others that context is something that is dynamic and at least partly constituted and shaped by the agents involved (Chouinard & Milley, 2016).

It is also important to note that in addition to working across different systems, the findings suggest that mediators may work between formal and informal spaces within systems, for example, between both overt and tacit processes. Pelling et al. (2008) describe these as the official and the “shadow” spaces of systems around collaborating groups (p. 5, citing Stacey, 1996); official spaces are sanctioned and explicit, “shadow” spaces exist but may be purposefully unrecognized. These authors argue that the two need to be bridged to support adaptation by social groups. I revisit this in the next chapter under directions for future research.

Transitions and tensions in collective sensemaking. The study findings largely support the presence and the challenges of important transitions in a collective sensemaking process, and they describe ways mediation can enable these transitions. I will touch briefly on these and then discuss them in the context of AL.

Doubt and confidence. Sensemaking research suggests that effectively moving through the process requires both doubt and confidence: doubt by actors that they already know the optimal path forward, because this prompts inquiry, yet also a sense of confidence in their capacity to act (e.g., Weick, 2006 see Chapter 2). In their study on mediated sensemaking, Strike and Rerup (2016) highlight the critical importance of the mediator's role in creating not only doubt, but also confidence, which they call a “key puzzle” in sensemaking theories (p. 882).

Doubt was evident in the starting conditions of the case studies: at least some (though not all) actors in each case knew they were moving outside their usual way of working and were aware that they had no “manual” or “roadmap” for where they were going (c11; c12). Attention to maintaining a level of doubt over time is also evident, for example, in Case 1 “to hold our ideas about the initiative lightly” and “never assume that what we were doing was the right thing” (c12). Some level of doubt was maintained by the evaluators by explicitly challenging assumptions, helping actors think through the significance of their ideas for the intervention, and testing theories of change, for example. This type of critique and support to critical thinking is an assigned role for evaluators.

Alongside doubt, the study findings describe how the evaluators also supported confidence or a sense of capacity to act. This study describes how progress was documented in the DEs to:

- help actors see their “early wins” through the messiness of implementation (c22)
- make their learning visible, for example, through iterating theories of change which became “a source of pride” (c11), and
- raise confidence that the process can be effective, by producing concrete results that actors valued and which encouraged further engagement.

This forms a part of the strategies described under ‘gaining traction’ above (see Table 8.5). This is similar to findings by Lam and Shulha (2015) who observed in a DE case study the role of evaluation data to offset uncertainty and provide momentum in an innovative initiative.

Individual to collective. Collective sensemaking requires an ability to transition from individual-level to intersubjective or collective interpretation (see Figure 3.3). The study findings

offer examples of evaluators enabling this transition as well. The case studies highlighted explicit attention to improving both individual *and* collective capacity in detail. Individuals' "facility" (c41) to make sense of and use data was understood as a skill that needs to be developed, in the belief that "to make meaning and have that inform next steps is a skill everybody should have" (c42). In the case accounts, the use of runways helped individuals and small groups to develop their thinking, and then to work in larger groups to arrive at a shared interpretation. The use of "epistemic objects" as tools to mediate interactions, enable collective interpretation and to codify, or 'anchor' the results was a feature of multiple cases. Some were used to support collective memory within the team, to track progress, and as boundary objects to communicate with more distant stakeholders. Articulating or codifying meaning in a tangible object as a collective effort also represents a further transition to distributed sensemaking (i.e., across tools and actors in a system) (see Chapter 2).

Interpretation of data to action, via judgement of significance. My conceptual framework recognized a need to actively transition between individual and collective interpretation, and implied that collective interpretation leads to action. It did not account for an important intermediate step. There is a subtle but important distinction to be made between interpreting data to arrive at shared understanding and determining significance for action. It involves a move from retrospective thinking (what happened?) to prospective thinking (what should we do next?). Evaluative judgement has to do with significance in this way (see Kushner, 1996 for a discussion). Determining significance also requires us to attend to the consequential validity of proposed actions in a context (e.g., Kane, 2001). This is why some argue that evaluation plays a critical role to bridge the "problem space" and the "solution space" for social programs (Tannahill & Sridharan, 2013). Such a shift from interpretation of data to significance for action is reflected in the question framework used in at least two of the cases: "What, So What, Now What"? (see Eoyang & Halladay, 2013) and the use of theories of change as frameworks for inquiry, because they can anchor data analysis and interpretation in the anticipated future consequences of actions. Significance is also thought to be a characteristic of knowledge as opposed to information. Knowledge includes the judgement of the significance of information *for a particular need* (Tsoukas & Vladimirov, 2001, cited in Greenhalgh, 2010). In short, this is missing in my framework, but is an important bridge between interpretation and action in a way that takes context into account; it is also another area of specialty for evaluators.

Experience to abstraction. Closely related to the above is a move from direct experience (i.e., individual experience, particular data) to shared models, which was another transition in the conceptual framework. As noted above, the use of tools (question frameworks, epistemic objects) are deeply implicated in this transition. Others have proposed that tools may serve to anchor ways in which actors have ‘reframed’ their thinking through a learning process (Engeström, 2006) and may also underpin a transition from retrospective to prospective thinking (Stigliani & Ravasi, 2012).

Shared models do not create themselves. This study’s findings highlighted challenges involved in creating shared models, including tensions that the process may generate. In the case studies, developing shared models required teams to confront internal differences in their ideas and/or to extend their thinking beyond proximal action to broader and longer-term systemic effects (see e.g., TOC development to incorporate systems thinking in Case 4). The case studies also included examples of ways this transition can be accomplished with the help of specialized support. These included what not to do, for example, *not* standing up “with a whiteboard and a marker and say[ing], ‘ok what’s your theory of change?’ Because it’s torture for everyone” (c11), rather, asking questions of individuals and supporting them in a collaborative modeling process over a sequence of steps.

Weick and Sutcliffe (2006) warned of risks from developing abstractions, because they disengage actors from experience and can introduce superficiality and distortion (see Chapter 2). They create new frames that can be both helpful and limiting. This reinforces the importance of paying close attention to the tools that we construct and use because of their potential to influence sensemaking. At least three of the case studies (1, 2, 4) and the CM describe a process of iterative modeling and real-world testing, aimed to explicitly critique models by contrasting them with new data from the field. Whether this serves to mitigate the risk proposed by Weick and Sutcliffe depends on the quality of that process and the strength of the overall learning system.

Models may serve to privilege some interpretations over others and / or to embed a preferred account in future decision making (see Chapter 2). My conceptual framework made no allowance for power dynamics; how to resolve competition among ideas and interests in the course of a collective inquiry process. This surfaced, however, as an important factor in both

strands of the study (see e.g., the interplay of Clusters 3 and 4 in Domain 2, Chapter 7). Midgley and Lindhult (2017) have cautioned that it is a mistake to assume, “that stakeholder participation in dialogue will allow the better argument to prevail, which ignores (or overly minimises) problems of coercion, groupthink, deceit, ideological framing and disempowerment” (p 11). Kushner (2015b) adds that, among multiple theories of change that may be at play in a program, some may be unrecognized by actors themselves even though they are operational. Others may be recognized but hidden TOCs that guide action even while actors use a ‘mandated’ TOC “as a subterfuge” (n.p.). The likelihood that theories about change will be multiple and diverse is recognized in these findings; actively surfacing assumptions and implicit theories about how change happens is described in the case studies and is an explicit function of the activities in Cluster 4 of the concept map. Whether these activities can surface intentional “subterfuge” depends on other factors relating to collaborative capacity, including trust and commitment to principles of collaborative inquiry. Activities in Cluster 3 exist to help productively resolve differences.

Van de Ven and Johnson (2006) call conflict "the generating mechanism of a dialectical process of inquiry... [and] an inevitable part of work among diverse investigators who hold pluralistic views of a given reality" (p. 809). By extension, if conflict is suppressed, so are inquiry and learning. For these reasons, conflict and its management are part of the focus for a mediator of an AL process. In keeping with the literature on diversity and conflict (see Chapter 2), Van de Ven and Johnson (2006) argue for some encouragement of task-related conflict along with "creative management" of interpersonal conflict (p. 809, citing Jehn, 1995; Flyvbjerg, 2001; see also De Dreu, 2006). Their proposal is reflected in the Cluster 3 / 4 interaction (see Chapter 7).

Sensegiving. This study did not document processes of sensegiving (see Chapter 2). However, opportunities for sensegiving to occur are evident. Where power differential may result in some accounts being emphasized over others, a commitment to and skill to promote equidistance in the inquiry may be critical (see the technical expert role, Chapter 6). As noted above, the attention to different perspectives is essential. Perspective affects the boundaries that are set on an inquiry influencing what is considered significant (what is attended to), how information is filtered and interpreted, and how significance for action is judged. This is said to

be important in the “fuzzy front end” of innovative initiatives (Midgley & Lindhult, 2017, p. 3) and is part of ongoing alignment (see Table 8.3). This study described case teams working to maintain a commitment to inclusion, voice, and balance, which might help to protect against sensegiving. Principles may offer heuristic value if advanced by a mediator. The possibility that mediators themselves may engage in sensegiving is taken up below under implications for practice.

In sum, the findings of this study lend support to the importance of multiple transitions in a collective sensemaking process. The findings describe these transitions occurring iteratively and suggest how they can be supported by mediation. The findings describe ways in which evaluators attempted to balance doubt and confidence, and abstract models and field-level data, for example. The findings also suggest omissions in the original framework that are important for an effort to balance multiple perspectives in collective sensemaking. Awareness and attention to these may help to strengthen the process as individual interpretations are integrated into models and tools that guide collective action.

As noted above, research on collective sensemaking offers insight into group learning processes. Stepping back out to a more abstract level, the study findings suggest three broad conditions are needed for adaptive learning. In the next section, I propose a tentative model of AL through the integration of these conditions.

Integrating Three Conditions for Adaptive Learning: A Proposed Model

The study findings describe four roles for the evaluator: as technical expert, guide, capacity builder, and learner. As described in Chapter 8, the technical expert’s activities include both opening up and grounding the inquiry, and the guide helps to challenge established ideas and provoke new thinking about an intervention, and also to bound thinking by framing up working models and course correcting over time. Activities in these roles can be compared to efforts to advance both divergent and convergent thinking as complementary drivers of adaptive learning (e.g., Cropley, 2006). These are the first two proposed conditions for adaptive learning.

In addition, the capacity builder and learner roles support development of underlying conditions for productive collaboration (see Figure 8.5) in terms of, for example, skills, habits, structures, power balance, identity and understanding of context. These align with a framework

for understanding collaborative capacity developed by Foster-Fishman et al., (2001). These authors developed their framework in the context of coalitions for community and systems change interventions. The framework includes four categories of capacity: Member Capacity (e.g., relevant skills and knowledge of individual members and their motivation and attitudes); Relational Capacity (e.g., working climate, shared vision, shared power, value placed on diversity); Organizational Capacity (e.g., effective communications, procedures, resources); Programmatic Capacity (e.g., goals, systemic validity). Collaborative capacity is suggested as a third proposed condition for AL.

Divergence and convergence in partnership. Divergent thinking is implicated in the generation of novelty and convergent thinking in the evaluation of novelty (Cropley, 2006; see also Smith & Smith, 2017). The two are present in the definition of an “intelligent collaboration”, which is characterized by an ability to generate diverse ideas as well as to consider them critically and productively (B. Davis & Sumara, 2006, see Chapter 3).

Creativity is often attributed to divergent thinking (Smith & Smith, 2017; see also Schalock, Verdugo, & van Loon, 2018 in the context of evaluation), and convergent thinking is sometimes thought to inhibit creativity. However, both divergent and convergent thinking together may be important to a creative process (Cropley, 2006). This study’s findings suggest a tight interplay between the two that may be generative, in the sense that convergent thinking, in addition to enabling selection among novel ideas, may also elicit divergent thinking. For example, helping to deconstruct and delineate ambiguous concepts and articulate them in terms of significance for action can prompt new thinking and new ideas. This took place in Case 2 in which a participant reported difficulty “seeing” ambiguous concepts until they were actively defined, which then prompted new lines of thinking about options and opportunities. Something similar has been observed by others. For example, Lam and Shulha (2015) report in a case study of DE the use of analysis to help a group “unlearn” inaccurate assumptions and then set them aside (p. 365); this was essential for the group to then raise and explore alternative ideas. Peurach, Glazer and Lenhoff (2016) link divergent and convergent thinking to processes of exploration and exploitation, respectively. They assert that both exploration and exploitation are important for collaborative learning under complex conditions and propose that DE has potential to support their interaction. Table 9.1, below, provides some examples of divergent and

convergent thinking as defined by Cropley (2006) compared with activities in the technical and guide roles identified in this study. Further below, I propose that the study findings support a perspective that divergent and convergent thinking together can advance AL.

Table 9.1. Examples of divergent and convergent thinking.

	Cropley (2006)	Technical Expert	Guide
		<i>Opening up</i>	<i>Provoking, Momentum</i>
Divergent Thinking	• Seeing the known in a new light • Shifting perspective • Seeing new possibilities	• Allow actors to see the unexpected. • Help to illuminate insight of people around the table.	• Help the team to view the initiative through a new frame. • Enable conversations that explore ambiguity and push for new insights.
		<i>Grounding</i>	<i>Staying on Track</i>
Convergent Thinking	• Combining what “belongs” together • Achieving accuracy and correctness • Testing and checking	• Bring more rigour to observations. • Use evaluation tools to bring more certainty to stakeholder hypotheses; walk on more solid ground.	• Help surface what the findings are pointing to. • Consolidate and make sense of the data; what it means to the organization for implementation. • Frame up emerging working models.

Figure 9.1, below, builds on a model by Cropley (2006). It proposes pathways arising from combinations of divergent and convergent thinking supported by collaborative capacity. For simplicity, only three pathways are included. One of these fosters AL. The other two result in either nonadaptive learning or maladaptive (superstitious) learning. I outlined the literature on superstitious learning in Chapter 2. This occurs through a “compelling” but misleading experience of learning – misleading because relationships between actions and outcomes are misapplied (Levitt & March, 1988, p. 325). As long as actions continue to be perceived as successful, this can lead to overconfidence without a corresponding increase in competence. Deliberate learning processes may inhibit a reinforcing loop (Zollo, 2009). Cropley’s (2006) model proposes results and risks of creative processes in which divergent thinking occurs either with or without convergent thinking. In Cropley’s model, either option can result in overconfidence. In Figure 9.1, I build on and modify Cropley’s model to speak to AL.

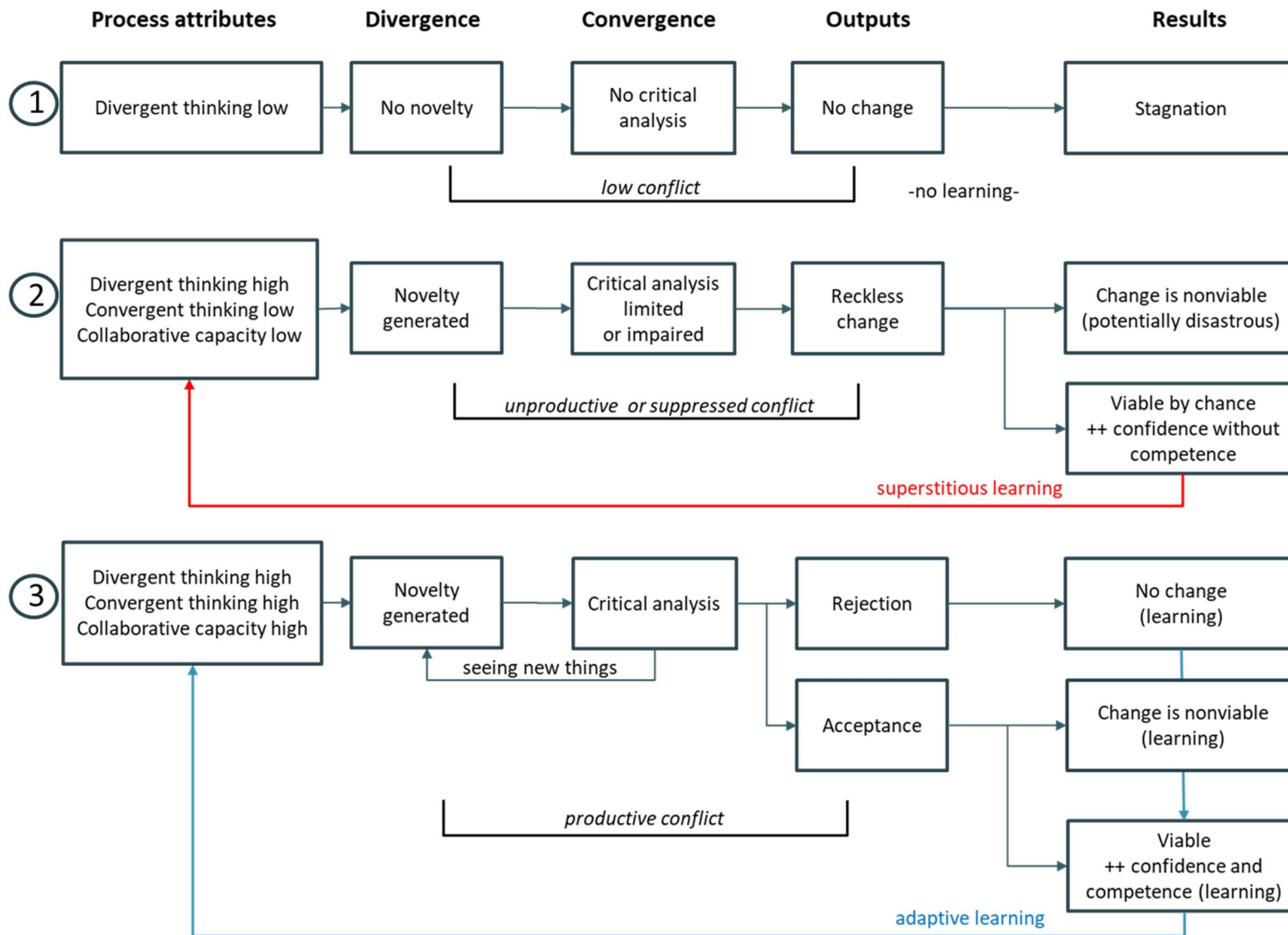


Figure 9.1. A model of proposed pathways.

Three pathways are depicted in the model. Low levels of divergent thinking are illustrated in Pathway 1. In this pathway, learning is limited to the refinement of the status quo and is nonadaptive to changing conditions. Pathway 2 illustrates strong divergent thinking in the absence of convergent thinking. Where successful by chance, this pathway introduces a loop of increasing confidence with an impaired learning process. The pathway represents something like Campbell's (1960) model of creativity through blind variation and selective retention (cited by Smith & Smith, 2017). Obviously, this approach is a concern for social interventions, given the potential of negative consequences for vulnerable people; even more so if it is reinforced through a maladaptive learning loop. The third pathway is proposed to be the most likely to support AL. It contains both divergent and convergent thinking, as well as an internal cycle between them that may be generative, provisionally labelled: "seeing new things". Three proposed outcomes of this pathway include warranted nonaction, nonviable action and viable action, all carrying potential for learning, which leads to a feedback loop supporting AL.

Figure 9.1 also makes note of the importance of collaborative capacity. In this model, conflict is proposed to emerge or intensify as novelty is generated, consistent with this study's findings (see interpretation of Domain 2, Chapter 7). As Pathway 1 does not generate novelty, conflict remains low. Pathways 2 and 3 do generate novelty. Low collaborative capacity is illustrated in Pathway 2, and consequently conflict is proposed to either be suppressed or to otherwise interfere with the remainder of the process. Pathway 3 illustrates high collaborative capacity which enables productive conflict. (This pathway would be similarly dysfunctional if collaborative capacity were low). The proposed model illustrates how all three conditions are necessary to AL.

Figure 9.2, below, is a simplified depiction of the three conditions. At least these three seem necessary to create space for AL to occur. The findings from this study have described ways in which mediation can support these three conditions. It is also important to reiterate the importance of boundaries to this process. In the integration analysis, boundaries on the inquiry surfaced as a key element of making space. They define the process and affect where it is located by actors among multiple possible places. These two models are proposed as input for future research, as elaborated in the next chapter.

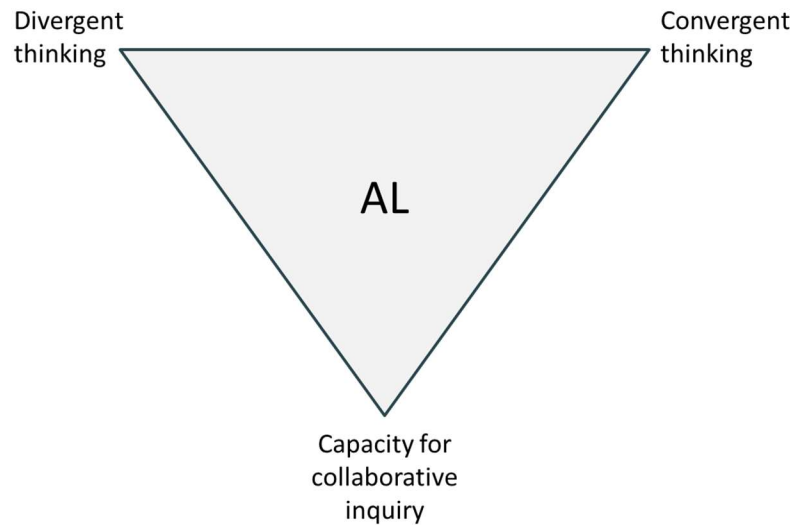


Figure 9.2. Three conditions to create space for AL.

Overall, this study reinforces the importance of attention to the learning system surrounding an intervention that is actively adapting. It supports a view that the quality of the learning system it is integral to what the intervention becomes. The learning system functions, in effect, as an agent (e.g., Salthe, 1993).

Implications for Practice

The study's findings highlight important considerations for practice. I will note three that stand out for DE and which also seem useful to a broader understanding of mediation in AL. The first concerns readiness of program actors. The second revisits an observation that aspects of the practice may be “invisible” to stakeholders. The third relates to boundaries on the role.

Developing readiness. DE explicitly aims at adaptive learning with the aid of a specialist intermediary. It can be an intensive process and an unfamiliar way of working for some actors, which raises the importance of what is described as stakeholder readiness (Cabaj, 2014; Patton, 2016a). The study's findings reinforce an assertion made in the DE literature that to be successful, DE needs to fit well with the nature of the intervention (does it need development?), the purpose of the evaluation (is there an explicit developmental purpose?), and characteristics of

the context (e.g., stakeholder adaptive capacity and readiness to engage in evaluation) (Cabaj, 2014; Patton, 2016a)³¹. With respect to stakeholder readiness, Patton (2016a) provides examples of readiness characteristics, including stakeholders' tolerance for ambiguity, commitment to adaptive learning and "preparation to stay the course" of long-term engagement (p. 13). McKegg and Wehipeihana (2016) write of dispositions, such as an inquiring mindset and perseverance.

This study's findings suggest that up-front assessment of actors' readiness is likely to be imperfect, particularly for multi-agency collaborations. Individual readiness may vary widely, and differences may not be apparent until actively surfaced. Others have noted that a group's capacity for collaborative evaluation can also decline over time (e.g., due to leadership turnover, unforeseen political events) (Shulha et al., 2016). This begs questions: should a DE not proceed if (some of) the actors are not ready? How ready do they need to be? If readiness is a dynamic condition, what then?

Cabaj (2014) suggests that if a group's readiness for DE is low, a decision to move forward should include a commitment to intentionally strengthen weaknesses. McKegg and Wehipeihana (2016) write that developmental evaluators need to "carve out space" for people to collaborate in evaluative inquiry in complex initiatives and that this needs to be a "key responsibility" for evaluators in DE (p. 284). This argument has been made for participatory and collaborative evaluations more generally, that is, that space needs to be actively created and maintained, because this space "hold[s] the potential for transformation and change" (Chouinard & Milley, 2016, p. 3).

This study's findings support a view that actively developing and sustaining capacity (making space) is necessary but very challenging to achieve, requiring time, effort and skill. Particularly for multi-agency AL, a specialized intermediary can make an important contribution. If we do not assume that "readiness" is a latent or fixed factor of context, then these findings suggest a practitioner also needs to look for conditions that can sustain the work while readiness is developed. McKegg and Wehipeihana (2016) suggest some starting points, such as ensuring that comfort with ambiguity is at least present with internal champions and leadership.

³¹ Cabaj (2014) provides a checklist for assessing readiness in DE. Other guidance for assessing readiness for evaluation are available, notably Preskill and Torres' (1999) instrument to assess readiness for organizational learning and evaluation. These focus on stakeholder readiness. The evaluator's readiness is usually assumed, but in the context of DE in particular should be questioned, given the duration and intensity of the process and the skillsets involved (McKegg & Wehipeihana, 2016).

Possibilities noted in these findings include structural interdependency, funder commitment to collaboration, active participation by funders, and availability of champions (see Chapter 8). These are likely to vary by context. Of note, interdependency has been proposed as important for collective sensemaking (Weick & Roberts, 1993, cited in Drazin et al., 1999) and for productive problem solving in diverse teams, including in evaluation (see Chapter 2). The role of senior organizational members as champions to enable evaluation capacity building (ECB) has also been described outside of DE (Cousins & Bourgeois, 2014).

Evaluation capacity building (ECB) literature typically positions ECB as a separate activity from carrying out an evaluation itself, and generally at the level of a single organization (Labin et al., 2012). The ECB literature also tends to emphasize skills and knowledge coaching (Ensminger et al., 2015), although some frameworks explicitly recognize the importance of aspects of organizational culture, attitudes and leadership in the organization (e.g., Preskill & Boyle, 2008) or organizational structures as antecedents to ECB (Cousins et al., 2014). This study describes ECB intentionally integrated with evaluation, and as an effort akin to development coaching as described by Ensminger et al. (2015). Development coaching is aimed directly at increasing capacity in organizations to be willing to learn, to value evaluation and take on evaluation tasks, and to create structural changes in the organization to embed evaluation. In the chapters above, the findings include ways the evaluation teams supported adaptation of system structures to incorporate evaluation, as noted in the section ‘evaluating quietly’. They helped develop relationships within the learning system, in part through demonstrating a commitment to co-creation and reciprocity and altering information flows (e.g., by establishing shared access to data across organizations and levels). They influenced group identity with respect to learning. Accounts emphasized building capacity to ask questions and to interpret and use data in addition to, or even more than, technical skills. Their efforts go beyond a conventional focus by evaluators on the technical aspects of a program, such as program implementation and program theory, to consider more fully the program’s “sociology” (Kushner, 1996, p. 194), by intervening in the system around the intervention. Ensminger et al. (2015) argue that development coaching requires more time, effort and a different skillset than conventional ECB. They describe this as “the most advanced form of coaching”, but still undeveloped in the ECB literature (p. 133). They call for more work in the field to develop competencies for this type of ECB.

The invisible DE. This study presents the work of DE as integrated and distributed: it is integrated (ideally) into the intervention's day-to-day operations and it is distributed over time and among program actors. Some of its most important outcomes are intangible. DE is about "the thinking that people do along the way. That's often very invisible" (c11). As noted in Chapter 7 and Chapter 8, entire groupings of activities, such as those in Cluster 3, may be difficult for stakeholders to recognize. Finally, the evaluator's performance of multiple roles and their multiple relationships may make their function more ambiguous than in other approaches.

This is consistent with findings by Strike and Rerup (2016) on mediated sensemaking. They cite Obstfeld (2014) to say that mediated sensemaking is "invisible", because of its distribution in practice over time and space: "and therefore less available for instantaneous observation ... [It] is inherently more complex and ... difficult for ... the social scientist to detect, track over time, and theorize" (p. 898). Strike and Rerup (2016) suggest the ambiguity of a mediator's role might explain why mediators have been almost entirely overlooked in research on sensemaking. Its ambiguity may make DE more challenging for stakeholders to understand, including those commissioning evaluations and those considering investing their time and energy to participate in DE. At a practice level, intermediaries may need to invest substantial effort to help program actors to "see" the process as part of a capacity building effort.

Yet at the same time, ambiguity may be helpful. The capacity to move across roles may be integral to mediation in AL, and such freedom of movement may be enhanced by a combination of a "named" role (e.g., to offer critique) and ambiguous positioning. Research on mediated sensemaking locates the mediator in between other, better documented functions. This research describes an unusual location that spans social boundaries, which exposes mediators to heterogeneous information and allows them to broker it across siloes (Heaphy, 2017; Stigliani & Ravasi, 2012; Strike & Rerup, 2016). Strike and Rerup (2016) describe mediators "oscillating" between locations across boundaries, and therefore being never always "inside" nor "outside" nor always at the periphery. This resonates with findings from this study. It is visible for example in participant accounts of relationships of different types and at varying degrees of distance from the core development team, and in accounts that portray the location of the evaluator variably, shifting between "inside" and "outside" the intervention (Chapter 5).

This appears to help an intermediary, because they can operate outside of formal structures; they can take on activities that are not performed within or by the formal organization and they are less restricted by established hierarchies, offering freedom or flexibility that other actors do not have (see Chapter 2). Boundaries are often thought of as walls or places of conflict. Others have suggested that boundaries can be generative and that they are an active part of a system (Byrne & Callaghan, 2014, p. 32 referencing Zeleny, 1996). The mediator may be specializing in this boundary space. Cooper (2012) calls this the “white space” in organizations, arguing that this is where knowledge intermediaries function (p. 37, from Rummeler & Brache, 1991). The implication is that a too-narrowly defined function and role may disadvantage an AL mediator.

Pelling et al. (2008) write that relational spaces that support adaptive capacity are shaped by an interplay between formal and informal systems of rules and norms. Formal systems are relatively easy to establish for a new collaboration (e.g., via agreed terms of reference, official procedures) compared with informal systems. Moreover, informal systems are sometimes deliberately unrecognized by the formal system (Pelling et al., 2008), which makes them harder to discern. Working between, brokering and linking these systems presents opportunities to develop adaptive capacity, but this is an “unregulated space” in which trust provides a “quality control function” (Pelling et al., 2008, p. 31). Along with the prospect that ambiguity may be beneficial to a mediator’s role, this brings us to questions about how role may be appropriately and constructively bounded.

Bounding the role. Although lacking formal authority, mediators intervene more extensively in the system than implied by labels such as “facilitator” or “broker”. They potentially have greater leverage at their disposal, and more influence, particularly as represented by the guide role (Chapter 8). Aspects of this role have been recently noted in the evaluation literature: Baldwin and Lander (2018) report findings from a DE case study in which the evaluators’ activities included “pushing and compelling [program implementers] to act in new ways especially as a collaborative project team” (p. 17). Chouinard and Milley (2016) remind us that evaluation itself carries an “inherent cultural authority” (p. 7), which extends an evaluator’s potential influence in a less overt way. This presents ethical questions for the practitioner

because it challenges a traditional view of the evaluation function as detached and neutral (Greene, 2000b).

Actors are believed to engage in sensegiving behavior when they perceive a gap in the sensemaking of others (Maitlis & Lawrence, 2007). Of note, at least two empirical studies have identified examples of sensegiving by mediators (Heaphy, 2017; Stigliani & Ravasi, 2012). Another study denies a sensegiving role, but reports mediators selecting which issues to raise – that is, judging which issues were most important or highest priority, and reducing the amount of information overall to allow actors to focus more deeply on issues that are considered by the mediator to be “core” (Strike & Rerup, 2016, p. 895). This scenario involves selection and framing of information; the mediator may or may not be aware of the nature and degree of their influence. Cilliers and Preiser (2010) argue that selection and framing are unavoidable:

Since we cannot deal with complexity in its complexity, we have to reduce that complexity when we try to understand it. There is no objective way to do this reduction – that would imply a meta-position which can deal with complexity fully – thus we cannot reduce our encounters with complexity to calculations. There are always choices, and therefore always normative elements involved (p vi-vii).

From a complexity perspective therefore, forms of sensegiving exist in any inquiry concerning a complex intervention.

Self-imposed boundaries on the evaluator role are noted in the findings from the case studies (see Chapters 5 and 6), for example with respect to caution about their influence on program decision making. This indicates awareness and concern by the evaluators for boundaries on their role. This awareness is also raised in the DE literature, for example by Patton (2011), who describes: “an obligation on the part of the evaluator to represent the standards and principles of the profession as well as his or her own sense of morality and integrity” (p. 314). Both Murphy (2016) and Asher et al. (2016) illustrate intensive work by DE evaluators to observe boundaries, by supporting program development teams through the questions of “what is happening?” and the “so what does that mean?” but keeping the “now what do we do?” semi-separate and more in the hands of program actors to determine. This marks a line of distinction between problem and solution spaces in evaluation identified by Tannahill and Sridharan (2013),

and the point of judging significance, which lies in between interpretation and action as noted in Chapter 2. This also aligns with the accounts of informal bounding of role reported in Chapter 6. This awareness and attention are important, and can be aided by more explicit and concrete discussions about boundaries for intermediaries, including evaluators, and especially how they might be maintained through such a complex process as AL. They may help to build awareness of and reflection on sensegiving by intermediaries in learning processes. Such discussions can be supported by continuing the open debate in the evaluation field about our commitments to, and assumptions about, neutrality and impartiality.

Neutrality and impartiality. The idea of impartiality or neutrality in evaluation is paralleled in the field of mediation, where Rifkin et al. (1991) called it “a folk concept”, writing that “there are tacit understandings of what it means and how it works in practice”, which “mystify” the concept (p. 151-2). These authors report that most mediators equate the terms neutrality and impartiality and describe them as about maintaining unbiased relationships (not favouring any group over another) and not being influenced in their actions by their own emotions, values or agendas (p. 152). Rifkin et al. discuss the concept of “equidistance” as an intentional practice relating to neutrality, noted above in the findings in Chapter 6.

In the case of evaluation, neutrality and impartiality are used to convey a position an evaluator adopts, that is, one that is not partial to the program or program operators, with intent to bring a neutral or detached perspective on the program. As observed by a participant in this study (see Chapter 5) preserving both impartiality and the perception of impartiality are important and may be essential to trust in an evaluation. Langlois, Blanchet-Cohen and Beer (2012) note that, unlike program actors, a developmental evaluator can be seen as having no agenda, which is important to trust and credibility in DE (p. 49). Baldwin and Lander (2018), suggest a developmental evaluator’s influence may be resisted by program actors if they are not seen as neutral, and Mostert et al., (2007) report a similar finding from a study on facilitators of participatory processes in environmental initiatives. Patton (2016a) notes that an evaluation’s credibility is a key determinant of use, and that credibility has much to do with how the evaluator approaches the work and collaboration with stakeholders.

A number of published case narratives on DE have described structures to support impartiality in the evaluation, such as hybrid internal / external evaluator teams, with external

evaluators performing key tasks (e.g., Togni et al., 2016). This solution was also reported in two cases in this study (Case 1 and 4), in which external researchers or evaluators were involved, especially for summative work. Such a structure is also suggested for the evaluation of place-based approaches by Horizons Canada "with an evaluator close to the project conducting an evaluation for the learning process, and a separate evaluator fulfilling a challenge function, drawing on both internal and external expertise" (Bellefontaine & Wisener, 2011, p. 19).

These are examples of efforts to find balance and draw on strengths of both inside and "outside" positions. It reflects efforts to navigate a shift in the field of evaluation, recognized by Greene (2000b), to critically question the idea of detachment of evaluators. Greene (2000b) highlights the multiple locations in which evaluators have positioned themselves in the history of the profession. One is as impartial servants of decision-making, privileging objectivity and truth and relying on methods. Another is as a facilitator, modelling a way of thinking (evaluative thinking, problem-solving), and emphasizing understanding of context and stakeholder needs to help program actors improve their practice. A third position is as a change agent:

As change agents, we attend to the process of inquiry – its opportunities for meaningful participation, its attention to power and privilege, its insistence on equal and reciprocal communications and relations among all stakeholders... we structure our evaluation process and its substantive foci to enable meaningful opportunities for action and change... beyond [providing] information, we enable action and change (p. 9).

In her reference to how evaluators structure the evaluation process, Greene calls attention to a view that there is, in effect, no place outside the "systems of mutual causality that [we] participate in" (Midgley, 2008, p. 60). And, as asserted by Chouinard and Milley (2016) a "participatory space", which evaluators help to shape, is never neutral (p. 3). This makes it imperative that an intermediary involved in inquiry attend to power, privilege and participation, and recognize and consider their own influence and how that might manifest, even unintentionally, in the learning process.

In the next chapter, I close this dissertation with a few thoughts on the study's contributions, its limitations, and the importance of continued research on the topic.

Chapter 10 Conclusion

I am writing this at a time when complex or “wicked” social problems are at the forefront of our awareness. We are witnessing intensifying political, social and environmental challenges and waning confidence in our ability to engineer solutions. We are also seeing parallel growth in large scale, multi-agency interventions that seek change at the level of systems. We are reimagining possibilities – for how we organize ourselves, how we interact and how we learn, driven by a sense of urgency. We are engaging with complexity on multiple fronts.

This study was motivated by awareness of the challenges in the design and development of complex interventions, and the importance of meeting these challenges. There is some research to suggest that diversity in perspectives, if productively engaged, can feed learning in complex situations; it is described as a “source of intelligence” in a learning system (B. Davis & Sumara, 2006, p. 858). Accordingly, approaches to social intervention like collective impact and social innovation explicitly seek to enable participation from people in different sectors, organizations, roles and social groups for situated learning and development. The challenge is to move from a recognition of the potential of diversity, to an understanding of how this might be realized to effect positive change in the world (see e.g., Pelling et al., 2008); in other words, to contribute to what some call a practice of pluralism (Global Centre for Pluralism, 2017).

Third-party intermediaries may play an important role, and this makes empirical research on what they do and how their work unfolds vitally important. The purpose of this study was to contribute by exploring mediation in adaptive learning as applied to social interventions. Although a virtually undocumented field in its own right, AL mediation is visible in existing professions. Evaluation is one of them. The evaluation field, in many ways, is at the forefront of thinking in these terms.

I focused on developmental evaluation because it is an approach in which an intermediary (the evaluator) actively intervenes with actors around a complex intervention, with an express purpose to support its development through an adaptive learning cycle. The multi-case component of the study explored the role of evaluators in multi-agency collaborations that were innovating in their sphere. The concept mapping component outlined and explored the structure of the practice. Through a systematic integration of the two components, this study

produced a rich description of AL mediation through the lens of DE. The findings point to several important implications for practitioners and researchers. Much of what is important for practitioners has been elaborated in Chapter 9. In this final chapter, I focus more on implications for research and theory, but first I provide a brief summary of the study's main contributions as well as its limitations.

A Summary of the Study's Contributions

First, the study provides empirical support for and elaborates the challenges diverse groups can face to engage productively in an adaptive learning process. The study outlined situational factors that can impede the process, including conflicting norms and attitudes toward learning and weak levels of trust. With respect to evaluation, members of the group may also need to “unlearn” preconceptions of what evaluation is and what evaluation does.

The study then described ways in which intermediaries made space for collaborative inquiry, through strategies such as cooling, aligning, normalizing and pacing the inquiry process, and using and adapting structures. Importantly, this also included modeling principles and creating opportunities for actors to experience both the process and results that actors valued; in other words, demonstrating worth in concrete and relevant ways. Making space was a primary theme in the cross-case analysis and spanned all domains of the concept map.

The importance of boundary-setting and alignment of the inquiry over time was highlighted in the integration analysis. Boundaries define an inquiry space; they are chosen and maintained, sometimes implicitly. Some assert that, from a systems perspective, it is a responsibility of evaluators to critique these boundaries (Gates, McNall, & Williams, 2018; Hargreaves, 2018). This study documented evaluators walking a fine line to respect boundaries set by others as part of a process of making space. It also suggested the importance of heightening actors' awareness of the effects of boundaries and negotiating their locations in ways that can advance the inquiry.

Four components of alignment were described related to focus, timing, subject and rationale. These were proposed as a framework for understanding alignment as an iterative and bidirectional process, that is, tailoring of the inquiry to actors' needs and developing actors' capacity to engage with this form of inquiry.

The concept mapping process pointed to five categories of activities for advancing the development of the initiative itself, including collective analysis and reflection, course correcting and framing models. These categories were elaborated in the integration analysis with attention to how these activities were expressed in the cases. They shed light on another facet of the mediation function focused on the intervention, and they touch directly on closing the gap between provision of data and data-informed program development (referred to as the “know-do” gap in KMb, see e.g., Greenhalgh & Malterud, 2017). The findings provide concrete examples of ways that a mediator might gain leverage and momentum for change, included soliciting the aid of champions, clarifying ambiguous ideas, prototyping and testing, and anchoring progress in tools and structures. The findings also draw attention to the use of data, goals and principles as signposts for navigation.

The study further documents roles for intermediaries as indicated in these data, which vary in their degree of directiveness and their focus on the inquiry vs the intervention itself (Figure 8.5). These roles suggest a strategic range that can enable shifts between nondirective and more directive positioning, attention to multifaceted relationships and balance of power, and influence on both the inquiry and the intervention. They represent a potential to work across boundaries vertically within organizations and laterally across them, and between explicit and implicit systems (the formal and the informal systems) surrounding the intervention. This links to findings elsewhere (e.g., Pelling et al., 2008) and hints at an intriguing direction for future research.

Four specific roles were documented in detail. For example, activities of the technical role involved tailoring methods and advocating for the data at the decision-making table. Activities point to efforts to open up the process of inquiry and foster exploration, and also to focus, frame and anchor it, particularly across the two more directive roles (guide and technical expert). While this may seem contradictory, it signals value placed on both divergent and convergent thinking as integrated parts of the process. The other two roles, capacity builder and learner, concentrated on establishing enabling conditions to keep the process from running aground, or to provide options to free it when stuck. The capacity builder role in particular was associated with key turning points in case trajectories. Through these roles the evaluators paid particular attention to relationships among members of the development team, that is, “creating a

different kind of conversation” among them that was more inquisitive, egalitarian and learning-motivated (c12).

Role fluidity and ambiguity are an opportunity and a challenge for an intermediary. Meeting the challenge has implications at multiple levels. At an individual level, awareness of and ongoing attention to role boundaries is warranted and was evident in the data. At a structural level, this is one rationale for multi-member DE teams, not only to provide complementary skillsets to span multiple roles (see e.g., Cabaj, 2011), but also to support each other to navigate roles and their boundaries. Teams in these case studies sometimes included hybridized structures or contracted periodic external support to add impartial and/or more distanced perspectives on the evaluation. Within the field of evaluation, locating appropriate boundaries to an evaluator’s role is a topic of ongoing discussion, even controversy, and continued open debate is important.

Agency came to the fore in the critique of the original conceptual framework, as the findings demonstrated actors’ ability to influence their context as well as to be influenced by it, given time, patience, effort and skill. Otherwise, the findings provide insight on aspects of collective sensemaking theory, including identified but relatively undeveloped tensions and transitions. These may act as mechanisms to drive the process forward in a productive way. At least two of these are firmly in the comfort zone of evaluators: (1) critique and support to critical thinking that help to sustain doubt, while building confidence by identifying and documenting patterns of progress; and (2) helping link interpretation of retrospective data to judgement of significance for action in the future. A close study of any one of the transitions discussed in the findings could yield insight on mechanisms for AL by multi-agency/cross-sectoral teams.

Lastly, a birds-eye view on these findings suggests a revised definition of adaptive learning which foregrounds the development of adaptive capacity as a primary objective. I proposed a tentative model for thinking about adaptive and maladaptive learning in collaborations. This model integrates divergent and convergent thinking with collaborative capacity.

The process of AL can be disruptive and uncomfortable. In fact, something is probably wrong if it is *not*. It is likely to reduce efficiency (see Chapter 2), and some argue social learning processes in general carry more serious risks. For example, they can lead to entrenched conflict

or provoke active suppression of dissent (Mostert et al., 2007; Muro & Jeffrey, 2008). Entering into the process requires a commitment (see e.g., Cabaj, 2014) that is likely to be resource intensive and long term. This study's findings support this view and suggest concrete opportunities for ways of acting on that commitment, including the use of guiding principles for negotiating the process. For evaluation specifically, the findings invite us to consider how current conceptions of ECB can be extended to cover multi-agency collaborations, deliberate integration of ECB with evaluation work as a matter of course for complex interventions, and expansion of its scope to go beyond emphasis on technical skills to address other important fundamental conditions. Assessing readiness is likely to be fraught with challenges but vitally important. If readiness is dynamic and variable within a diverse group, then for such groups it is not practical to approach as fixed and pre-ordinate whole-group condition, or something that must be achieved before real evaluation work can begin. This *is* real evaluation work.

Overall, this study's findings strongly suggest that an adaptive learning system needs to be built and sustained. It is not likely to be a naturally occurring part of a collaborative initiative or to generate spontaneously from our desire for it. Effectively, adaptive learning is its own wicked problem. There is hope in these case data, which demonstrate the possibility of developing systems for learning that are different from what existed at the outset, and for some participants, different from what they continued to have outside of their collaborations.

Limitations of the Study

This study was designed to document an elusive practice in an emerging area of research, through the study of exemplar cases and the contributions by experienced practitioners. This was an exploratory study with important limitations to consider.

For each of the case studies in Strand 1, I drew on only two sources of data: interviews and document/video content analysis. Feasibility prevented other forms of data collection (e.g., surveying a wider range of participants). Instead, the multi-case study was designed to be systematically integrated with the parallel GCM study as a way to challenge and strengthen findings.

The two strands of the design were not completely independent however: the multi-case participant group contributed in part to the initial phase of the GCM process. The strands are also

linked by me as the sole researcher. Although steps were taken to preserve the distinctiveness of the strands (see Chapter 4), some consistency of findings between them might have arisen because of these linkages. In addition, although the cases in Strand 1 unfolded in different places and focused on different social issues, they were purposefully drawn from a set of examples published in the same edited volume (see Chapter 4). Two cases (1 and 3) shared an evaluator. Two other cases (1 and 4) shared an external champion. This may explain some consistency across cases.

In terms of a mixed methods design, although the two strands of the study were treated with equal weight (see Chapter 4), overall the study placed greater weight on qualitative, interpretive methods than quantitative methods. As well, an exploratory quantitative method was used (hierarchical cluster analysis). While consistent with the study's overall approach and, as a spatial method, consistent with the study of complex phenomena (Wolf-Branigin, 2009), other options exist (e.g., Krause, 2018) that would add different perspectives. A hypothesis-testing, confirmatory method would be valuable to challenge components of the models and frameworks proposed through this study.

It is important to note that these findings are drawn from study of an approach (DE) that explicitly aims at adaptive learning and in which intermediaries perform an active and specialized function. The study draws on “exemplar” cases and input from participants with substantial experience. For reasons discussed in Chapter 4, the findings are meant to inform understanding of a robust, mediated AL process. This study was not designed to explore non-mediated AL processes, or examples where the process is likely to be weak. This would be valuable to pursue in future research.

Evaluation as a field carries motivations, expectations and professional culture that are different from other disciplines. For these reasons, these findings may be limited in their applicability to other contexts. Lastly, having drawn across multiple fields for this study, and given the study's exploratory design and limited other empirical research in this area, it is important to reiterate that the frameworks proposed in this study are intended as preliminary offerings. They are here to be challenged, modified and/or elaborated by future research.

An Agenda for Future Research

This study touched lightly on several topics that deserve their own in-depth and focused attention, and it has raised many questions that it cannot answer. Below I outline a few additional lines of inquiry that to me seem especially promising.

Research on the use of tools as mediators in adaptive learning processes invites further study, and in particular how they are developed, by whom, and in what way. Research on epistemic objects from the organizational learning literature can provide a starting point for investigation of their uses and influences in evaluation and other adaptive learning processes. The possibility that tools themselves may have agency has been suggested by others (e.g., Engeström & Blackler, 2005) and this idea warrants study in the context of evaluation.

As noted above, a comparative study of cases of AL that do and do *not* involve specialized intermediaries could be fruitful. This could also include comparative study of cases adopting collaborative forms of evaluation, including but not limited to DE, and those using non-evaluator intermediaries. Collective impact interventions, for example, commonly employ professional facilitators at events and/or dedicated coordination staff from a “backbone” agency (Kania & Kramer, 2011). This study has helped to trace some features of the terrain that may be useful to research in this direction.

Two of the cases in this study (3 and 4) had very different trajectories, and they suggest themselves as possible contrast cases. This study was not designed to pursue this type of inquiry, but a future study might purposefully select and study cases that vary in their outcomes. This could provide insight into the role of key factors on trajectory, such as degree and nature of interdependency among actors. Considering the proliferation of interventions adopting collaborative learning approaches (Hargreaves, 2018), a case-based comparative method for medium-sized samples, like qualitative comparative analysis (QCA) might a viable and interesting option.

The case examples in this study varied in their intensiveness and duration, but they were all multi-year efforts that required greater time and involvement of stakeholders than is typical in evaluation. This study was not intended to address questions about the relative costs and benefits of processes such as described in this study, or under what conditions they may or may not be

warranted. This is a question to explore in greater depth, especially given possible risks of engaging in AL as noted above. It links to questions raised in this study about what practitioners need to consider with respect to stakeholder readiness and process alignment.

Also with respect to alignment, this study documented extensive efforts to tailor the inquiry to needs in the context. Yet Hawe (2015) argues that interventions that over-emphasize feasibility or are tailored too far to context (e.g., to perceived capacity limits in the context) may lead to superficial and ‘minimally disruptive’ interventions with weak outcomes. If mediation is a worthwhile intervention, to what extent does it need to disrupt the existing system? How might a mediator find the right balance for the context? Overall, much more research is needed on how robust adaptive learning systems are built.

Given the challenges of making space and providing traction for collective, adaptive inquiry, there could be a risk of losing sight of stakeholders beyond the immediate development team. The data here indicate that the evaluators and their program partners maintained a commitment to inclusion, voice, and balance, and they expressed the importance of principles of engagement, for example, reciprocity, co-creation, to support it. The DE literature includes at least one study on the development of guiding principles for a collaboration through DE (Murphy, 2016) and a book on principles-focused evaluation has recently been released (Patton, 2018): these invite research on how principles are developed, modelled and applied over time in real-world interventions, which perspectives they consider, and about their influence on collaborative capacity and intervention trajectory. The potential benefits (or drawbacks) of explicit application of principles developed outside the intervention context, such as Shulha et al.’s (2016) principles for Collaborative Approaches to Evaluation could also be explored.

Lastly, I proposed a tentative model of pathways and minimal conditions for an AL process (Figures 9.1 and 9.2), building on work by Cropley (2006) and the findings from this study. I included this as a potential input to future research. Much more work is needed to understand conditions for AL and their implications for how we conceptualize and define the process. This perspective on AL might also be useful as a lens for more granular level research in collective sensemaking. For example, it suggests study of how iteration and feedback may operate in a process of modifying frames of thinking, influencing actors’ attention, selection and interpretation of data in ways fundamental to a sensemaking process.

Closing

As we face an uncertain future, we need to stretch our adaptive capacity, our repertoire of options, and our: “ability to experiment and foster innovative solutions in complex social and ecological circumstances” (Armitage, 2005, p. 703). Our learning systems feed our adaptive capacity. They are not yet up to the challenge.

In the previous chapter, I drew attention to an idea that a learning system functions as an agent, in the sense that the quality and the integrity of the learning system is integral to what the intervention becomes. We sometimes describe evaluation and social interventions being ‘done to’ people, or maybe ‘done with’ people. To borrow from Wehipeihana, McKegg, Thompson and Pipi (2016), we also ‘do as’. The quality and the nature of our learning system determines what we create for social change.

This study responds to calls for researchers to help advance the learning capacity of our systems to enable their (and our) adaptation (Krucken, 2003). In a rapidly changing, interdependent world, learning how to practice pluralism is a piece of this. As we move forward, we will accumulate more stories of “imagining and seeing differently” (Weick, 2006, p. 1725); hopefully these will also be stories about arrival at breakthrough understanding.

References

- Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of Educational Research*, 81(2), 132–169.
- Alhadeff-Jones, M. (2008). Three generations of complexity theories: nuances and ambiguities. *Educational Philosophy and Theory*, 40(1), 66–82.
- Alkin, M., & Christie, C. (2004). An evaluation theory tree. In M. Alkin (Ed.), *Evaluation Roots: Tracing Theorists Views and Influences*. Thousand Oaks, CA: Sage Publications.
- Amo, C., & Cousins, J. B. (2007). Going through the process: an examination of the operationalization of process use in empirical research on evaluation. *New Directions for Evaluation*, 116, 5–26.
- Argote, L. (2013). *Organizational Learning: Creating, Retaining and Transferring Knowledge* (pp. 31–56). New York, NY: Springer US.
- Argote, L., & Miron-Spektor, E. (2011). Organizational Learning: From Experience to Knowledge. *Organization Science*, 22(5), 1123–1137.
- Argyris, C., & Schön, D. (1978). *Organizational Learning*. Reading, MA: Addison-Wesley.
- Argyris, C., & Schön, D. (1996). *Organizational Learning II: Theory, Method and Practice*. Boston: Addison-Wesley.
- Armitage, D. (2005). Adaptive capacity and community-based natural resource management. *Environmental Management*, 35(6), 703–715.
- Asher, J., Foote, N., Radner, J., & Warren, T. (2016). Science and how we care for needy young children: the frontiers of innovation initiative. In *Developmental evaluation exemplars: principles in practice* (pp. 103–124). New York, NY: The Guilford Press.

- Bakker, A., Kent, P., Hoyles, C., & Noss, R. (2011). Designing for communication at work: a case for technologically-enhanced boundary objects. *International Journal of Educational Research*, 50, 26–32.
- Baldwin, C., & Lander, R. (2018). Developmental evaluator functional role activities and programmatic developments: a case study analysis. *American Journal of Evaluation*.
<https://doi.org/1098214017743586>
- Bandura, A. (1977). *Social Learning Theory*. New Jersey: Prentice Hall.
- Barbrook-Johnson, P. (2017, January). Models as “interested amateurs.” Retrieved January 30, 2017, from <https://i2insights.org/2017/01/24/models-as-interested-amateurs/>
- Befahr, K., & Trochim, W. M. K. (2011). Conflict in small groups: the meaning and consequences of process conflict. *Small Group Research*, 42(2), 127–176.
- Behfar, K., Mannix, E., Peterson, R., & Trochim, W. M. K. (2011). Conflict in small groups: The meaning and consequences of process conflict. *Small Group Research*, 42(2), 127–176.
- Bellefontaine, T., & Wisener, R. (2011, June). The evaluation of place-based approaches: questions for further research. Policy Horizons Canada. Retrieved from www.horizons.gc.ca
- Berta, W., Virani, T., Bajnok, I., Edwards, N., & Rowan, M. (2014). Understanding whole systems change in health care: insights into system level diffusion from nursing service delivery innovations – a multiple case study. *Evidence & Policy: A Journal of Research, Debate and Practice*, 10(3), 313–336.
- Biesta, G. (2010). Pragmatism and the philosophical foundations of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *The Sage Handbook of Mixed Methods in Social & Behavioral Research* (2nd ed.). Thousand Oaks, CA: Sage.

- Biesta, G., & Burbules, N. (2003). *Pragmatism and educational research. Philosophy, theory, and educational research*. Lanham, MD: Rowman & Littlefield.
- Blumer, H. (1954). What is wrong with social theory? *American Sociological Review*, 19(1), 3–10.
- Brown, A., Stacey, P., & Nandhakumar, J. (2008). Making sense of sensemaking narratives. *Human Relations*, 61(8), 1035–1062.
- Brueggeman, J., & Boswell, T. (1998). Realizing solidarity: sources of interracial unionism during the great depression. *Work and Occupations*, 25(4), 436–482.
- Byrne, D. (2005). Complexity, configurations and cases. *Theory, Culture & Society*, 22(5), 95–111.
- Byrne, D. (2009). *The Sage Handbook of Case-based Methods*. (C. Ragin, Ed.). London: Sage Publications.
- Byrne, D. (2013). Evaluating complex social interventions in a complex world. *Evaluation*, 19(3), 217–228.
- Byrne, D., & Callaghan, G. (2014). *Complexity Theory and the Social Sciences: The State of the Art*. London: Taylor & Francis Group.
- Byrne, David, & Callaghan, G. (2014). *Complexity Theory and the Social Sciences: The State of the Art*. London: Taylor & Francis Group.
- Cabaj, M. (2011). *Developmental evaluation: The experience and reflection of early adopters*. University of Waterloo, Waterloo, Canada.
- Cabaj, M. (2014). Developmental evaluation: Diagnostic checklist. Retrieved from <http://www.tamarackcommunity.ca/library/developmentalevaluation-diagnostic-checklist>

- California Evidence-Based Clearinghouse. (2019). Nurse-Family Partnership (NFP). Retrieved January 4, 2019, from <http://www.cebc4cw.org/program/nurse-family-partnership/>
- Callaghan, G. (2008). Evaluation and Negotiated Order. *Evaluation, 14*(4), 399–411.
- Campbell, D. T. (1970). Considering the case against experimental evaluations of social innovations. *Administrative Science Quarterly, 15*(1), 110–113.
<https://doi.org/10.2307/2391194>
- Chouinard, J., & Milley, P. (2016). Mapping the spatial dimensions of participatory practice: a discussion of context in evaluation. *Evaluation and Program Planning, 54*, 1–10.
- Cronbach, L. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*(2), 116–127.
- Cilliers, P., & Preiser, R. (2010). Why difference. In P. Cilliers & R. Preiser (Eds.), *Complexity, Difference and Identity: An Ethical Perspective* (pp. v–ix). Heidelberg: Springer.
- Clariana, R., Koul, R., & Albright, K. (n.d.). Using Pathfinder KNOT analytic tools for comparing and combining concept maps. Retrieved from
http://www.personal.psu.edu/rbc4/CMC_2006.htm
- Clarke, A. (2003). Situational Analyses: Grounded Theory Mapping After the Postmodern Turn. *Symbolic Interaction, 26*(4), 344–358.
- Connelly, J. (2007). Evaluating complex public health interventions: theory, methods and scope of realist enquiry. In *Journal of Evaluation in Clinical Practice* (Vol. 13).
- Cooper, A. (2012). Knowledge mobilization intermediaries in education: a cross-case analysis of 44 Canadian organizations (doctoral dissertation). University of Toronto. Retrieved from
https://tspace.library.utoronto.ca/bitstream/1807/32688/12/Cooper_Amanda_201206_PhD_thesis.pdf

- Corbin, J., & Strauss, A. (2014). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Los Angeles: Sage Publications.
- Cornelissen, J., Mantere, S., & Vaara, E. (2014). The contradiction of meaning: the combined effect of communication, emotions and materiality on sensemaking in the Stockwell shooting. *Journal of Management Studies*, 51(5), 600–736.
- Cousins, J. B. (2004). Minimizing evaluation use as principled practice. *American Journal of Evaluation*, 25(3), 391–397.
- Cousins, J. B., & Bourgeois, I. (2014). Cross-case analysis and implications for research, theory, and practice. *New Directions for Evaluation*, 141, 101–119.
- Cousins, J. B., & Chouinard, J. (2012). *Participatory evaluation up close: An integration of research-based knowledge*. Charlotte, NC: Information Age Publishers.
- Cousins, J. B., & Earl, L. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, 14(4), 397–418.
- Cousins, J. B., Goh, S., Elliott, C., & Bourgeois, I. (2014). Framing the capacity to do and use evaluation. *New Directions for Evaluation*, 141, 7–23.
- Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, 18(3), 391–404.
- Dahler-Larsen, P. (2016). How evaluation contributes to the formation of concepts. Presented at the European Evaluation Society Conference, Maastricht, NL.
- Dave, G., Frerichs, L., Jones, J., Kim, M., Schaal, J., Vassar, S., ... Lovelady, N. (2017). Conceptualizing trust in community-academic research partnerships using concept mapping approach: A multi-CTSA study. *Evaluation and Program Planning*, 66, 70–78.

- Davies, H., Nutley, S., & Walter, I. (2008). Why “knowledge transfer” is misconceived for applied social research. *Journal of Health Services Research & Policy*, 13(3), 188–190.
- Davis, B., & Sumara, D. (2006). *Complexity and Education*. London: Lawrence Erlbaum Associates.
- Davis, B., & Sumara, D. (2010). “If things were simple . . .”: complexity in education. In *Journal of Evaluation in Clinical Practice* (Vol. 16, pp. 856–860).
- Davis, T. (2009). Diversity practice in social work: examining theory in practice. *Journal of Ethnic and Cultural Diversity in Social Work*, 18, 40–69.
- De Dreu, C. (2006). When too little or too much hurts: evidence for a curvilinear relationship between task and innovation in teams. *Journal of Management*, 31(1), 83–107.
- Denrell, J. (2008). Superstitious behaviour as a byproduct of intelligent adaptation. In G. Hodgkinson & W. Starbuck (Eds.), *The Oxford Handbook of Organizational Decision Making*. Oxford: Oxford University Press.
- Dewey, J. (1930). *The quest for certainty: A study of the relation of knowledge and action*. London: George Allen & Unwin Ltd.
- Donnelly, J. (2017). A systematic review of concept mapping dissertations. *Evaluation and Program Planning*, 60, 186–193.
- Dozois, E., Langlois, M., & Blanchet-Cohen, N. (2010). DE 201: A practitioner’s guide to developmental evaluation. J.W. McConnell Foundation.
- Drazin, R., Glynn, M., & Kazanjian, R. (1999). Multilevel theorizing about creativity in organizations: a sensemaking perspective. *Academy of Management Review*, 24(2), 286–307.

- Druckman, D., & Wall, J. (2017). A treasure trove of insights: sixty years of JCR research on negotiation and mediation. *Journal of Conflict Resolution*, 61(9), 1898–1924.
- Edwards, N., Rowan, M., Marck, P., & Grinspun, D. (2011). Understanding Whole Systems Change in Health Care: The Case of Nurse Practitioners in Canada. *Policy, Politics, & Nursing Practice*, 12(1), 4–17.
- Eisenhardt, K. (1989). Building theory from case study research. *Academy of Management Review*, 14, 532–550.
- Engeström, Y. (2001). Expansive learning at work: toward an activity theory reconceptualization. *Journal of Education and Work*, 14(1), 133–156.
- Engeström, Y. (2006). *Activity theory and expansive design*. Retrieved from http://projectsfinal.interactionivrea.org/2004-2005/SYMPOSIUM%202005/communication%20material/ACTIVITY%20THEORY%20AND%20EXPANSIVE%20DESIGN_Engestrom.pdf
- Engeström, Y., & Blackler, F. (2005). On the life of the object. *Organization*, 12(3), 307–330.
- Engeström, Y., & Middleton, D. (1996). *Cognition and communication at work*. Cambridge, UK: Cambridge University Press.
- Ensminger, D., Kallemeyn, L., Rempert, T., Wade, J., & Polanin, M. (2015). Case study of an evaluation coaching model: exploring the role of the evaluator. *Evaluation and Program Planning*, 49, 124–136.
- Eoyang, G., & Halladay, R. (2013). *Adaptive action: leveraging uncertainty in your organization*. Palo Alto, CA: Stanford University Press.

- Fam, D. (2017). Facilitating communities of practice as social learning systems: a case study of trialling sustainable sanitation at the University of Technology Sydney. *Knowledge Management Research & Practice*, 15, 391–399.
- Ferlie, E., Crilly, T., Jashapara, A., & Peckham, A. (2012). Knowledge mobilisation in healthcare: A critical review of health sector and generic management literature. *Social Science & Medicine*, 74(8), 1297–1304.
- Finegood, D. T. (2012). The Importance of Systems Thinking to Address Obesity. In A. Drewnowski & B. J. Rolls (Eds.), *Nestlé Nutrition Institute Workshop Series* (Vol. 73, pp. 123–137). Basel: S. KARGER AG. Retrieved from <http://www.karger.com/doi/10.1159/000341308>
- Fleischer, D., & Christie, C. (2009). Evaluation use: results from a survey of US American Evaluation Association members. *American Journal of Evaluation*, 30(2), 158–175.
- Fleischer, J., & Zumeta, Z. (n.d.). Preventing conflict through facilitation: a comparison of mediation and facilitation. Retrieved from <http://collaborateatwork.com/resources/facilitation.php>
- Foreman-Peck, L., & Travers, K. (2015). Developing expertise in managing dialogue in the “third space”: Lessons from a responsive participatory evaluation. *Evaluation*, 21(3), 344–358.
- Foster-Fishman, P., Berkowitz, S., Lounsbury, D., Jacobson, S., & Allen, N. (2001). Building collaborative capacity in community coalitions: a review and integrative framework. *American Journal of Community Psychology*, 29(2), 241–261.
- Gamble, J. (2008). A developmental evaluation primer. J.W. McConnell Foundation.

- Gary, L. (2005). Thought leadership; Ronald Heifetz - The challenge of adaptive leadership. *New Zealand Management*, Aug 2005, 46–48.
- Gates, E., McNall, M., & Williams, B. (2018). Are we part of the problem or solution? Using systems approaches to speak truth(s) to power(s) for better or worse. Presented at the American Evaluation Association Annual Conference, Cleveland, OH.
- George, A., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.
- Gerrits, L., & Verweij, S. (2015). Taking stock of complexity in evaluation: A discussion of three recent publications. *Evaluation*, 21(4), 481–491.
- Gioia, D., & Chittipeddi, K. (1991). Sensemaking and sensegiving in strategic change initiation. *Strategic Management Journal*, 12(6), 433–448.
- Gioia, D., & Thomas, J. (1996). Identity, image and issue interpretation: sensemaking during strategic change in academia. *Administrative Science Quarterly*, 41, 370–403.
- Global Centre for Pluralism. (2017). Intersections: practicing pluralism. Global Centre for Pluralism. Retrieved from <https://www.pluralism.ca/practising-pluralism/>
- Goldman, A., & Kane, M. (2014). Concept mapping and network analysis: an analytic approach to measure ties among constructs. *Evaluation and Program Planning*, 47, 9–17.
- Green, J., Roberts, H., Petticrew, M., Steinbach, R., Goodman, A., Jones, A., & Edwards, P. (2015). Integrating quasi-experimental and inductive designs in evaluation: a case study of the impact of free bus travel on public health. *Evaluation*, 21(4), 391–406.
- Greene, J. C. (2000a). Challenges practicing deliberative democratic evaluation. *New Directions for Evaluation*, 85, 13–26.

Greene, J. C. (2000b). Reconsidering the role of the evaluator in evaluation and society.

Presented at the American Evaluation Association Annual Conference, Honolulu, Hawaii.

Greene, J. C. (2007). *Mixed Methods in Social Inquiry*. San Francisco: Jossey Bass.

Greene, J. C., & Hall, J. (2010). Dialectics and pragmatism. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (pp. 119–142). Thousand Oaks, CA: Sage Publications.

Greenhalgh, T. (2010). What is this knowledge we seek to exchange? *The Milbank Quarterly*, 88(4), 492–499.

Greenhalgh, T., & Malterud, K. (2017). Systematic reviews for policymaking: muddling through. *American Journal of Public Health*, 107(1), 97.

Greenhalgh, T., & Wieringa, S. (2011). Is it time to drop the “knowledge translation” metaphor? A critical literature review. *Journal of the Royal Society of Medicine*, 104, 501–509.

Hall, N. (2013). Pragmatism, evidence, and mixed methods evaluation. *New Directions for Evaluation*, 138, 15–26.

Hargreaves, M. (2018). Leveraging systemic change: evaluating what works. NORC @ the University of Chicago.

Hawe, P. (2015). Lessons from complex interventions to improve health. *Annual Review of Public Health*, 36, 307–323.

Hawe, P., Bond, L., & Butler, H. (2009). Knowledge theories can inform evaluation practice: What can a complexity lens add? *New Directions for Evaluation*, 2009(124), 89–100.

Heaphy, E. (2017). Dancing on hot coals: how emotion work facilitates collective sensemaking. *Academy of Management Journal*, 60(2), 642–670.

- Howard, J., Flores, A., & Hambleton, R. (2015). Evaluation as the co-construction of knowledge: case studies of place-based leadership and public service innovation. In *Advances in Program Evaluation* (Vol. 15, pp. 207–227).
- Jabeen, S. (2016). Do we really care about unintended outcomes? An analysis of evaluation theory and practice. *Evaluation and Program Planning*, 55, 144–154.
- Jain, M., Heyland, D., Dhaliwal, R., Day, A., Drover, J., Keefe, L., & Gelula, M. (2006). Dissemination of the Canadian clinical practice guidelines for nutrition support: Results of a cluster randomized controlled trial. *Critical Care Medicine*, 34(9), 2362–69.
- Jehn K. A., & Bendersky C. (2003). Intragroup conflict in organizations: a contingency perspective on the conflict-outcome relationship. *Research in Organizational Behavior*, 25, 187–242.
- Johnson, R., & Onwuegbuzie, A. (2004). Mixed methods research: a research paradigm whose time has come. *The Educational Researcher*, 33(7), 14–26.
- Jones, N., Suurdt, J., Ouellette-Kuntz, H., & Heyland, D. (2007). Implementation of the Canadian Clinical Practice Guidelines for Nutrition Support: A multiple case study of barriers and enablers. In *Nutrition in Clinical Practice* (Vol. 22, pp. 449–57).
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.
- Kane, M., & Rosas, S. (2017). *Conversations about group concept mapping: applications, examples and enhancements*. Thousand Oaks, CA: Sage Publications.
- Kane, M., & Trochim, W. M. K. (2007). *Concept mapping for planning and analysis*. Thousand Oaks, CA: Sage Publications.

- Kania, J., & Kramer, M. (2011). Collective impact. *Stanford Social Innovation Review*, Winter 2011, 36–41.
- Kemmis, S., & McTaggart. (2008a). Participatory action research: communicative action and the public sphere. In N. Denzin & Y. Lincoln (Eds.), *Strategies of Qualitative Research* (pp. 271–330). Los Angeles: Sage Publications.
- Kemmis, S., & McTaggart. (2008b). Participatory action research: communicative action and the public sphere. In N. Denzin & Y. Lincoln (Eds.), *Strategies of Qualitative Research* (pp. 271–330). Los Angeles: Sage Publications.
- Khazraee, E., & Gasson, S. (2014). Epistemic objects and embedded paradigms. *Academy of Management Proceedings*, 2014(1), 15832.
- King, E. (2017, April). International development, sustainable development goals and pluralism. Global Centre for Pluralism. Retrieved from <https://www.pluralism.ca/intersections-practicing-pluralism/>
- King, J., & Stevahn, L. (2002). Three frameworks for considering evaluator role. In K. Ryan & T. Schwandt (Eds.), *Exploring evaluator role and identity*. Charlotte, NC: Information Age Publishers.
- Kleiboer, M. (1996). Understanding success and failure of international mediation. *Journal of Conflict Resolution*, 40(2), 360–389.
- Kolb, D. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Upper Saddle River, New Jersey: Prentice Hall.
- Kolb, D. (2014). *Experiential Learning: Learning as the Source of Learning and Development* (2nd ed). Upper Saddle River, New Jersey: Pearson Education Ltd.

- Krause, H. (2018). RCTs: not necessarily the gold standard: good alternatives for measuring impact. Presented at the American Evaluation Association Annual Conference, Cleveland, OH.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. London: Sage.
- Krucken, G. (2003). Learning the “new, new thing”: On the role of path dependency in university structures. *Higher Education*, 46, 315–339.
- Kushner, S. (1996). The limits of constructivism in evaluation. *Evaluation*, 2(2), 189–200.
- Kushner, S. (2015a). Chasing the McGuffin: toward the continuing reform of program evaluation. *Evaluation*, 21(2), 204–215.
- Kushner, S. (2015b, July 8). Re: Resources on “theory of change”; Message posted to EVALTALK. Retrieved from <https://listserv.ua.edu/archives/evaltalk.html>
- Lahtinen, T., Guillaume, J., & Hamalainen, R. (2017). Why pay attention to paths in the practice of environmental modelling? *Environmental Modelling & Software*, 92, 74–81.
- Lam, C. Y., & Shulha, L. M. (2015). Insights on using developmental evaluation for innovating: A case study on the cocreation of an innovative program. *American Journal of Evaluation*, 36(3), 147–169.
- Langlois, M., Blanchet-Cohen, N., & Beer, T. (2012). The Art of the Nudge: Five Practices for Developmental Evaluators. *Canadian Journal of Program Evaluation*, 27(2), 39–59.
- Langlois, M, Blanchet-Cohen, N., & Beer, T. (2012). The Art of the Nudge: Five Practices for Developmental Evaluators. *Canadian Journal of Program Evaluation*, 27(2), 39–59.
- Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=90019890&site=ehost-live>

- Lanham, H. J., Leykum, L. K., Taylor, B. S., McCannon, C. J., Lindberg, C., & Lester, R. T. (2013). How complexity science can inform scale-up and spread in health care: Understanding the role of self-organization in variation across local contexts. *Social Science & Medicine*, 93, 194–202.
- Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: University Press.
- Lemke, J. (2000). Across the scales of time: artifacts, activities, and meanings in ecosocial systems. *Mind, Culture and Activity*, 7(4), 273–290.
- Levinthal, D., & Rerup, C. (2006). Crossing an Apparent Chasm: Bridging Mindful and Less-Mindful Perspectives on Organizational Learning. *Organization Science*, 17(4), 502–513.
- Levitt, B., & March, J. G. (1988). Organizational Learning. *Annual Review of Sociology*, 14, 319–340.
- Liedka, J., Salzman, R., & Azar, D. (2017). *Design thinking for the greater good: Innovation in the social sector*. New York, NY: Columbia University Press.
- Lipschitz, R., Friedman, V., & Popper, M. (2007). *Demystifying Organizational Learning*. Thousand Oaks, CA: Sage Publications.
- Maitlis, S., & Christianson, M. (2014). Sensemaking in Organizations: Taking Stock and Moving Forward. *The Academy of Management Annals*, 8(1), 57–125.
- Maitlis, S., & Lawrence, T. (2007). Triggers and enablers of sensegiving in organizations. *Academy of Management Journal*, 50(1), 57–84.
- Maitlis, S., & Sonenschein, S. (2010). Sensemaking in crisis and change: inspiration and insights from Weick (1988). *Journal of Management Studies*, 47(3), 551–580.

- March, J. G., Sproull, L. S., & Tamuz, M. (1991). Learning from Samples of One or Fewer. *Organization Science*, 2(1), 1–13.
- Marchal, B., Westhorp, G., Wong, G., Van Belle, S., Greenhalgh, T., Kegels, G., & Pawson, R. (2013). Realist RCTs of complex interventions - an oxymoron. *Social Science & Medicine*.
- Mark, M. M. (2002). Toward better understanding of alternative evaluator roles. In K. Ryan & Schwandt (Eds.), *Exploring evaluator role identity*. Greenwich, CT: Information Age Publishers.
- Mark, M. M., Henry, G., & Julnes, G. (2000). *Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs*. San Francisco, CA: Jossey-Bass.
- Marx, A., Rihoux, B., & Ragin, C. (2014). The origins, development and application of Qualitative Comparative Analysis: the first 25 years. *European Political Science Review*, 6(1), 115–142.
- Maxwell, J. A. (2013). *Qualitative Research Design: An Interactive Approach* (3rd ed). Los Angeles: Sage Publications.
- McCullum, C., Pelletier, D., Barr, D., Wilkins, J., & Habicht, J. (2004). Mechanisms of power within a community-based food security planning process. *Health Education & Behavior*, 31(2), 206–222.
- McKegg, K., & Wehipeihana, N. (2016). Developmental evaluation in synthesis: practitioners' perspectives. In M. Patton, K. McKegg, & N. Wehipeihana (Eds.), *Developmental evaluation exemplars: principles in practice* (pp. 271–288). New York, NY: The Guilford Press.

- McLinden, D. (2016). And then the internet happened: thoughts on the future of concept-mapping. *Evaluation and Program Planning*, 60, 293–300.
- McMaster, R. (2013). *Sensemaking in Emergency Response Command and Control*. University of Birmingham, UK. Retrieved from <http://etheses.bham.ac.uk/4945/>
- Meindl, J., Stubbart, C., & Porac, J. (1994). Cognition within and between organizations: five key questions. *Organization Science*, 5(3), 289–293.
- Mertens, D. (2015). Philosophical assumptions and program evaluation. *Spaziofilosofico*, 13, 75–86.
- Merton, R. (2012). On sociological theories of the middle range. In C. Calhoun, J. Gerteis, J. Moody, S. Pfaff, & I. Virk (Eds.), *Classical sociological theory*. John Wiley & Sons.
- Michalski, G., & Cousins, J. B. (2000). Differences in stakeholder perceptions about training evaluation: a concept mapping / pattern matching investigation. *Evaluation and Program Planning*, 23, 211–230.
- Midgley, G. (2006). Systemic Intervention for Public Health. *American Journal of Public Health*, 96(3), 466–472.
- Midgley, G. (2008). Systems Thinking, Complexity and the Philosophy of Science. *E:CO*, 10(4), 55–73.
- Midgley, G. (2016, July 7). Co-creation without systems thinking can be dangerous. Integration and Implementation Insights. Retrieved from <https://i2insights.org/2016/07/07/co-creation-and-systems-thinking/>
- Midgley, G., & Lindhult, E. (2017). *What is systemic innovation?* Hull, UK: Hull University Business School.

- Miettinen, R., & Virkkunen, J. (2005). Epistemic objects, artifacts and organizational change. *Organization*, 12(3), 437–456.
- Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed). Los Angeles: Sage Publications.
- Milley, P., Szijarto, B., Svensson, K., & Cousins, J. B. (2018). The evaluation of social innovation: a review and integration of the current empirical knowledge base. *Evaluation*, 24(2), 237–258.
- Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford: Oxford University Press.
- Mohrman, S., Gibson, C., & A. Mohrman. (2001). Doing research that is useful to practice: a model and empirical exploration. *Academy of Management Journal*, 44(2), 357–375.
- Morgan, D. (2007). Paradigms lost and pragmatism regained: methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research*, 1(1), 48–76.
- Mostert, E., Pahl-Wostl, C., Rees, Y., Searle, B., Tabara, D., & Tippet, J. (2007). Social learning in European river-basin management: barriers and fostering mechanisms from 10 river basins. *Ecology and Society*, 12(1).
- Mowles, C. (2014). Complex, but not quite complex enough: The turn to the complexity sciences in evaluation scholarship. *Evaluation*, 20(2), 160–175.
- Mowles, C., Stacey, R., & Griffin, D. (2008). What contribution can insights from the complexity sciences make to the theory and practice of development management? *Journal of International Development*, 20(6), 804–820.

- Muro, M., & Jeffrey, P. (2008). A critical review of the theory and application of social learning in participatory natural resource management processes. *Journal of Environmental Planning and Management*, 51(3), 325–344.
- Murphy, N. (2016). Nine guiding principles to help youth overcome homelessness: a principles-focused developmental evaluation. In *Developmental evaluation exemplars: principles in practice* (pp. 63–82). New York, NY: The Guilford Press.
- Nonaka, I., & Konno, N. (1998). The concept of “ba”: building a foundation for knowledge creation. *California Management Review*, 40(3), 40–54.
- Oborn, E., Barrett, M., & Racko, G. (2012). Knowledge translation in healthcare: incorporating theories of learning and knowledge from the management literature. *Journal of Health Organization and Management*, 27(4), 412–431.
- Obstfeld, D., Borgatti, S., & Davis, J. (2014). Brokerage as a process: decoupling third party action from social network structure. *Contemporary Perspectives on Organizational Social Networks*, 40, 135–159.
- Orlikowski, W. (2002). Knowing in Practice: Enacting a Collective Capability in Distributed Organizing. *Organization Science*, 13(3), 249–273.
- Pandey, S., & Srivastava, S. (2016). Knowledge brokers in service design: lessons from organizational studies. In *Proceedings of the ServDes. 2016 Conference*. Amsterdam: Linköping University Electronic Press.
- Patrizi, P., Thompson, E., Coffman, J., & Beer, T. (2013). Eyes wide open: learning as strategy under conditions of complexity and uncertainty. *Foundation Review*, 5(3), 50–65.
- Patton, M. (1994). Developmental evaluation. *Evaluation Practice*, 15(3), 311–319.
- Patton, M. (2007). Process use as usefulism. *New Directions for Evaluation*, 116, 99–112.

- Patton, M. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: The Guilford Press.
- Patton, M. (2016a). State of the art and practice of developmental evaluation: answers to some common and recurring questions. In M. Patton, K. McKegg, & N. Wehipeihana (Eds.), *Developmental Evaluation Exemplars: Principles in Practice* (pp. 1–24). New York, NY: The Guilford Press.
- Patton, M. (2016b). The developmental evaluation mindset: eight guiding principles. In *Developmental Evaluation Exemplars: Principles in Practice* (pp. 289–312). New York, NY: The Guilford Press.
- Patton, M. (2018). *Principles-focused evaluation: the GUIDE*. New York, NY: The Guilford Press.
- Patton, M., McKegg, K., & Wehipeihana, N. (Eds.). (2016). *Developmental Evaluation Exemplars: Principles in Practice*. New York, NY: The Guilford Press.
- Pawson, R., & Tilley, N. (1997). *Realistic Evaluation*. Thousand Oaks, CA: Sage Publications.
- Pawson, R., & Tilley, N. (2004). *Realist evaluation*. DPRN thematic meeting 2006 report on evaluation. Retrieved from http://www.communitymatters.com.au/RE_chapter.pdf
- Pawson, R., Wong, G., & Owen, L. (2011). Known Knowns, Known Unknowns, Unknown Unknowns: The Predicament of Evidence-Based Policy. *American Journal of Evaluation*, 32(4), 518–546. <https://doi.org/10.1177/1098214011403831>
- Pelling, M., High, C., Dearing, J., & Smith, D. (2008). Shadow spaces for social learning: a relational understanding of adaptive capacity to climate change within organizations. *Environment and Planning*, 40(4), 867–884.

- Petrucci, C., & Quinlan, K. (2007). Bridging the research-practice gap: concept mapping as a mixed-methods strategy in practice-based research and evaluation. *Journal of Social Service Research*, 34(2), 25–42.
- Peurach, D., Glazer, J., & Winchell Lenhoff, S. (2016). The developmental evaluation of school improvement networks. *Educational Policy*, 30(4), 606–648.
- Poggi, G. (1965). A main theme of contemporary sociological analysis: Its achievements and limitations. *The British Journal of Sociology*, 16(4), 283–294.
- Poth, C., Pinto, D., & Howery, K. (2012). Addressing the challenges encountered during a developmental evaluation: Implications for evaluation practice. *Canadian Journal of Program Evaluation*, 26(1), 39–48.
- Preskill, H., & Beer, T. (2012). Evaluating social innovations. FSG. Retrieved from <https://www.fsg.org/publications/evaluating-social-innovation>
- Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation*, 29(4), 443–459.
- Preskill, H., Gutierrez, E., & Mack, K. (2017). Facilitating intentional group learning. FSG. Retrieved from <https://www.fsg.org/tools-and-resources/facilitating-intentional-group-learning>
- Preskill, H., & Torres, R. (1999). Building capacity for organizational learning through evaluative inquiry. *Evaluation*, 5(1), 42–60.
- Ragin, C. (2009). Reflections on casing and case-oriented research. In D. Byrne & C. Ragin (Eds.), *The Sage Handbook of Case-based Methods* (pp. 522–535). London: Sage Publications.

- Ray, T., & Little, S. (2001). Communication and context: collective tacit knowledge and practice in Japan's workplace. *Creativity and Innovation Management*, 10(3), 154–164.
- Rezania, D., & Blyth, L. (2009). Making Sense of Experiential Learning in Management Education. In P. P. Daly & D. D. Gijbels (Eds.), *Real Learning Opportunities at Business School and Beyond* (pp. 147–162). Retrieved from http://link.springer.com/chapter/10.1007/978-90-481-2973-7_10
- Richardson, K. (2004). Systems theory and complexity: Part 1. *Emergence: Complexity and Organization*, 6(3), 75–79.
- Rifkin, J., Millen, J., & Cobb, S. (1991). Toward a new discourse for mediation: a critique of neutrality. *Mediation Quarterly*, 9(2), 151–164.
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2), 155–169. <https://doi.org/10.1007/BF01405730>
- Roberge, M.-É., & van Dick, R. (2010). Recognizing the benefits of diversity: When and how does diversity increase group performance? *Human Resource Management Review*, 20(4), 295–308.
- Rogers, E., Medina, U., Rivera, M., & Cody, J. (2005). Complex adaptive systems and the diffusion of innovations. *The Innovation Journal: The Public Sector Innovation Journal*, 10(3), 1–26.
- Rorty, R. (1982). *Consequences of pragmatism*. Minneapolis, MN: University of Minnesota Press.
- Rosas, S., & Kane, M. (2012). Quality and rigor of the concept mapping methodology: A pooled study analysis. *Evaluation and Program Planning*, 35, 236–245.

- Rudolph, J. W., & Morrison, J. B. (2008). Sidestepping superstitious learning, ambiguity, and other roadblocks: a feedback model of diagnostic problem solving. *The American Journal of Medicine*, 121(5), 34–37.
- Ryan, K., & Schwandt, T. (Eds.). (2002). *Exploring evaluator role and identity*. Greenwich, CT: Information Age Publishers.
- Saari, K., & Kallio, K. (2011). Developmental impact evaluation for facilitating learning in innovation networks. *American Journal of Evaluation*, 31(2), 227–245.
- Salthe, S. (1993). *Development and evolution: complexity and change in biology*. Cambridge, MA: MIT Press.
- Sandberg, J., & Tsoukas, H. (2015). Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. *Journal of Organizational Behavior*, 36, 6–32.
- Schalock, R., Verdugo, M., & van Loon, J. (2018). Understanding organization transformation in evaluation and program planning. *Evaluation and Program Planning*, 67, 53–60.
- Segerholm, C. (2002). Evaluation as responsibility, conscience and conviction. In K. Ryan & T. Schwandt (Eds.), *Exploring evaluator role and identity* (pp. 87–102). Greenwich, CT: Information Age Publishers.
- Senge, P. (1990). *The Fifth Discipline*. New York, NY: Doubleday.
- Shulha, L. M., Whitmore, E., Cousins, J. B., Gilbert, N., & al Hudib, H. (2016). Introducing evidence-based principles to guide collaborative approaches to evaluation: results of an empirical process. *American Journal of Evaluation*, 37(2), 193–215.
- Skolits, G., Morrow, J., & Mehalic Burr, E. (2009). Reconceptualizing evaluator roles. *American Journal of Evaluation*, 30(3), 275–295.

- Sloman, S., & Fernbach, P. (2018). *The Knowledge Illusion: Why We Never Think Alone*. New York, NY: Penguin.
- Smith, J., & Smith, L. (2017). The nature of creativity: mayflies, octopi, and the best bad idea we have. In B. Beghetto & B. Sriraman (Eds.), *Creative contradictions in education: cross-disciplinary paradoxes and perspectives* (pp. 21–35). Switzerland: Springer.
- Snowden, D. (2011). Naturalizing sensemaking. In K. Mosier & U. Fischer (Eds.), *Informed Knowledge: Expert Performance in Complex Situations*. New York, NY: Taylor & Francis Group.
- Snowden, D. (n.d.). Naturalizing sensemaking. In K. Mosier & U. Fischer (Eds.), *Informed by Knowledge: Expert Performance in Complex Situations*. New York, NY: Taylor & Francis Group.
- Snowden, D., & Boone, M. E. (2007). A leader's framework for decision making. *Harvard Business Review*, 85(11).
- Social Programs that Work. (2019). Nurse-Family Partnership. Retrieved January 4, 2019, from <https://evidencebasedprograms.org/programs/nurse-family-partnership/>
- Stacey, R. (2001). *Complex Responsive Processes in Organizations: Learning and Knowledge Creation*. London: Taylor & Francis Group.
- Stacey, R. (2005). *A Complexity Perspective on Researching Organizations: Taking Experience Seriously*. London: Routledge.
- Sterman, J. D. (2006). Learning from Evidence in a Complex World. *American Journal of Public Health*, 96(3), 505–514.
- Stevahn, L., & King, J. (2005a). Establishing essential competencies for evaluators. *American Journal of Evaluation*, 26(1), 43–59.

- Stevahn, L., & King, J. (2005b). Managing conflict constructively in program evaluation. *Evaluation, 11*(4), 415–427.
- Stevahn, L., King, J., Ghore, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation, 26*(1), 43–59.
- Stevenson, W., Zinzow, H., & Sridharan, S. (2003). Using event structure analysis to understand planned social change. *International Journal of Qualitative Methods, 2*(2), 1–18.
- Stewart, D., Shamdasani, P., & Rook, D. (2007). *Focus groups: theory and practice*. Thousand Oaks, CA: Sage Publications.
- Stigliani, I., & Ravasi, D. (2012). Organizing Thoughts and Connecting Brains: Material Practices and the Transition from Individual to Group-Level Prospective Sensemaking. *Academy of Management Journal, 55*(5), 1232–1259.
- Strike, V., & Rerup, C. (2016). Mediated Sensemaking. *Academy of Management Journal, 59*(3), 880–905. <https://doi.org/amj.2012.0665>
- Strubing, J. (2007). Research as pragmatic problem-solving: The pragmatist roots of empirically-grounded theorizing. In *The Sage handbook of grounded theory* (pp. 580–602).
- Tannahill, C., & Sridharan, S. (2013). Getting real about policy and practice needs: evaluation as a bridge between the problem and solution space. *Evaluation and Program Planning, 36*, 157–164.
- The World Bank. (2003, June). Scaling-up the impact of good practices in rural development, Report Number: 26031. The International Bank for Reconstruction and Development.
- Thomas, G. (2011). A typology for the case study in social science following a review of definition, discourse, and structure. *Qualitative Inquiry, 17*(6), 511–521.

- Togni, S., Askew, D., Rogers, L., Potter, N., Egert, S., Hayman, N., ... Brown, A. (2016). Creating safety to explore: strengthening innovation in an Australian Indigenous primary health care setting through developmental evaluation. In *Developmental evaluation exemplars: Principles in practice* (pp. 234–251). New York, NY: The Guilford Press.
- Trochim, W. M. K. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning*, 1989(12), 1–16.
- Trochim, W. M. K. (1993). The reliability of concept mapping. Presented at the American Evaluation Association Conference, Dallas, TX. Retrieved from <http://www.socialresearchmethods.net/research/Reliable/reliable.htm>
- Trochim, W. M. K. (2017). Hindsight is 20/20: reflections on the evolution of concept mapping. *Evaluation and Program Planning*, 60, 176–185.
- Trumpower, D., Filiz, M., & Sawar, G. (2014). Assessment for learning using digital knowledge maps. In D. Ifenthaler & R. Hanewald (Eds.), *Digital Knowledge Maps in Education* (pp. 221–237). New York, NY: Springer.
- Tsoukas, H. (2009). A Dialogical Approach to the Creation of New Knowledge in Organizations. *Organization Science*, 20(6), 941–957.
- Tyre, M., & Orlikowski, W. (1994). Windows of opportunity: temporal patterns of technological adaptation in organizations. *Organization Science*, 5(1), 98–118.
- University of Colorado Boulder. (2019). Nurse-Family Partnership. Retrieved December 21, 2018, from <https://www.blueprintsprograms.org/factsheet/nurse-family-partnership>
- Urban, J., Hargraves, M., & Trochim, W. M. K. (2014). Evolutionary evaluation: implications for evaluators, researchers, practitioners, funders and the evidence-based program mandate. *Evaluation and Program Planning*, 45, 127–139.

- Urry, J. (2005). The complexity turn. *Theory, Culture & Society*, 22(5).
- Van de Ven, A., & Johnson, P. (2006). Knowledge for theory and practice. *Academy of Management Review*, 31(4), 802–821.
- Vangen, S., & Huxham, C. (2003). Enacting leadership for collaborative advantage: dilemmas of ideology and pragmatism in the activities of partnership management. *British Journal of Management*, 14, S61–S76.
- Vogus, T., Rothman, N., Sutcliffe, K., & Weick, K. (2014). The affective foundations of high-reliability organizing. *Journal of Organizational Behavior*, 35, 592–596.
- Wall, J., Stark, J., & Standifer, R. (2001). Mediation: a current review and theory development. *Journal of Conflict Resolution*, 45(3), 370–391.
- Ward, V., Smith, S., House, A., & Hamer, S. (2012). Exploring knowledge exchange: a useful framework for practice and policy. *Social Science & Medicine*, 74, 297–304.
- Wathan, N., Wyatt, S., & Harris, R. (2008). *Mediating health information: The go-betweens in a changing socio-technical landscape*. New York, NY: Springer.
- Wehipeihana, N., McKegg, K., Thompson, V., & Pipi, K. (2016). Cultural responsiveness through developmental evaluation: Indigenous innovations in sport and traditional Māori recreation. In *Developmental evaluation exemplars: principles in practice* (pp. 25–44). New York, NY: The Guilford Press.
- Weibe, E. (2010). Temporal sensemaking: Managers' use of time to frame organizational change. In T. Hernes & S. Maitlis (Eds.), *Process, sensemaking, & organizing* (pp. 213–241). Oxford: Oxford University Press.
- Weick, K. (2006). Faith, evidence and action: better guesses in an unknowable world. *Organization Studies*, 27(11), 1723–1736.

- Weick, K. (2009). *Making Sense of the Organization*. Chichester: West Sussex: John Wiley & Sons Ltd.
- Weick, K., & Sutcliffe, K. (2006). Mindfulness and the Quality of Organizational Attention. *Organization Science*, 17(4), 514–524.
- Weick, K., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science*, 16(4), 409–421.
- Weiss, C. (1995). Nothing as practical as a good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. Connell (Ed.), *New Approaches to Evaluating Community Initiatives. Concepts, Methods, and Contexts*. Queenstown, MD: Aspen Institute.
- Weiss, C., Murphy-Graham, E., Petrosino, A., & Ghandi, A. (2008). The fairy godmother - and her warts: making the dream of evidence-based policy come true. *American Journal of Evaluation*, 29(1), 29-47.
- Weiss, R. S., & Rein, M. (1969). The evaluation of broad-aim programs: a cautionary case and a moral. *The ANNALS of the American Academy of Political and Social Science*, 385(1), 133–142. <https://doi.org/10.1177/000271626938500112>
- Wenger, E. (2010). Communities of practice and social learning systems: the career of a concept. In *Social Learning Systems and Communities of Practice*. London: Springer Verlag.
Retrieved from trayner.com/resources/publications/cops-and-learning-systems/
- Westhorp, G. (2013). Developing complexity-consistent theory in a realist investigation. *Evaluation*, 19(4), 364–382.
- Williams, B. (2015). Prosaic or Profound? The Adoption of Systems Ideas by Impact Evaluation. *IDS Bulletin*, 46(1), 7–16.

Williams, B., & Van t'Hof, S. (2016). *Wicked Solutions: A Systems Approach to Complex Problems* (2nd ed). Bob Williams. Retrieved from <http://bit.ly/1SVoOH3>

Wolf-Branigin, M. (2009). Applying complexity and emergence in social work education. *Social Work Education*, 28(2), 115–127.

Wolf-Branigin, Michael. (2009). Applying complexity and emergence in social work education. *Social Work Education*, 28(2), 115–127.

Zollo, M. (2009). Superstitious Learning with Rare Strategic Decisions: Theory and Evidence from Corporate Acquisitions. *Organization Science*, 20(5), 894–908.

Appendix A. Content Analysis Template, Strand 1

Adapted from Clarke (2003).

Individual actors (key individuals, significant people):
Nonhuman actors (e.g., infrastructure, technology, information):
Collective elements (e.g., organizations, groups):
Silent actors:
Constructs, ideas, common assumptions (“discursive constructions”); major debates; expectations of actors:
Political / economic influences:
Sociocultural influences:
Temporal influences (e.g., seasonality, history, trajectory):
Spatial influences (e.g., geography, locations):

Appendix B. Interview Guide Template

Introduction

First, I'd like to thank you for taking time to participate in this interview. You'll have seen on the consent form that your information will be kept confidential, and that, with your permission I'd like to audio record the interview to be sure that I don't miss anything. Are you comfortable proceeding? Do you have any questions before we begin?

For most of the interview, I'd like to ask you about your experiences with developmental evaluation, with reference to the evaluation work for [intervention]. But first I'd like to ask a very broad question. This is more of a brainstorming question and I'm hoping you'll feel free to offer as many ideas as come to mind. [Relate to study description and concept mapping].

Concept Mapping Prompt

1. Based on your experience, what are the most important things a developmental evaluator does?

Case: When? (e.g. pacing, transitions)

1. I understand from [case documents] that this evaluation began a year after the [intervention] was launched, is that correct?
 - The DE continued for 4 years and then was scaled back, is that right? In a broad sense, how did the DE unfold for you over those years? *How did you get started? What was it like at the beginning? Did it follow a particular set of activities the entire time, or how did it change over time?*
2. Can you tell me about an experience with this DE that was particularly significant? What happened? *When did it happen?*
3. Can you tell me about a time when you thought the DE was going really well? What happened?
4. Was there a time when you encountered a major challenge with DE? What happened then? *When? Why?*
5. Do you remember any turning points, where things suddenly changed? What happened? *When? Why?*

Who/what? (actors, tools) & How? (methods, mechanisms) / prompts for Where (spaces, channels)

6. You met with the external evaluator/internal team multiple times a year, I think? What were those meetings like typically?
7. Can you remember any particular meeting? What stood out for you about this one?

8. The external evaluator is / you are described in the published case as a 'coach'. How did they/you coach?
9. I understand there was an internal team for the [intervention] that was closely involved in the evaluation. Why was this formed? Were they all from [orgs]? How many were on the team? How was the team structured?
10. The internal team held what were called 'sensemaking and analysis' sessions. Why did you have them? How often? What were the sessions typically like?
 - Were there constraints on your ability to achieve what you wanted?
11. Is there a particular session you remember? What happened?
12. [For program participant]: In the published case, you're described as [role] for the DE. What did that mean?
 - What kinds of things did you do in this role? Why? (and when, where, what were the opportunities or boundaries on what they did or could do?)
13. Was there a particular situation you remember when you needed to act in this way?
14. Another interesting idea in the published case is what was called the 'unobtrusiveness' of the DE process, I think it said something like evaluating without distracting people from the work of running the [intervention]. Does that resonate? What does that mean? How was that done (ask for example)?
15. Are there other characteristics of DE that are important like that? What would someone miss about DE if they only looked at the reports? *[prompt also on the descriptions in documentation about space and breathing room for the team to explore]*?
16. In the published case, it says that the DE contributed to development of 'a more adaptive team culture of evaluation and learning that affected all our work'. From your perspective, what does it mean to have an adaptive team culture of evaluation and learning? (Prompt, how do you think that came about?).
17. DE was described in your case as 'a mechanism for managing risk'. What does that mean in your view? How is that important in DE? *[prompt also on the descriptions in the text about making space and breathing room for the team to explore]*?
 - Can you describe a specific example when risk management became important? What happened? How was it done?
18. In the chapter there's a statement [quote from chapter]. What does that mean to you in the context of this experience?
 - Can you think of a specific example?
 - i. What role did the theory of change play? It mentions that the theory of change was important point of focus for all the stakeholders? Why?
 - ii. The theory of change was diagrammed – how were the diagrams used?
 - iii. How did this relate to the 'contribution story' mentioned in the chapter?
19. Is there anything else important about this experience that we didn't cover?

Other possible questions (time permitting):

1. The chapter also mentions prototypes. Can you provide an example? How were these used? Why?
2. Did you learn from the DE process? What (kinds of things) did you learn?

3. Was there anything that got in the way of learning from the process or applying that learning? Was there a time when something learned was 'lost'? Why?
4. Throughout the DE period, you were also actively gathering input from outside (expert panel review, external case studies). Why? What was the intent? *How did you / did you find it easy to integrate the external information?*

Appendix C. Letter of Informed Consent: Interviews

Title of the study: Mediated Sensemaking: Adaptive Learning with Social Interventions.

About this form: As someone with extended experience using developmental evaluation, your input is important and much appreciated. Thank you for your willingness to take part. As you may have questions about the study and your participation, additional information is provided below, including how to contact me, my supervisor and the university.

What is the study's purpose? This study seeks to improve understanding of the role of mediators or facilitators in learning for the development of social interventions. The explicit focus will be on the role of developmental evaluators as examples of mediators. The study will attempt to shed light on the process and how it may influence 'adaptive learning' in collaborative groups.

What does participation involve? Your involvement will consist of participating in an independent, private interview about your experiences with developmental evaluation. The interview will last approximately 1 – 1.5 hours and will be held by telephone or Skype at a time convenient to you. With your permission, interviews will be audio-recorded to ensure the accuracy of the information collected. A case summary to which the interview data will contribute will be provided to you for your review. A summary transcript of the interview will be made available to you on request. Please be aware that security of a transcript sent by email cannot be guaranteed.

Your participation is voluntary and you may withdraw from the study at any time and/or refuse to answer any questions, without any consequences to you. If you decide to withdraw, all data gathered until the time of withdrawal will be destroyed.

What are the risks and benefits of participating? There are not likely to be any risks from participating. I hope that this study will contribute to the development of DE and related learning-focused approaches to evaluation.

How will my privacy be protected? All information gathered in the interview will remain strictly confidential and will be accessible only to me and my thesis supervisor, Dr. J. Bradley Cousins. However, it should be noted that the book "Developmental Evaluation Exemplars: Principles in Practice, edited by M.Q. Patton, K. McKegg and N. Wehipeihana would be referenced in reports of this research and it would be noted in the methods section of any reports that authors of some of the chapters (3-4) from this book had been interviewed. Also, please note that more than one author of [author's article] may be interviewed; a coauthor may be able to identify your comments in the case summary, even if these comments are de-identified, as a result of your previous close collaborative work.

Confidentiality and conservation of data: The information gathered will be securely stored for five years in electronic form on password-protected computers and servers at the University of Ottawa.

Contact information:

Principle Investigator: Ms Barbara Szijarto, PhD Candidate, University of Ottawa.

Thesis Supervisor: Dr. J. Bradley Cousins, Centre for Research on Educational and Community Services, University of Ottawa, 136 Jean Jacques Lussier, Vanier Rm 5002, Ottawa, ON, K1N 6N5.

Acceptance:

☐ I, _____ [*Name of participant*], agree to participate in interviews, as described above.

If you have any questions about the study, you may contact me or Dr. Cousins.

If I have any questions about the ethical conduct of this study, you may contact the Office of Research Ethics and Integrity, University of Ottawa, Tabaret Hall, 550 Cumberland Street, Room 154, Ottawa, ON K1N 6N5

Tel.: [613] 562-5387, Email: ethics@uottawa.ca

There are two copies of the consent form, one of which is yours to keep.

Participant's name	Signature:	Date:
Researcher's name	Signature:	Date:

Sincerely,

Barbara Szijarto
PhD Candidate
University of Ottawa

Appendix D. Letter of Informed Consent: GCM Sorting and Rating

Project title: Mediated Sensemaking: Adaptive Learning with Social Interventions

Invitation to Participate: You are invited to participate in a research study on developmental evaluation led by Ms Barbara Szijarto. Your participation in this research is voluntary; thank you for considering this invitation.

Purpose of the Study: This study seeks to improve understanding of the role of sensemaking mediators or facilitators in learning for the development of social interventions. The explicit focus will be on the role of developmental evaluators as examples of mediators. The study will attempt to shed light on the process and how it may influence ‘adaptive learning’ in collaborative groups.

Participation: Your involvement will consist of sorting a set of statements about the role of the developmental evaluator, assigning labels to the groups you create, and rating each statement in terms of its importance. You would have at least 2 weeks to do these activities, and can do them on your own at a time convenient to you. The activities are anticipated to take about 1 hour altogether.

Potential risks and benefits: There are not likely to be any risks from participating. I hope that this study will contribute to the development of DE and related learning-focused approaches to evaluation.

Privacy of participants: Your participation and the results of your sorting and rating will remain strictly confidential and will be accessible only to me and my thesis supervisor, Dr. J. Bradley Cousins. All findings from this phase of the study will be reported in aggregate form.

Confidentiality and conservation of data: The information gathered will be securely stored for five years in electronic form on password-protected computers and servers at the University of Ottawa.

Voluntary Participation: You are under no obligation to participate in any of the research activities and if you choose to participate, you may withdraw from the study at any time and/or refuse to answer any questions, without suffering any negative consequences. If you choose to withdraw, all data gathered until the time of withdrawal will be destroyed.

Acceptance:

☐ I, _____ [*Name of participant*], agree to participate, as described above.

If you have any questions about the study, you may contact me or Dr. Cousins.

If I have any questions about the ethical conduct of this study, you may contact the Office of Research Ethics and Integrity, University of Ottawa, Tabaret Hall, 550 Cumberland Street, Room 154, Ottawa, ON K1N 6N5, Tel.: [613] 562-5387, Email: ethics@uottawa.ca

Participant's name	Signature:	Date:
--------------------	------------	-------

Researcher's name	Signature:	Date:
-------------------	------------	-------

Sincerely,

Barbara Szijarto
PhD Candidate
University of Ottawa

Appendix E. Letter of Informed Consent: GCM Expert Panel

Project title: Mediated Sensemaking: Adaptive Learning with Social Interventions

Invitation to Participate: You are invited to participate in a research study on developmental evaluation led by Ms Barbara Szijarto. Your participation in this research is voluntary; thank you for considering this invitation.

Purpose of the Study: This study seeks to improve understanding of the role of mediators or facilitators in learning for the development of social interventions. The explicit focus will be on the role of developmental evaluators as examples of mediators. The study will attempt to shed light on the process and how it may influence ‘adaptive learning’ in collaborative groups.

Participation: Your involvement will consist of participating in an expert panel to discuss the results of the concept mapping component of the study. This will be a virtual discussion with 4-5 other developmental evaluators using the GoToMeeting platform (<https://www.gotomeeting.com>). The discussion is expected to last 45 minutes.

Potential risks and benefits: There are not likely to be any risks from participating. I hope that this study will contribute to the development of DE and related learning-focused approaches to evaluation.

Privacy of participants: Participants in the panel will be asked to respect the confidentiality of all members of the group however we cannot guarantee that all members of the panel will honour this request. The data gathered through the panel discussion will be kept confidential. No panel participants will be identified in reports arising from this research; all findings from this phase of the study will be reported in aggregate form only.

Confidentiality and conservation of data: The information gathered will be securely stored for five years in electronic form on password-protected computers and servers at the University of Ottawa.

Voluntary Participation: You are under no obligation to participate in any of the research activities and if you choose to participate, you may withdraw from the study at any time and/or refuse to answer any questions, without suffering any negative consequences. If you choose to withdraw, all data gathered until the time of withdrawal will be destroyed.

Acceptance:

☐ I, _____ [*Name of participant*], agree to participate in the virtual panel discussion, as described above.

If you have any questions about the study, please contact me, or you may also contact Dr. Cousins.

Mediation Social Change

If you have any questions about the ethical conduct of this study, you may contact the Office of Research Ethics and Integrity, University of Ottawa, Tabaret Hall, 550 Cumberland Street, Room 154, Ottawa, ON K1N 6N5. Tel.: [613] 562-5387. Email: ethics@uottawa.ca.

There are two copies of the consent form, one of which is yours to keep.

Participant's name	Signature:	Date:
--------------------	------------	-------

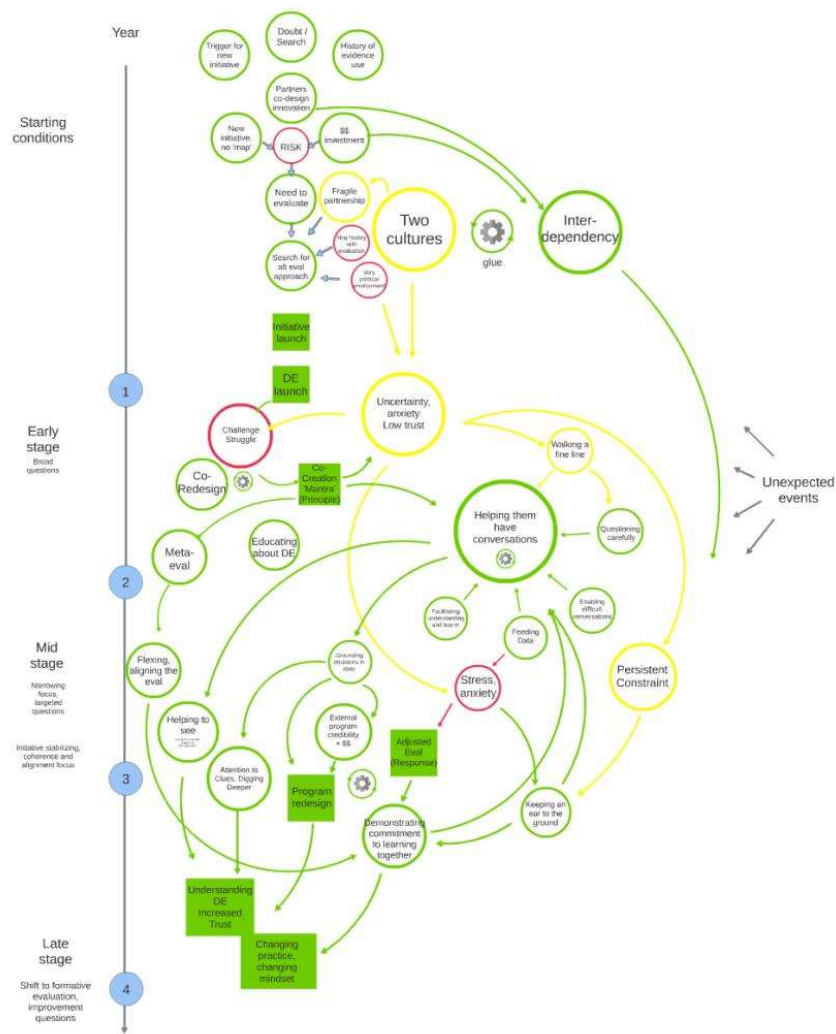
Researcher's name	Signature:	Date:
-------------------	------------	-------

Sincerely,

Barbara

Barbara Szijarto
PhD Candidate
University of Ottawa

Appendix F. Case Trajectory Diagram, Case 2



Appendix G. GCM Statement List by Cluster

Cluster 1

1. Ensure DE will be a good fit for the purpose of the evaluation.
51. Ensure the people involved are ready / really want to do DE.
76. Be clear about what kind of evaluation they are doing. Not drifting back to formative or summative without intending to.
26. Be a good strategy coach.*
77. Help actors make evaluation part of their work.
48. Co-create with the collaborators.
35. Iterate the evaluation design.
72. Find the relevance between the project being evaluated, its relation to the organization, and to the broader community.*
75. Create some kind of framework to start with, pulled out of what the project says they're doing.

Cluster 2

4. Develop good trusting relationships with people around the table.
41. Have good judgement.
24. Create the right environment for different groups (e.g. members of the community) to contribute to the evaluation.
33. Have robust people skills.
22. Have cultural competence to bring to the data collection.
78. Employ robust analytical skills.*
50. Have robust reporting skills.*
17. Be a credible instrument in the evaluation process.
19. Have a deep understanding of the work by being integrated in the work.
79. Collaborate with clients on an ongoing basis.
47. Engage with collaborators in multiple ways.*

Cluster 3

3. Figure out the sources of tensions in the group.
36. Mediate power dynamics.
28. Be really aware of power dynamics.
12. Shift how people connect with one another.
70. Offer actors in a collaboration a new way to operate with each other.
9. Enable the creativity of participants to emerge.
44. Be comfortable in a place of discomfort.

- 25. Help move the group through when there's discomfort.
- 16. Employ robust facilitation skills.
- 6. Know when not to intervene too much with their ideas, and when to intervene because people are getting stuck.
- 37. Have honest and challenging conversations.
- 29. Surface tough issues.
- 11. Be really careful about how things are communicated, to help stakeholders have a productive conversation.

Cluster 4

- 13. Listen and pay attention to what is going on.
- 67. Help to illuminate insight of people around the table.
- 61. Help people to recognize sources of tensions in the group (for example, differing theories of change).
- 15. Enable conversations that explore ambiguity and push for new insights, in a space separate from a focus on closure or decisions.
- 62. Ask good questions.
- 63. Know the right questions to ask at the right time.
- 69. Facilitate conversation through questions.
- 54. Help actors to see their work differently.
- 27. Help actors to be open minded when thinking about how to work toward intended outcomes.

Cluster 5

- 32. Take time to get really clear on the focus of the program and how it all fits together; how do the actors think change happens?*
- 59. Support opportunities for rapid experimentation.
- 21. Bring more rigour to observations.
- 49. Help to navigate the path forward.
- 65. Frame up emerging working models of the intervention.
- 74. Help to course correct.
- 40. Help to hone the strategy.
- 81. Describe the rationale or evidence for key decisions in the development of the intervention.
- 2. Point out where decisions are being made without (or despite) considering key data.
- 88. Support the group to iterate on the intervention.
- 58. Identify major forks in the road or key developments.
- 73. Detail the initiative's evolving theory of change for critical review.
- 39. Remind stakeholders where they are going when they get caught up in day to day operations; of things they need to pay attention to.

- 87. Clearly articulate what is being developed.
- 34. Hold any kind of structure loosely (for example, the program objectives) knowing it may not actually be what's happening.
- 53. Document the evolution of the intervention (what has happened over time, what has been achieved).
- 52. Nudge and challenge stakeholders' thinking where that needs to happen.
- 90. Help the team to view the initiative through a new frame.
- 55. Help to build a culture of curiosity and inquiry about their work.
- 57. Help actors keep in touch with what is happening on the ground.

Cluster 6

- 5. Provide rich feedback from the field.
- 10. Use evaluation tools to bring more certainty to stakeholder hypotheses; walk on more solid ground.
- 68. Use data to help actors understand what the progress is at different levels.
- 8. Keep everybody thinking: "are we getting to where we said we were going", "is the model shaping up the way we thought?", "what are the things we're going to change?".
- 85. Gather data from multiple sources on an ongoing basis.
- 86. Evolve the group's understanding of the problem (e.g. what is poverty?).
- 80. Bring data to the table to support conversations.
- 60. Encourage openness to new sources of data and perspective.
- 45. Generate real time feedback.
- 7. Illuminate assumptions using data and evaluation tools.

Cluster 7

- 84. Highlight patterns of success and failure.
- 43. Consolidate and make sense of the data; what it means to the organization for implementation.
- 14. Allow actors to see the unexpected.
- 30. Help the team to spot new patterns.
- 23. Facilitate conversations around data.
- 18. Facilitate participation in analysis of data.
- 38. Support collaborative meaning making or sensemaking.
- 46. Provide a systematic framework for collective learning.
- 83. Provide a systematic structure for reflection.
- 56. Facilitate regular reflection.
- 31. Help surface what the findings are pointing to.
- 42. Present information to inspire dialogue.
- 89. Help the group to ask good questions.
- 66. Develop evaluative thinking.

- | |
|--|
| <ul style="list-style-type: none">20. Use evaluation as a vehicle for learning.71. Remind actors if they've been down a certain road before and what they learned; help them avoid going in circles.64. Bring a different perspective on the work.82. Bring together outsider and insider perspectives. |
|--|

* “Bridging statements”: statements with bridging values above 8.0.

Appendix H. Focus Group Guide

1. **Greetings & introductions** (10 minutes)
 - Prompt for questions, about process, informed consent, other? Confirm permission to audio record. Reminder to respect confidentiality of other group members. Instructions for use of the chat box, what to do in the event of audio issues, use of toll free line option, etc.
 - Background on GCM and first 2 steps in the process.
2. **Quick tour** of the map (2 minutes)
3. **Broad questions** (5 minutes)
 - What strikes you off the top? What's your initial impression?
 - Overall, does this resonate with your experience? Does anything surprise you? Is there anything that doesn't make sense?
4. **Cluster by cluster** (15 minutes)
 - For each cluster: What do you think this cluster is about? Are there important subgroups within the cluster? What would you name it?
5. **Whole map** (15 minutes).
 - Think about the important things that you do in your work. Is there anything that's missing here?
 - How do the parts of the map interconnect?
 - Are there regions on the map where more than one cluster seems to fit together, or any larger themes in parts of the map?
 - What are the implications of what we've discussed so far? For DE practice? For learning and adaptation in programs?
 - If you were to make changes to the map, what would you change? Why?
6. **Closing.** Solicit questions. Provide information about next steps, contact information. Thank you.

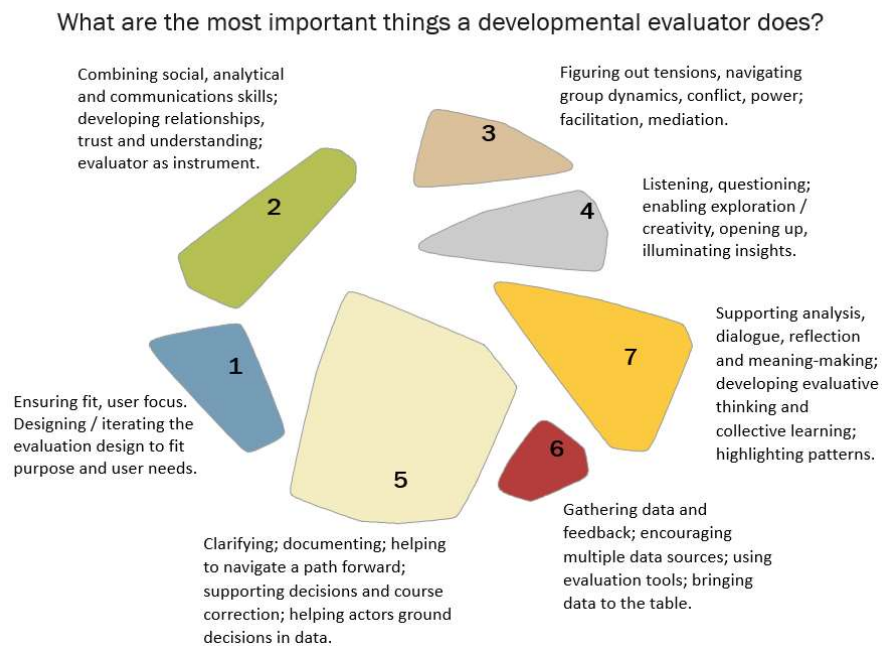


Figure H.1. Initial map provided for interpretation

Appendix I. Full Case Study Reports

Case 1

This case summary is organized in rough chronological form corresponding to broad stages of a four-year evaluation period: starting conditions, structuring the evaluation, mid stage and late stage. Themes derived through initial coding and analysis are briefly elaborated with each ‘stage’ to illustrate the progression of the evaluation, however most themes are relevant to across the whole case.

1. Starting conditions

The initiative is aimed at fostering learning and knowledge sharing among youth education service agencies by developing a community of practice in a large North American urban area. A non-governmental organization leads the initiative and commissioned the evaluation when the initiative was about 1 year old. The organization is large and well established, and expresses interest in fostering sustained, system-level community change. It acts primarily as a funder and capacity builder, rather than in direct program delivery (d, d-v).

Problem and opportunity

Based on research and its own experience in the sector, the organization considered client outcomes to be adversely affected (at least in part) by fragmentation among agencies in the sector and barriers for practitioners to access information (1d0). A large donation triggered the initiative by allowing the organization ‘to go beyond funding individual programs’ (c12, 1d0). The idea was supported by champions on the board of directors even though the organization had not tried this type of intervention before (d, c12).

Norms related to learning

The organization is self-described as having a culture of evidence use and commitment to internal learning (d, d-v). A tension is described between a value placed on learning and an otherwise action-oriented culture in the sector generally. A view that ‘doing and learning’ are

separate things; that learning is ‘a luxury’ that involves ‘sitting around reflecting’ is sometimes expressed (c12).

Search for evaluation approach

Doubt, permission to experiment and ‘capacity to act’ (Weick, 2006, p. 1724) appear to have been important to opening up the search for an evaluation approach (instead of adopting their ‘traditional’ approach) and the selection of DE.

Doubt. Actors were uncertain about (a) the initiative, (b) its objectives, and (c) how to evaluate it.

Doubt about the initiative. This was a new way of working for the organization (d, c12, c11), described as an “idea ahead of the evidence” (d), ‘without manuals or roadmaps and very little established practice to guide the design’ (d), which meant that the organization was “driving off the map” (c11) and “entering uncharted territory”, ‘stepping into an unknown future’ (d). The initiative was also seen as risky. The perception of risk was described as a “very big contextual challenge” (c12) because: “any donor money always went directly to a program - like an actual tangible thing that served people in the community - so this was a risk to say actually we want to use this money to create a better system and a place for practitioners to connect and get better at what they do. That was a very challenging sell because we had to tie that to real change for young people” (c12).

Doubt about objectives. Expectations were high: of lowering youth drop out rates in the urban area (c11, c12) and building knowledge about effective practices in the sector (d). A participant described reacting to envisioned outcomes, saying: “And I was like, what?... it was a significant gift, but... I think we were really missing the complexity of the issue we were trying to tackle” (c12).

Doubt about evaluation. Project participants ‘worried about their traditional evaluation approach’ (d/c12) - that it might not give the understanding needed or provide information quickly enough (d). “We didn’t know how to evaluate this thing” (c12); ‘we knew it would be fluid and unpredictable to try to make sense of’ (d/c12).

Permission, capacity, confidence. *Permission.* Because the initiative was new for the organization and there was no established model, it brought “a certain level of permission” to consider “different ways of thinking about and approaching evaluation” (c11); “it might have been trickier to apply [DE] to a stream of programming or funding that had been done for decades” (c11). Also, the team leader described themselves as ‘not an evaluator’ without “preconceived very technical ideas about how evaluation is to be done” which is likely to have contributed to willingness to experiment with a novel approach (c12). The team deliberated on how to approach evaluation. An evaluation staff person suggested DE (d, c12).

Other factors considered in the choice of approach included the need to develop trust, provide rapid feedback, and support adaptation, all of which align with DE, along with an understanding that DE would allow them “to demonstrate results and to understand the contribution of the initiative to the landscape and issue” (c12). The initiative depended on collaborative work among agencies to openly share and discuss opportunities and challenges. However, in the early days some agencies appeared concerned that the initiative would be used to ‘scrutinize them’ (d-er). This suggested a need to develop trust. Program actors knew that this was considered a ‘risky’ venture, and the evaluation had to provide accountability. They would need to: “keep track of whether we were being successful and report to the donor and the board and the community on this” (c12). They would have to demonstrate progress “in short order”; there was a “keen desire to demonstrate progress on stated goals quickly” (d). As the initiative needed to get underway quickly, it was already ‘in action even while underlying ideas [were] in fledgling state’ (d). Due to this sense of urgency, staff said “we needed to get going quickly, and adapt as we went along” which the team felt DE could support (d).

Capacity and confidence. DE was becoming better known as viable option at this time because of recent publications. Trusted insiders and outside experts added to confidence. A staff person: “had been reading a lot at that point from Michael Quinn Patton’s new book, she was really jazzed about DE” (c12). An opportunity arose to invite Patton to come to the organization (c12). Staff attended the session (d, c12) and DE sounded like a good fit for the initiative (c12). The team leader asked the advice of another colleague who knew and recommended a developmental evaluator (d, c12).

2. Structuring the evaluation

The evaluation began in the year after the initiative launched. The project team had already developed an initial theory of change (TOC) (c12, d). The team and evaluation consultant discussed the TOC at their first meeting and used it as a basis for an evaluation framework (c12, d).

Guiding principles and intentions influencing design

Some principles and intentions informed the structuring of the evaluation. These included intentions to build evaluation capacity, the appropriate positioning of the evaluation consultant vis a vis the initiative, and ‘making evaluation part of the work’.

Evaluation capacity building (ECB). The organization wanted to develop internal evaluation capacity (c11). The evaluation was therefore set up with an ‘internal/external’ hybrid structure. Evaluation functions were shared between the consultant and the internal staff and volunteers (C11, d). The consultant was positioned as a coach to the project team, usually working at a distance but meeting face to face with the team at regular intervals, and participating in some data collection/participatory analysis events (c11, c12, d). Many of the evaluation activities were conducted directly by the internal project team, with the support of the consultant (c11, c12, d).

Documents mention technical experience gained by team members, e.g. survey development (d). However, ECB was understood to go beyond technical skills. An emphasis was also placed on ‘building capacity for analysis and use’ of data (c11). This was reported to be valued by the organization, said to be unusual:

[A] principle perhaps that I’ve come to see over time... rigorous methods, valid survey design, interview process, yes that’s important. But there are diminishing returns. And I would rather see a group get really good at shared analysis and the ability to take conclusions from that analysis and let that inform how their work is unfolding, than getting increasingly better at methods you know. That’s the missing piece for a lot of organizations. And the essential one... it wasn’t the case with [this organization] which

really stands out for me - a lot of organizations put all their marbles in the data, and don't invest in building the capacity around analysis and use (c11).

Accordingly, evaluation work was said to be aimed at fostering “the thinking that people do along the way” that shapes what they do (c11), with formal reporting “a pause along a path of discovery” (d-er/c11). The original evaluation framework was not adhered to strictly, but used as “a helpful conversation that got us more focused on the fact that we needed to be gathering data regularly, and so [the consultant's] role was really to be a coach and to help us learn how to do DE in this initiative” (c12).

Positioning of the evaluation consultant. Although the structure of the evaluation was described as an ‘internal/external hybrid’ this was not understood to mean that the consultant was truly ‘external’ in the sense of being independent of the initiative. The evaluator was ‘not really a 3rd party’ (c12). DEs are described generally in case documents and in interviews as ‘embedded in the work or integrated in the work’ (c12, d). Because of this, they are said to have a ‘deep understanding of the work’, and able to develop trusting relationships, which are important because they can ‘nudge or challenge the thinking of the people around the table where that needs to happen’ (c12). Participants gave examples of ‘nudging thinking’ in this initiative, noted below (c12, c11).

Making evaluation part of the work. Documents describe ‘integrating evidence and critical perspective in a way that was unobtrusive to activities and participants’ (d). This is called “a delicate dance – weaving evaluative thinking into what is happening in a way that does not distract from it” (d). ‘Making the evaluation part of the work’ is evident in how the work was distributed; an emphasis on reciprocity, simplicity and opportunism; an affirmative stance; and treatment of practice and theory as interrelated.

Distribution of work among actors. An ‘internal/external hybrid’ suggests a binary structure, however the evaluation functions in this case were distributed among multiple actors and groups at varying degrees of distance from the initiative. The consultant, the internal team members and volunteers, and external advisory groups and researchers were actively involved (d, c12, c11). The evaluator assisted with the overall evaluation design, designed data collection instruments,

supported team members to design instruments, conducted data collection, supported team members to collect data, and facilitated regular analysis and ‘sensemaking sessions’ (d). Staff gathered data, tracked progress, engaged in sensemaking, critical thinking and analysis (d). Advisory groups acted as ‘sounding boards’, participated in data analysis and provided feedback; external researchers conducted case studies to answer specific evaluation questions (c12, c11, d).

Distribution over time and place. The evaluation work was also ‘distributed’ over time and place so that it could be integrated into the work of the initiative. This was an intentional aspect of design to support sustainability and use of the evaluation. It allowed the timing of data and reflection to better fit key decision moments, said to enhance use (d). A team member said: “we tried to make sure that evaluation was the work and the work was the evaluation” (c12). For example, the evaluation integrated simple data collection into regular events and activities of the initiative, to form ‘a continuous loop of evaluation’ (d-fr, d-v). A participant described, as an example, incorporating a ‘mini evaluation’ at every learning event: “where we just had people write on a sticky note the answer to 1 question” (c12). Another participant noted: ‘the more [evaluation is] just an extra thing the greater the likelihood that it gets dropped. ‘Oh we ran out of time to do that today’, or ‘we couldn’t get around to it’. That’s more likely to happen.’ (c11); ‘[Evaluation] should be part of their work more than this add-on thing, the more that’s a mantra you take, the more you have something useful and sustainable’ (c11).

Collaboration, reciprocity. ‘Making it part of the work’ also involved a commitment to collaboration and reciprocity, reflected in how the evaluation was framed and modeled (c11). A participant said: “I think it’s very much about that [framing] and modeling that. So instead of saying, ‘I’m going to collect some info’ and then disappear, ‘we’re going to collect some info together and we’re going to use it here now for something that’s probably relevant to you” (c11). As an example, a questionnaire was used to collect data before agency events, so that at the event agencies could participate in analysis, ensuring that members could have joint involvement and use of the analysis and the results of the data (c11). Findings were said to be ‘often shared back in presentations and exchanges’ with agency members (d), which helped to create ‘a sense of shared ownership’ (d). Intentionally selecting questions that are also useful to the people contributing data (c11), indicates reciprocity was modeled. At events with member

agencies, the initiative supported development of a learning agenda for the coming year which helped to shape the initiative and the evaluation around the learning goals of members; ‘mechanisms such as the learning agenda helped to develop trust and openness over time’ (d). The evaluation was also aimed at being useful outside the intervention: it provided material for the team to publish on the experience, to help others who want to develop a similar initiative (c12). This interest in ‘reporting out’ as well as within the initiative is echoed in a Year 4 event, a key part of which was extracting lessons to assist other initiatives in other places (d-v). The Year 4 evaluation report also stated a purpose to share lessons with others (e.g., funders, other organizations) (d-er).

Simplicity and opportunism. Keeping evaluation in reasonable proportion with the needs of the initiative was another noted guiding principle (c11) and which seems to have influenced how the evaluation was ‘made part of the work’, for example, by emphasizing simplicity and leveraging opportunities as they arose. A team member said that a valuable part of the evaluation ‘coaching’ was ‘helping us learn... to think about all the opportunities when we could be gathering data in very simple ways’ (c12), for example “the coordinator also learned a lot about data collection, she learned about - like she was always thinking about how we can use this upcoming event as an opportunity to gather information, what is it that we need to learn now?” (c12).

The team did invest in more intensive data collection and analysis on occasion, including conducting a survey, focus groups, and through case studies, although: “we didn’t do a lot of that, we tried to make sure that evaluation was the work” (c12). For example, the case studies were driven by both need and opportunity. They were undertaken at a time when important questions were raised about linkages between activities and anticipated outcomes in the theory of change. The team could not answer these questions with the data at hand and so the questions were brought to the expert advisory panel. The case studies were proposed by a member of the panel, who also connected the team with resources to conduct the studies (c12, d).

Affirmation, permission to fail. Also relevant to ‘making it part of the work’ was an approach to learning collaboratively, with failure accepted as a learning opportunity. This is reflected in comments by participants and in the evaluation design. For example, a data collection instrument that ‘did not go well at all’ was described as ‘a good learning moment’ for the team, both as an

opportunity for team members to learn technical skills (question construction) and to think differently about the initiative because it illuminated priorities (c11). A participant noted that this early ‘failure’ resulted in no harm to the initiative or evaluation overall (c11). Such a ‘safe-fail’ experience is enabled by an evaluation designed with multiple opportunities for data collection and of sufficient duration to try out and adapt instruments. A second participant also expressed an openness to failure as opportunity:

I see all the time when people come together, they come together to talk about what doesn’t work, and how we need more money, totally legitimate. This is a different kind of conversation... building relationships around learning that are more egalitarian, that are more inquisitive as opposed to judgemental, helps to build that sense of shared ownership. It’s not about whether we get it right or wrong, its ‘oh that was a total failure, what did we learn from that?’ [laugh]... it was more fun I think” (c12).

‘Doing is part of theory making’. Accounts suggest actors saw an important relationship between practice and theory which underpinned the evaluation and relates to ‘making the evaluation part of the work’. The phrase ‘doing is part of theory making’ was used in documents (d). Participants described this in terms of implicit theory that is developed through learning in practice, which is reflected in practitioners’ decision-making, and which can be elicited and abstracted to a theory of change (c11); and also as a belief that the work of the initiative itself could generate or elicit ‘practice-based evidence’, or theories at a more abstract level, to inform the work of others (c12). Both action and reflection were described as interconnected in a learning process. Speaking of the sector, a project member stated: “I don’t think we value reflection and learning. I think we value doing, and we see those things as separate, and I don’t think they’re separate” (c12). A participant described working to identify the ‘theory that is implicit in decisions’ and ‘framing up emerging working models’ of actors (c11) as the initiative unfolded. This participant clarified that: “the way not to do [it] is to stand up in front of a lot of people with a whiteboard and a marker and say, ‘ok what’s your theory of change?’ Because it’s torture for everyone” (c11); more effective is: ‘asking questions, getting their interpretation of data and applying back to the [formalized] theory of change’ (c11). This requires intention and skill: “for a lot of practitioners they’re not wired to think ‘this is our theory of change’ but it’s

there and its sort of visceral and you can bring it out and its helpful to them and you probably get a more real, accurate theory of change as a result” (c11). Transitioning between direct experience and abstraction is considered problematic in the literature (e.g., Kolb, 2015; Weick, 2009). This account suggests a process of abstracting practice experience to develop theory, and also grounding theory in practice, through a facilitation process.

3. Mid stage (acting & responding)

An evaluation report for the initiative describes the interaction between emergent outcomes and deliberate actions to be characteristic of complex systems (d-er). Actors adapted the evaluation to both guide and respond to events over the four years in ways that illustrate such interactions between intentions, emergence and response. Examples appear in accounts of the use of questions, tools, and perspectives in the evaluation. These helped the team to test assumptions, actively monitor, attend to the unexpected, and adjust stakeholder expectations as the initiative matured. Throughout the period, actors worked to sustain the evaluation work by ‘holding the space’ for the approach.

Action and response: the use of questions, tools and perspectives

Questions. Good (i.e., well-constructed) questions were described as important for eliciting useful data, adding energy, lending structure, and influencing how issues are framed.

Useful data and energy. Probing questions were said to help the team to make sense of what was happening and ‘avoid running headlong into implementation of the initiative’ (d-er/c12).

Describing the effect of revising a set of questions that hadn’t worked well, the use of new questions marked an early turning point. “That next meeting was outstanding, like really a lot of insights, reshaped some stuff that was very different from... originally anticipated would be happening, gave us direction on how to respond, injected a lot of energy and momentum into the work, and... some really good strategy decisions based on early data and evidence” (c11).

Structure. A question framework guided the team’s reflective practice sessions, held 2-3 times a year. Its general format was: “what have we been up to, what have we learned, what do we do going forward?”. These questions were described as “simple, but the thing that’s often missing

from organizational work” (c11). Documents refer to a similar set of questions that helped the team systematically monitor progress (d-er, d); and which also elicited past, present and future perspectives on the initiative.

Questions were reported to be important for ‘making space’ or creating ‘a kind of structure’ that supports collaborative analysis and reflection. As stated by a project member: “DE requires you to make use of your data, so you have to have a space to make sense of what you’re gathering... you have a set of questions that you’re trying to answer [that] can’t be answered by an external evaluator. They’re answered through the emergence of the work and the learning from the work, and so it’s just inherent that you have to have a space to talk about what’s emerging from our work” (c12). Guided by these questions, the reflective practice sessions were reported to have contributed to rethinking the initiative’s theory of change in a responsive way (d). The sessions reportedly also led to instrumental use in the form of updates to plans for the initiative for the subsequent months (c12). More generally, “they were used to hone strategy and course correct” (c12), in other words both proactively and responsively.

Framing. Questions can influence how issues related to the initiatives are framed, potentially operating as a lens through which issues are considered. For example, to help evolve the group’s understanding of the nature of the problem they are targeting (e.g., ‘what is poverty’) (c11), helping the team to consider what success would look like, and helping to frame the intervention itself (c11); in one instance questions were reported to have ‘helped people to see better’ (c11).

Over the course of the evaluation, new questions were raised in response to what was described as persistent uncertainty and changing conditions in the initiative. For example, membership in the initiative grew very rapidly and unexpectedly over the first 2 years, in both size (eventually reaching nearly 10x expected participation) and diversity of participation (from primarily front line practitioners as the intended target to participants at multiple levels of the system). This resulted in a new question: ‘why are we recruiting so many people?’ (d-er). Considering this question led to a decision to continue growing the initiative and welcoming diverse participation, based on data and alignment with the principles of the initiative (d). This question about recruitment, and other important questions, such as: “who is not at the table but should be?” (d) were raised in response to emergent events (would not have occurred to participants at the onset

of the initiative). They were open questions that suggest an interest in keeping the thinking frame wide but avoiding ‘drift’. A participant said: “developmental evaluation is really about asking good questions and helping to ask good questions and helping people to be curious about their work” (c12).

Use and iteration of tools. Evaluation tools were adapted in a responsive way, but not indefinitely. For example, the set of questions used for agency events (mentioned above) was revised once, and then used 2-3 times a year throughout the rest of the evaluation. The team also used a structured document to record their internal reflective practice sessions. In the end this ‘made a lot of really good data accessible and efficient to get at (c11); it ‘revealed to us the whole arc of the experience, how the thinking evolved over time, what we figured out, how they responded’ (c11). Consistent use of these tools gave continuity and provided data for the evaluation report four years later.

Use of the agency event tool over the extended period was attributed both to the quality of the tool but also to the effort invested to get it right: “[O]ften evaluations create this form of tracking tool and... in my experience there’s a range in how people actually engage with them and do them” (c11). The challenge of the first event ‘revealed to us what we wanted’ (c11); ‘maybe because it was a struggle, and then this moment of clarity and usefulness early, we really stuck with it’ (c11).

A tool that was iterated throughout the evaluation was the theory of change. The original TOC was drafted prior to the start of the DE. It was a focus of discussion at the first DE meeting, and was reported to have prompted regular data gathering moving forward (c12). As the initiative unfolded, challenges arose to the team’s conceptualization of the initiative. Early ideas about how to do things had to be adapted to respond to unexpected change (d). A participant also described ‘nuances to the thinking around the TOC that became clearer over time’ (c11). In the documentation there are reports that “the pace and sequence of the initiative turned out to be different than originally anticipated. Several key assumptions about the initiative changed based on new learning and insight” (d-er). For example, reaching agreement on a common set of objectives among the participating agencies had been a planned part of implementation, but proved impossible to achieve (c12). The first realization that the TOC needed to be changed was

disruptive: “oh my gosh we actually need to rewrite our theory of change” (c11). This was another turning point. Over time, iterating the TOC became normalized and a ‘source of pride’, because it was an indication of learning: “in the end, we were quite proud we were on version 9 of the theory of change” (c11).

The TOC was illustrated in visual form. It was used for external communication about the initiative, as a boundary object, in other words, “it helped people talk with others about [the initiative]” (d11). It was also a platform for reflection within the team: “to check what they were experiencing against their assumptions and expectations” (d-er) (as illustrated in the TOC); “working methodically, the team ... continually related learning to the theory of change in ways that encouraged the team to consider emerging ideas and to share what they were learning” (d). It was also said that: “The sense that [the initiative] was evolving kind of lived there” (d11), and “it clarified something that was ambiguous” (c11) suggesting the TOC played an anchoring role (a thread of continuity). Yet the team avoided locking in or constraining the initiative through the TOC. The team had “developed an approach to our work that was open and we never assumed that what we were doing was the right thing” (c12). Overall, the approach ‘enabled us to hold our ideas about the initiative lightly’ (c12, d).

Eliciting other perspectives. As noted above, doubt and uncertainty were experienced through the period covered by this case study. The team did not have an existing model to guide the initiative, but they used other tools to try to resolve doubt. Drawing on past, present and future perspectives through question structure is noted above. The team also employed mechanisms to elicit feedback from diverse groups of actors outside the internal team.

Agency groups. The evaluation relied on a core group of agency members who were consulted regularly, sometimes through organized ‘strategy sessions’: “They were our key group that we were going back to regularly and saying ‘how are things going, here’s what the data is telling us, here are our thoughts about that, how do you make meaning of that, what do you think, what should the next year look like” (c12). At times the team also reached out to ‘peripheral’ agencies for feedback, for example to understand if there were barriers to participation (d).

Expert advisory panel. An expert panel of 4 individuals was formed in Year 3. This panel was used as ‘a sounding board for big issues’ (c12). There were at least 3 needs met by the panel: lending diverse perspectives, as due diligence (assurance in uncertainty) and adding credibility.

The panel was intentionally constructed to be a diverse group with different professional and academic backgrounds (d) (i.e., rather than diverse in socioeconomic status or another characteristic). When asked why the panel was formed, a participant first pointed to the “benefit from having recommendations coming from a team of people who had a range of expertise” (c12). Another participant echoed this and linked it to complexity theory: ‘in a complicated situation you need expertise. In a complex situation, you need diversity of perspectives’ (c11); ‘their diversity was the critical factor above anything else’ (c11).

One participant described the panel as a way to ensure due diligence in the ‘responsible stewardship’ of the initiative and funds given the uncertainty surrounding the initiative, to help “build a business case for what we could really reasonably do” (c12). They were positioned to lend credibility to the initiative and to the evaluation. In documentation, the expert panel members were called ‘independent leaders’ (d-er) and ‘independent experts’ (d) and the panel was said to have given an ‘independent assessment of the initiative’ that influenced the decision in Year 4 to continue funding (d). Informants agreed that the panel members ‘brought currency’ (c11) and ‘had an awful lot of credibility’ (c12). This was important, because on their own, the staff could “draw conclusions about this and make the case but that might not be heard” (c11).

The panel was described as ‘a really useful evaluation tool’ (c11) and ‘an evaluation intervention’ (c12). The evaluator acted as chair for the panel “and developed really great relationships with them (c12)”. The panel members had a relationship with the organization as well and one member served on a board committee (c11). When asked if the expert panel was comfortable with the emergent approach, both participants concurred. One responded:

They were, that was part of the on-boarding of them, they didn’t just come in at the end, they probably met annually for a couple of hours at least, and so they understood the background and context. It set the stage for them... if it was like, ok we’re about to wrap

something up, you come and be an expert panel, I'm not sure they would have had the same mindset (c11).

Panel members were therefore not completely independent or 'outside' the initiative, but could be described as located at the boundary. They were able to develop knowledge of context over time and through relationships, but only met to assist with 'big issues' about once a year. The other key group—the core agency members—was closer to the initiative. Members participated regularly and were said to have a sense of 'ownership' of the initiative. Both groups were friendly to the initiative and its approach.

Testing, monitoring and attending

Testing the TOC. Documents describe evaluation activities used to 'test the theory of change with many small prototypes and regular data gathering' (d), for example, 'prototyping an approach to building a network' (d-er), prototyping ideas for engaging front-line workers (d), and learning through observation and soliciting feedback from agency members. Participants gave an example of data gathering to 'test' the TOC. The TOC suggested that individual participation would lead to systems change, but 'we didn't know how to make systems change happen'. To better understand the possible link between individual involvement in the initiative to organizational and system change, the team consulted core members of the initiative and the expert panel. A member of the expert panel recommended using explanatory case studies and connected the team to post-doc interns to run the studies. The case study data indicated that these links in the TOC were not operating as imagined, and that 'if organizational change was necessary to the theory of change, then more energy needed to be put there' (c11). These data prompted another turning point, described as a 'shift in direction':

we had some very interesting learning from that which was actually we didn't have really strong evidence of [system impact], and we realized that we hadn't really designed very well for that. And that what we needed to do... [was] to figure out what are the right activities and interventions to get to those kinds of outcomes, and do we actually have the levers to do that, and how would we do that? (c12)

This was then brought to the core group of agencies. “When we brought this to the core team, they really shifted my thinking about what systems change was” (c12). As a result, the team took action to better support this type of change. “Together from all of that data gathering, and sensemaking, we were able to make some really good well informed recommendations to the board and do some program planning to kind of take some shifts in direction in the years that would come after that” (c12). By consulting the agency members:

we were really opening things up and saying here’s what we’re trying to figure out... they came back to us with some really useful analysis ... they were telling us stuff and interpreting the data that we saw in a way that the network effects are stronger than we thought they were, and the organizational change effects are actually weaker than we thought they were. That was really, really helpful. Because we needed to evolve our expectations and change what we’re doing or how we’re going about things (c11)

Overall, the developmental evaluation was described as a demonstration of the organization’s ‘commitment to test how reality aligns with rhetoric’ (d-er). Testing and data gathering were thought to enhance transparency and credibility and to enable deeper understanding of the initiative over time.

Active monitoring and course correcting. In concert with testing and data collection, a watching, responding stance is indicated in the data. This reflects a focus on staying on course, and course correcting. Described as ‘watching for deviations, reminding of the rationale, questioning shifts, pulling people back’ (c11); ‘using data to say, are we on the right track?’ (c12), or ‘pointing out if decisions were being made without, or in spite of, the data’ and by bringing actions back to the theory of change (c11). This work was said to have been applied to ‘making good choices’ (c12). DE is said to also be about ‘never losing sight of intended impact’ (c11): “there’s some confusion that people think [DE is] just about process, and to me you can never lose sight of the impact, the intended impact. Even though your understanding of what that is or what that looks like evolves over time” (c11).

Attending. While DE is described as helping the initiative ‘stay on course’, it also appeared to help raise awareness of and attention to the unexpected. While documentation points to the DE

'supporting the team in dealing with the unanticipated' (d-er), it also appears to have helped them proactively *see* or *notice* the unanticipated in the first place, to 'identify key developments and forks in the road' (c11). For example, when case study results indicated no evidence of systems change through the pathway articulated in the TOC, the team consulted the core agency group. A different story of change was told:

That there actually was systems change but it wasn't that. It was that there was now this group of 400 people across the city from all different types of organizations, at all levels of organizations who now know each other, and who now pick up the phone and call each other and get stuff done. And so front line workers know presidents of universities. And they sat together and talked about things. So if the front line worker wants to create some kind of collaboration with a local college, 'I know that guy I can phone him'. It broke down barriers and they said it really shifted power dynamics in the system, and so whether or not it changed agencies and programs, it changed something else that I think we hadn't really anticipated it was actually a very disruptive system level initiative that created new forms of relationships that would not have existed otherwise (c12).

This was not anticipated or part of the existing 'design' of the initiative, it was not part of a pre-set 'course to stay on', it was an emergent outcome (c11, c12, d).

Reflective practice sessions. Ongoing throughout the evaluation, the team held reflective practice sessions that were used to review progress against objectives and consider implications moving forward (d, c11). The sessions provided a forum with, 'a good intensity of activity around them' (c11).

Documenting. The importance of documentation was emphasized in the public reports of the initiative, usually with reference to credibility of the intervention: "The developmental evaluation enabled [the organization] ... to tell the story of the [initiative] at every stage of the process, including how it was developing, what progress was being made" (d). "In being actively observant in a systematic way, the team could construct a narrative to explain what was happening, as well as how and why, in a grounded and credible way" (d). Documenting progress

was also mentioned by informants. It enabled the team to ‘describe the rationale, basis or evidence for every key decision’ (c11).

Adjusting expectations about impact

Within the team. Over time, participants reported gaining a deeper understanding of the intended outcomes in the TOC (i.e., in terms of sequences and mechanisms) (c12). They also realized that they ‘had no idea how to get to’ some of these (c12). For example, the initiative was expected to impact drop out rates:

through DE we were able to have very honest conversations about ‘how doable is this?’ ‘Do we have the right goal?’ ‘Maybe that’s the right goalpost’. We learned about attribution and contribution and that’s what we were contributing to. It enabled us to actually, through the development of our theory of change and constantly testing and refining that, to learn that what we are doing and what can we hold ourselves responsible for was shifting the conditions in the youth serving sector. We were able to learn to be much more specific about what we’re doing and what our contribution is and manage expectations about that (c12).

A project member contrasts this with would occur through conventional evaluation: “evaluation says [to] ask ‘did we do that?’ and they say no. Actually [through DE] we learned we couldn’t do that. So what we learned was ‘oh, we’re we were a year in and we have the wrong goal’. Or that’s not the way we need to frame this” (c12).

Outside the team. The expert panel played a key role to help the team to reset stakeholder expectations around the outcomes of the initiative. For example, the goal of impacting graduation rates was taken to the panel by the team (c12, c11): “there was this constant tension of yes, we need to hold that out as the higher goal, but our ability to actually measure that was really very difficult, and could be a distraction of effort” (c11). This was “a methodological challenge in that I’m not sure you can measure that in the short term. There’s so much going on that drives change on that, could you isolate an intervention like this from everything else in a meaningful way? So, the solution was this expert panel” (c11). The expert panel supported the view that “the high school graduation goal - that should be in another room. That’s still our end

game... but we need to be much clearer about what's more proximal to us, and how do we describe that and how that is an important contribution to shifting the system" (c12). In general, the panel was said to have helped stakeholders make a distinction between contributing to a change vs an initiative causing a change: "this group helped [the organization] develop a more nuanced understanding that the [initiative] was one of numerous efforts by many players contributing to this broader, population-level goal [of higher graduation rates]" (d).

The core group of agency members was also influential in helping the organization reconsider intended outcomes. For example, the initiative had been intended to become self-managing over time, however it was later understood that this was neither a viable nor a desirable goal (d, c12) for reasons that were not anticipated by either the organization or the member agencies – likely did not exist - until the initiative had reached a later stage of development. For example, the members came to value the importance of having a neutral convener to building trust and 'space' for learning: "they valued that. It took time. It really took time to build that for people to say that, it did not start that way by any means" (c12). This was felt to be another turning point for the initiative. Being 'a neutral convener' meant "we weren't there to teach people how to do the work, we were there to just create a space where people could build trusting relationships and learn together how to do their work better" (c12).

Sustaining the approach: energy, reassurance and 'holding the space'

Energy. A team member pointed to aspects of the approach that helped it to lend energy and momentum over the four years covered by this case study. For example, openness to experimentation and to learning from failure made it 'more fun' (c12). They described telling agency members: "this is an experiment, you are helping us, like let's learn together, here's our hypothesis what do you think – and people said they found that actually quite interesting and it got them thinking about how to use that kind of approach in some of their own work as well" (c12). This participant reported that agency members also responded enthusiastically to idea of building 'practice-based evidence': "Some people get really excited..., they say 'we can build this evidence?' and I say 'yeah!'. You guys are building the evidence by doing this, and if you document this and you share this it's a contribution to the field of knowledge on how to do this stuff. So people, some people get really jazzed about that" (c12).

Reassurance and ‘holding the space’. To sustain support for the evaluation, and the initiative itself, actors report having to ‘hold the space’ for experimentation, reflection and learning. This connects to the tension indicated in comments about learning intentions but having an ‘action oriented’ culture in practice. Both informants described the need to make space, and this was also reflected in the documentation. Descriptions of ‘space making’ indicate that it was both internally and externally directed. Attention to communication and coaching, and input from respected outside experts, were applied to this purpose.

Inside the team. Frustration and stress within the team was noted in documents, and by both informants. “At times the process caused some tensions in the team, as this was a new way of working” (d). The team leader actively worked to reassure people 'one on one' (c12). Their effort was reinforced by the consultant evaluator (c12).

It was challenging to be honest for some staff. Because I think they didn’t have a history of working that way, and I think they felt quite threatened by that, that they felt that this constant asking about the work was interpreted by them as meaning ‘I’m not doing the right thing’, and ‘when can we ever just like land and say this is the way we’re going to do it’, why do we have to keep talking about it, and so I think it was really challenging for them... I think that for people who aren’t used to working in complexity and... are really good at implementing and getting stuff done, having to keep opening that conversation and revisiting was not comfortable for them, it created a lot of ripples and it took time to work through that. It was hard on the team. There were team members who were really into working that way, you know there was a real clash (c12).

This participant linked the situation in this initiative to other experiences (outside this case), in which DE is an uncomfortable approach for some people, stating that it can be a ‘really tough way of working’ for some (‘they find DE so irritating’) that it can lead them to ‘push back’ (c12). They suggested that individuals used to working in ‘very structured, black and white environments’ or who have a strong action-orientation, may find DE especially challenging (c12).

The team leader played an important role to ‘hold space’ for the DE internally. “[I]t’s easy in the early stage of an initiative for people to jump on to something new and different, like ‘sure let’s go for it’, and then you kind of over time you move back to ‘no we actually need to do what we’ve always done’ so I think [the team leader] held that space, and normalized it” (c11). The lead was said to be: ‘really strong and I think [DE] resonated for her, this way of working’ (c11). They were described as an ‘inside leader’ and an ‘anchor’ committed to making it happen (c11). This was thought to become easier over time: “I think in the later stages it was less a risk because the team saw value and benefit and it just became part of their practice, but I think [the team lead] did have to hold that space” (c11); the other participant concurred: it became easier as the team began to “see how their way just kept getting better because of working that way” (c12); The role of the evaluation consultant was noted as a ‘reminder of the rationale’ (c11) and “reinforcing that this is how DE works and this is why its important for us to work this way” (c12): At a meta level it was said that DE itself created space for learning (d, c12): a participant called this aspect of DE, ‘the greatest value that DE brings to my work... it creates a container or space for regular reflection, for relationship building, for learning together about the work’ (c12). This includes ‘figuring out sources of tensions. Collaborating groups often experience conflict. [It’s] not the evaluator’s role to resolve them, but to surface them, help people recognize what’s resulting in tension’ (c11). The documentation notes about the DE approach that there is ‘expectation of development not performance, stakes are lower, creating space for experimentation’ (d). Finally, the expert panel was also said to help inside the organization: “they saw this as a legitimate approach and people then within the organization said: ‘if they’re on board, then we are too’” (c11).

Outside the team (external). Addressing accountability needs, ongoing communications, coaching and drawing on outside expertise were important.

Accountability. Public documents emphasize the importance of DE addressing an accountability need. By “helping to meet internal and external accountability requirements in a way that did not constrain”, DE “reinforced the legitimacy of the approach being taken [by the initiative]” (d). This involved “[d]emonstrating to the donor, governance and the community they could trust

[the organization] to take a calculated risk and track how this unfolded with rigour and discipline” (d-er).

Communications and coaching. Ongoing communication with and coaching of senior leaders of the organization are also mentioned. “From the outset, [team members] considered messages and interactions that would facilitate legitimacy and familiarity for the evaluation” (d):

Developmental evaluation was new to many in the organization, so efforts were made to share the theory of change and introduce developmental evaluation to senior leaders, both at the start and at key points throughout the initiative. These discussions provided opportunities to describe the activities underway, to show how they fit into the theory of change, and to share learning gleaned from evaluation activities. Senior leaders heard about the initiative as it was unfolding, including what worked and what didn't, and what adjustments and improvements were made (d).

Documents highlight the role of the team leader in this regard: As an internal champion, the team leader “helped senior managers, board members, and the project's donor understand and support what was emerging, and managed expectations relating to the nature of innovation in an emerging initiative” (d). The TOC is also noted as an important communication device, helping to show the developing understanding of how the initiative was contributing to change (d).

Expert panel. External expertise was brought to bear in support of the overall approach to the initiative and the evaluation. A participant described the expert panel as strategic and important, so that the organization's executive and board would know that experts from a range of disciplines had been consulted (c12). Another participant echoed this: ‘the staff team felt that they could draw conclusions about this and make the case but that might not be heard. Who could objectively look at that for us and draw attention to or make conclusions about things for which we would have pretty ambiguous data?’ (c11). Along with ambiguous data, a ‘risk in perception’ was noted related to the structure of the evaluation and evaluator's role as a coach: “if it was purely the internal team or purely the internal team with [evaluator] as coach, I'm just not sure the conclusions would have had the same currency” (c11). Both informants reported that

the team recommendations alone would ‘probably have been ok’, the internal team had credibility; external expertise added weight (c11).

External experts: A Year 4 evaluation report linked internal data to research by authorities on collective action and communities of practice, such as Kania and Kramer (2011) and Wenger and Lave (1991) and Wenger (1998) (d-er). Patton also provided a review of the evaluation for this report, which was called an ‘independent review to determine whether the evaluation meets generally accepted standards of quality for the kind of evaluation’ (d-er). Patton references a personal connection with Wenger, noting that Wenger saw value in DE (d-er). Brenda Zimmerman, a respected complexity researcher, contributed content to the report as well (d-er). A participant reported that these contributions were: “trying to give [DE] legitimacy and currency but also some explanation around what this is”, so that readers could ‘situate it’ because this was not a traditional evaluation (c11). In the words of one participant: ‘Evaluation can mean 200 different things. People often talk past each other, you’re thinking performance measurement, I’m thinking adaptation in an innovation sense, and then something happens and one of is not happy with what happens. A lot of that within that report was setting expectations, explaining complexity, explaining developmental evaluation’ (c11). The content and format of these contributions in the report suggest they were intended to coach the reader on implications of complexity, for example how outcomes can be interconnected and mutually reinforcing, and that simple cause/effect relationships are rare (d-er). These data suggest that actors needed to explain, defend and protect the DE approach and its underlying assumptions (e.g., complexity) over time, and did so through careful communications, coaching, reassurance and support of external experts.

4. Late stage (learning, establishing, scaling back)

Learning outcomes.

Productive group learning and ‘intelligent collaboration’ There are multiple references in the data to decision-making informed by information generated by the evaluation. Some are noted above. As another example, when the team realized that front-line practitioners were underrepresented in the initiative, a process of idea-generation with existing members was

followed by small-scale experimentation and observation to assess effectiveness of the ideas. This process fed improvements to the initiative to facilitate front-line practitioner participation (d). As a further example, pressure for the initiative to adopt an advocacy role came from within the growing membership (d-er). This proposed direction was questioned and evaluated using participant feedback, with reference to principles of inclusiveness and the learning mission of the initiative, resulting in a decision to not pursue such a role (d-er). These examples conform to a definition of ‘productive group learning’ (Lipschitz, Friedman and Popper, 2007) and ‘intelligent collaboration’ (B. Davis & Sumara, 2006).

Evaluation capacity building. A participant described an example of successful ‘capacity transfer’ that enabled a staff person to move from ‘struggling’ with development of a data collection instrument to ‘doing a really outstanding job’ (c11). Another participant described the consultant’s coaching as ‘super helpful’ to staff learning technical skills related to evaluation. Agency members cited evaluation among the areas of learning they acquired through participation in the initiative (d-er). Expanded use of evaluation in their own programs was reported and attributed to their experiences (d-er), in particular, applying evaluation to their own program design and delivery (instead of only ‘a funder-driven afterthought’) and use of theory of change (d-er). In addition to technical skills, less tangible learning was reported, relating to evaluation theory (e.g., attribution/contribution) and conceptual rethinking of the initiative (c12, c11, d).

Culture shift. Both direct skill-building and fostering a ‘culture of evaluation’ or ‘a practice of inquiry’ in the team were attributed to participation in the evaluation (d/c12, c12, c11). A team member raised this: “I think for me one of the most valuable things that came out of that is it shifted our culture, and I’ve seen culture shifts in funder/grantee relationships and how people connect with one another, because [DE] is much less about having to prove that I’m doing the right thing or list all the activities that I’m doing and hope that you’re happy with them and much more using information and data to say ‘what does this mean to us collectively?’ and ‘are we on the right track?’” (c12); “it was very much like culture building, having a lot of inquiry and openness to learning about how we were doing” (c12).

‘Learning what we’re doing’. The organization and the team reset their expectations for the initiative over time. They let go of the idea of building best practice models, in favour of more diverse promising practices and strategies (d-er). They let go of reducing drop out rates, and replaced this with a broader objective of ‘educational attainment’. Overall, the focus was shifted to learning around broad objectives, rather than on attaining specific goals (d-er). This was attributed to the DE process which led to rethinking the initiative in a systematic way (in other words, not simply drift away from ambitious outcomes to something less). It was also seen as part of ECB, such as learning about contribution and attribution as noted above (c12).

Documentation echoes the shift to a contribution perspective, and realization that contribution to the envisioned change would be ‘indirect and long term’ (d).

Establishing the initiative. The DE process was important to establish the initiative. A participant recounted ‘using DE to demonstrate that the initiative was a worthwhile investment’ (c12). ‘We needed time under our belt to develop the initiative. Later, we invited the donor to the initiative’s learning events ‘and he was completely totally excited about it’ but ‘it took time to get there’ (c12). The evaluation is credited with helping the organization decide to extend the initiative beyond its initial term in Year 2, and then again in Year 4 (d).

Scaling back. The DE was scaled back after Year 4. This was attributed to a greater degree of comfort with initiative. The resource-intensiveness of the DE is also noted in documentation, along with concern about how to sustain DE over time (d). Some continuing (but less resource intensive) DE work was said to be focused on remaining gaps in understanding (d).

Case 2

This summary traces the course of a developmental evaluation. To provide context, the summary notes some events predating the evaluation. It then outlines the initial stage, middle stage and late stage of the evaluation’s first 4 years. Themes derived from initial coding and analysis are outlined with each ‘stage’ to illustrate how the evaluation progressed, however many themes are relevant to aspects of the whole case.

Description of the Initiative

The program being evaluated is a place-based initiative that supports low-income students to achieve post-secondary education (d0). It is a long-term, multimillion dollar investment, initiated and funded by a foundation, and designed and operated in partnership with a school district (d0, c21). The initiative also relies on extensive secondary relationships with organizations and individuals in the local community (d0, c21, d35, v8).

History / Starting Conditions

Changing approach

The foundation had been investing in local education initiatives for decades, however data were showing declining enrollment, graduation rates and test scores (d0, v11, v12). In response, the foundation commissioned an analysis of its grant making activities. It found no correlation between its investments and student achievement (d0, d35, v12). This marked a major turning point for the foundation, described as a ‘catalyst for change’, triggering a search for a new approach (d0, v12).

In the following year, the foundation signed a formal partnership with a school district to design and implement a new initiative (d0). Over the next 2-3 years, the partners developed a ‘multipronged approach’ to identify and then establish important conditions for student educational success (d0, v12b c23, d6, d5).

Uncertainty and risk

The initiative was different from others of its type because of its intent to comprehensively address multiple key determinants of student success over several years of students’ elementary and secondary education (d0, v12). The initiative was described as an ‘innovative experiment’ and a ‘systems change initiative’ (d0); ‘a game changer for students, families and the entire community’ (d3, d7, v1, v5). The initiative also required a very significant and long-term investment (d0), described by a participant as: “a huge risk... a huge investment”, in which the foundation is ‘putting itself out there in a way they never had before’ (c21). In addition to risk associated with a new model, its ambitious scale, and the financial commitment,

actors were faced with “a shifting environment and turbulent context” for implementation due to surrounding “[e]conomic, social, political, and educational dynamics” (d0, v12).

Doubt, search and ‘stumbling on DE’

The foundation made a commitment to fund an evaluation. The foundation saw itself as accountable to the community (c21), and both core partners were to be accountable to the initiative’s goal (d0). While it was said that ‘rigorous evaluation was required’ given the extent of the investment (d0), actors doubted that a traditional evaluation approach would be effective (d0, c21). Concerns about evaluation approach were linked to the ‘fluid strategy’, ‘fragile relationships’ involved, and overall complexity of the initiative (d0). Actors wanted “an approach to evaluation that would support and inform the initiative as it matured” (d0, also c21). After a lengthy search period, the leaders ‘stumbled on DE’ (c21). Patton’s 2011 DE text was ‘hand delivered’ by an acquaintance (d0). Actors felt it aligned closely with the initiative: “[i]t was as if [Patton’s] volume had been written with [this initiative] in mind” (d0). Actors began a search for a DE evaluator, and hired a firm from outside the community (c21).

Tension and glue: two interdependent organizations

At the heart of this initiative are two organizations in a close and interdependent partnership: the foundation and the school district. A participant explained that they were: “two entirely different organizational cultures trying to come together”; ‘we knew we would have to negotiate as we learned how to be partners together’ (c21, also c23, c22). Also, although both partners were interested in a learning approach to evaluation (d0, c21), the school district’s previous experience with evaluation had not always been positive. They were said to have experienced evaluation ‘as punitive, as testing, judgement, very often finding fault’ (c21). This history and the ‘highly political’ active context of this partner meant that actors tended to view the evaluation with skepticism and concern about what it would involve and how it would be used; this made evaluation a sensitive issue for the partnership (c21; c22). In contrast, the foundation is described as ‘privileged, for all sorts of reasons including governance structure, resources and control over resource use’, which enabled ‘an adaptive culture’ and a more secure position from which to embrace evaluation (c21, also c22). The foundation also had less

experience with evaluation generally (c21). An additional underlying tension existed in the ‘grantor’ and ‘grantee’ positions of the two partners (d0, also c21). Although differences between the two core partners in organizational history, structure, culture and role in the initiative were substantial, a high level of interdependency between them had been built into the initiative’s design (d38, c22, v8). Data suggest this interdependency helped to keep actors at the table and sustain the partnership over time. Moreover, the DE process was said to act as a ‘glue’ that advanced the partnership. This is discussed further below (c21).

Beyond the partnership between the foundation and the school district, a range of secondary relationships in the community are described as forming "a web of public-private partnerships that goes in every direction" (d35, also v8). A participant noted that diverse linkages throughout the community were important to ‘building a movement’ around the initiative, but also meant that there were limits to control: “when you catalyze something you can’t really control it anymore” (c21). An orientation to inclusiveness and awareness of a dynamic context is reflected here. In spite of limits to control, the compelling vision of the initiative is credited with motivating diverse actors to engage and align for a common purpose (c21, also v1). Alignment was also supported by the evaluation, discussed below.

Overall trajectory of the evaluation

The evaluation started in the same year as the initiative was first implemented (c23), called Year 1 for this summary. Overall, the evaluation has been described as entirely developmental in Years 1 and 2, with very broad questions that ‘touched on most aspects of the initiative’ (c23). Over time, questions became more focused, the evaluation was ‘narrowed down’ and deepened as the initiative stabilized and stakeholder interests changed (c22, c23, d). In years 2 and 3, the evaluation questions focused more on coherence and alignment in the initiative (d0), yet also branched out to investigate emergent, unforeseen events in the wider environment that had potential to affect the initiative, for example the impact of rising housing costs on families’ ability to remain in the program area and participate in the program (d0, c22, d23). In Year 3, the evaluation was taking a hybrid developmental / formative character, with greater focus on student progress and parent engagement (c22; c23). By Year 4 the model ‘had sort of come together’ and the formative aspect of the evaluation became primary (c23). Even as

the initiative was gaining coherence and stability, the data illustrate interplay between the initiative and its surrounding context which influenced the evaluation. The evaluation is ongoing and is projected to "remain a key source of learning" for the initiative for the future (d20).

Initial stage

Being excited, being uncertain

DE was 'unfamiliar territory' for the initiative's leaders when the evaluation began (c21). The foundation was described as transparent and excited about participating. The school district was described as more cautious (c22). Some 'skepticism and cynicism' was thought to relate to organizational history with evaluation (noted above) and to evaluators from outside the community (c21, c22). The evaluators needed to work to establish understanding of DE as a distinct approach (d0, c23) and develop trust (c21, c22, c23); this is a thread that continues through the case.

Bounding the evaluation

From the start, the scope of the evaluation was set to focus on the initiative, and not on the internal operations of the partner organizations. This placed some topics out of scope (c22). Navigating boundaries will be revisited below.

Engaging stakeholder groups

The evaluation was structured to enable the participation of three pre-existing groups of stakeholders: (1) a small 'core' team of individuals from the foundation and school district, which was joined in Year 1 by the evaluators; (2) an extended primary stakeholder group of about a dozen individuals from the foundation, school district, and participating schools and (3) an 'advisory group' that included representatives from the community organizations, parent groups, local businesses, post-secondary institutions and donors (d0, c22, c23). These groups differed in their 'altitude' and intensity of participation. The core team convened regularly throughout the year for ongoing decision-making. The primary stakeholder group met at least every 6 months and were involved in some of the 'nitty-gritty conversations' (c23, also d0). The advisory committee convened annually for higher-level reporting on evaluation findings, to

ensure transparency and accountability to the community (c21), and to solicit their perspectives and advice for the coming year (c23).

Designing the evaluation plan

Initial activities included a half-day meeting with the primary stakeholder group for planning and drafting learning questions (c21), and explaining how DE differs from traditional evaluation (d0). After this, the evaluation team spent about three months 'getting a sense of the landscape and different peoples' needs (c23) through interviews, document analysis, and literature review (d0). They produced a draft evaluation plan, including a draft systems map to be used in place of a theory of change (d0; c21).

Redesigning the plan. The draft plan was rejected (d0, c21, c23). It was described as 'too fully baked' (d0) and 'over-reaching' (c21). A core partner asked for more participation in developing a new plan (c21). This response came as a surprise (c21, c23); a participant described it being 'a blank spot for us' that the newness of the partnership and the initiative meant 'there wasn't enough trust in the room' for the usual approach to evaluation design (c23). It was described as 'our first sort of crisis' (c21).

A series of discussions among members of the core team resulted in agreement to rework the plan, despite the significant effort already invested and the impact on timelines (c21). The evaluators were given credit for their 'facilitative and negotiative skills, being able to hear both sides of the dialogue' in this situation, and willingness to invest in 'getting it right' (c21). A two-day working session was held with the primary stakeholder group to co-create a new plan. The group engaged in system mapping, discussion about expected outcomes, redrafting learning questions, and planning for data collection and use (d0). It was a 'much more of a co-creative process' (c21); "standard stuff an evaluator would do" but as more of an 'interactive workshop' (c23).

This 'first crisis' and its response were called an early turning point for the evaluation and for the initiative (d0; c21; c23). Three outcomes in particular were noted: changing how the initiative and the evaluation were framed, enabling some learning by team members, and contributing to relationship development:

Framing the initiative and the evaluation. In the short term, the systems mapping was described as helping everyone ‘get their heads up’ out of their own day-to-day activities to see a bigger picture of the initiative (c21). It ‘recast the players for everyone’ and it ‘got us into the same space’ (c21, also d0).

The session also involved resetting expectations for the evaluation. The development of evaluation questions was said to have been carefully guided by the evaluators to ensure the ‘nature and tone’ of the questions focused on understanding the initiative and how it was developing rather than on testing an established model or measuring impact (d0). A participant noted: “often you have evaluators coming into town and there are expectations of what is going to be produced and ‘are we going to know our impact on graduation rate?’”. This participant described the need for: “kind of talking people down from that” to think about “the next couple of years and... the sorts of things we’re interested in that can lead to ultimate outcomes” (c23). More broadly, the event ‘changed the dynamic’ in the group from concern that the evaluators were there to ‘judge what the players were doing’ to seeing the evaluation as ‘in service to us’; and ‘about stuff we’re curious about’ (c23).

Learning. The core team, including the evaluators, took away learning from this experience that they applied to the rest of the evaluation. For example, the ‘co-creative process became sort of our mantra: co-create, co-create, co-create’; “fortunately we learned that lesson right off the bat, that we really do need to be very engaged with the team in every aspect of the work” (c21, also d0, v12). In addition, as a ‘multi-sensory’ exercise, the systems mapping activity was said to facilitate learning by participants about the system surrounding the initiative (c21). The session was also credited with enabling front-line participants to understand the reason for data collection activities, and to support them later because ‘they made that connection back’ to what they had learned in the session (c21). The new plan was not radically different from the original (c21; c23), but the session was said to increase ‘understanding and buy-in for the plan’ (c23).

Relationship-building. Participants said that this first challenge seeded important long-term relationship development (c21; c23). It ‘helped build our muscle as a team’ (c21): “With every challenge you sort of come out of it stronger. I think the challenge of the reboot of the

plan, that challenge helped us gel more as a team and understand each others' perspectives better and how we're approaching this initiative. Sort of the process part of it helped to support the work" (c21).

Needing a runway. The working session was not an isolated event; it was a part of a larger set of activities which, although less formal, were important to this becoming a 'turning point'. For example, a participant explained that the systems map was not the result of "one magical meeting" but instead the map "was something we had played around with, showed folks in some of our individual conversations and made tweaks around it, and came to this meeting with a decent draft that then got worked on by folks. It does require some kind of a runway to get to a point where you can have a meaningful conversation" (c23). When asked to elaborate, they said:

it's hard for people to just come into a room fresh and just think about a system. Systems are so amorphous and big and there's lots of things going on. I think to do some initial conversations and preparatory work in putting some visuals together and then... going down the path a little bit both in terms of the actual map and in terms of the thinking process of different individuals helps, before you actually have this larger group conversation about a map (c23).

This example relates to a theme of pacing and framing information to facilitate use, which will be revisited below.

Underlying 'meta-strategies'

Some meta-strategies or principles appear to have influenced how the evaluation was structured and executed. These included being responsive, seeking perspectives and raising voice, positioning the evaluators, grounding the work in data, and systems thinking.

Being responsive. Being responsive to change and stakeholder needs was said to be important, and to align with supporting the development of a changing initiative: "given that the initiative is changing, the kind of DNA of evaluation has to fit the DNA of the initiative" (c23). It involves, "iterating the [evaluation] design as you go along and being flexible with it" (c23,

also v12, d0). Responsiveness was said to be important to maintain the relevance of the evaluation for stakeholders, and was linked to building trust and engagement of stakeholders (c22, d20). This in turn was said to allow the evaluation to go deeper, and to enable the use of findings (d20). Responsiveness is evident at multiple levels through both informal and formal activities, as follows.

Levels of responsiveness, micro, meso and macro. The evaluation was ‘flexed’ at a micro level each year. Activities were changed based on emergent needs and what was being learned through the process (d0). The evaluation was also deepened and focused over the four years from broad ‘what is this thing really?’ questions (c23) to questions about alignment and clarity of constructs such as ‘what is meant by ‘college culture’?’ (c23). At a more macro level, the evaluation changed from a developmental approach to a formative approach as the intervention took shape and stakeholders’ interests changed (c23; c22). It should also be noted that an effort was made at reflective practice by the evaluators for their own work, and explicit feedback on how the evaluation was working and what could be improved was solicited from stakeholders (d0).

Formal and informal activities. *Formal opportunities* for responsiveness were built into the evaluation’s structure. These allowed actors to reflect and rethink the direction of the evaluation and helped to provide space for the evaluation to change direction (c22). For example, a participant noted that the evaluation questions were intentionally revisited every 6 months to ask, ‘is this information still making sense?’ (c22), and: “Every year we go back to the evaluation questions that we had the year before, the activities, and within a very collaborative effort, last year we spend 4 or 5 hours, we... just figure out ‘what are we interested in learning this year?’... we’re never assuming we’ll continue activities” (c22). *Less formally*, bringing a complexity lens to this evaluation was said to enable responsiveness (d20). Related to a complexity perspective is developing a deep understanding of the local context and maintaining awareness of changes through active monitoring, through ‘lots of listening and paying attention to what is going on’ (c23, also c22), discussed further below.

Seeking perspectives, raising voice. Bringing multiple perspectives and voices to the evaluation, in a meaningful way, was described as important both for the quality of findings, and to ‘make the findings real’ for users of the evaluation (c22). Additional perspectives were said to directly improve quality of findings by helping to offset the degree of uncertainty about what is happening in the system (c22). As an example, an extensive effort was made throughout the evaluation to reach parents not represented in the data (d0; c21; c23), which was said to “model a commitment to co-creation/representation in the evaluation” (d35). As another example, a participant described discussing findings with different stakeholder groups, and comparing evaluation data with other data that stakeholders may bring to the table, because this helps ‘to come to a better understanding of what was really happening’ (c22). This was said to involve ‘having humility’ to realize data would always be partial (c22). It is also related to a complexity perspective.

Helping people to connect with the data through voice was said to be important. For example, case studies of individual students were used “to give a more ‘real’ feeling to the findings” (c22). Presenting findings to the advisory group was described as a way for members, who were also acting as informants to the evaluation, to ‘see their voices being raised up’ (c21). ‘Raising voice’ was said to indirectly feed back to data quality by demonstrating respect for informants’ needs and perspectives: “If you don’t take the time to help people understand that you’re there to help them, and to raise their voices, and they see that you’re just using data and doing your job and you don’t really show that you care, I don’t think the quality of the data is the same...” (c22); “when there’s an assumption ... that we... wanted to learn from their voices, they could be candid” (c22). This participant described the importance of returning results of the evaluation work back to those who contributed data as a form of reciprocity: “Otherwise they’re not going to see the value in talking to us. We’re just extracting information from them and we’re not giving them information back, so we have to find ways to communicate back... and make sure that they were feeling like they were being heard and that it was not just a one-way street” (c22). An evaluation with multiple rounds of data collection, analysis and reporting, such as this one, provides opportunity to develop a relationship with informants that may not exist in other evaluation designs. The way that questions are framed was also said to be important:

if you are giving people power, people can see that. Like if you ask a question and say... 'what are the things you haven't been able to achieve', instead of asking 'what's your vision for how programs should look like in your school', and 'what are the things the ... foundation is providing and what are the things [they] might not providing and might be able to provide'... if you shift the way you ask the question into an understanding that it's our client in some way that is as responsible for this as them, I think that's really helpful (c22)

Positioning of the evaluators. Some guiding principles were suggested about the evaluator's role. The positioning of the evaluators relative to the intervention was sometimes ambiguous. As mentioned above, the evaluators were not from the community; they travelled in and out for evaluation-related activities, and were described in interviews as 'a third party' or 'neutral broker' (c21). Yet they saw their role as members of the team, working "side-by-side" with the initiative's stakeholders (d20). A participant explained the evaluator 'is at the table with the client' (c23), and elaborated that: "you are 'on their side', right? [The evaluation is] not to put anybody down, it's not to point fingers, it's not to say 'gotcha'. It's about for everybody to say: 'how do we make this better?'" (c23). The evaluation was structured to enable close communication with core team members, deep understanding of the context and development of trust. Position shifts are indicated in interviews when evaluator participants referred variably to the initiative's team as 'we' or 'them'; to the initiative as 'our initiative' and 'their initiative' or to the students in the community as 'our kids', for example. Data suggest that pragmatic purposes are sometimes served by variable positioning, which will be revisited below. At the same time, awareness and deliberation about role boundaries were evident. Participants volunteered comments on positioning related to (1) maintaining scope of evaluation activities and (2) impartiality to enable critique. In terms of scope, a participant said:

Our job is to continue to provide this information, and if needed help brainstorm about possible solutions, but the solutions are implemented by the client... it's very clear in my mind boundaries about who is responsible for putting in a solution. Over the longer term, I think the evaluator's role is to bring it back year to year and say, 'hey we're still hearing some of these messages' or 'we're hearing something different from last year' and it might

suggest these kinds of activities, and we could go as far as, you know, putting people in the conversation who could say, ‘what do these findings imply for action on your part’ but they still have to figure out what the action is and take that action (c23).

In terms of impartiality, another participant said: “I call it being a neutral advocate for the effort. You’re advocating for the effort, of course I want this program to succeed, but at the same time I’m neutral. I’m going to be rigorous in my data and I’m going to tell you what I’m finding, even if it’s not positive” (c22).

Grounding in data. Participants point out that DE tends to be flexible and responsive, but this does not reduce the importance of rigour guiding the evaluation process (c22; d0; c23). Rigour was described by participants in terms of validating assumptions, quality design of data collection instruments, robust data, and transparency (c22, c23). “For all that developmental evaluation is about relationships and collaboration and facilitation, I think it’s still an evaluation, so you still need to ground your observations, insights and findings in data” (c23). “We have evaluation tools that we can use [and] we have to give them justice. We need to be able to stand behind our data for everything we’re saying” (c22). ‘Grounding the work in the data’ is a recurring theme, revisited below.

Systems thinking. Systems perspectives and approaches are evident in how the evaluation was structured and executed. This is a principle of DE, modelled in this case in multiple ways, for example in helping actors broaden their perspective of the initiative through mapping the system around the intervention. Other examples are available below.

Middle Stage

The initiative moved quickly into implementation in the fall of Year 1. The initiative’s path over subsequent years was said to have been shaped by the conversations held among stakeholders as part of the evaluation (c21, d0). Documentation describes these as productive, data-driven conversations that were a ‘primary mechanism’, enabling reflection, ‘collective sensemaking’, strategic learning and utilization (d). Helping stakeholders have conversations is a central theme through the middle stage of the evaluation.

Helping stakeholders have conversations

A participant described the evaluators “trying to come to the table with data to help [stakeholders] have a conversation and make sure that it’s a productive conversation” (c22). Both formal and informal activities provided support for these conversations, including: making space for conversations through structure; ‘grounding the work in the data’; and building capacity and ensuring buy-in. Less formal activities included positioning and ‘giving cover’; keeping an ear to the ground; ‘walking a fine line’; helping to see; and demonstrating a commitment to learning together.

Making space for conversations through structure. The evaluation was intentionally structured to make space for conversations with the core team, the larger primary stakeholder group, and with the advisory group representing the wider community. The conversations with the different groups took place at different ‘cycle speeds’ and intensity related to the group’s proximity to the intervention.

The conversations centred on feedback about the intervention provided by the evaluation. The processes to deliver this feedback were structured and adapted in responsive ways to help conversations to be productive. An example outlined below illustrates a collaborative effort to tailor feedback processes, so that they would fit the capacity of the core team to absorb and interpret data. It also illustrates how structuring can affect the way that feedback is received and used.

Structuring feedback processes. Year 1 was described as a period of intensive data collection and feedback to the core team (c21, c22, c23, d0, v12). Data collection and feedback were initially structured in 3 steps, to be repeated every 2-3 months: (1) on-site data collection, (2) a ‘learning memo’ sent by email to core members about two weeks after the visit, and (3) a teleconference with the core team to discuss the contents of the memo. The intent of the learning memo was to provide a ‘candid’ report of what the evaluators were observing (c22), to note and track patterns over time, identify urgent issues that might have been “completely off our radar”, and any other things “that might require a little fix” (c21, also c22). The feedback process was

supplemented with additional ‘check-ins’ by phone for updating or planning the next round of data collection (d0; c21; c23), as well as formal mid-year and end-year reports (v12).

Restructuring. The initial feedback process became overwhelming and resulted in pushback (c21, c22). The response was attributed in part to the newness of DE: “we explained what developmental evaluation was, but this was not the way they had worked, so for them, they were overwhelmed”; team members said they ‘could not digest everything being provided’ (c22). Challenges processing the volume of feedback in addition to their regular workload, trying to ‘make sense of it on the fly’ and the uncertainty about expectations for action were said to have produced stress (d0), especially for the school district given their ‘grantee’ position in the relationship (c21, also d0):

these learning memos every couple of months... they were *full*, there was so much there... There was so much there that needed attention... [a school district member] had this sense of anxiety that... [there was an] expectation that everything that needed attention would be checked, sort of check, check, check – that [they] would attend to all of those things” (c21).

This led to ‘a really uncomfortable conversation’ (c21). The evaluators were said to have: “understood the dynamic and they were very open to trying to get ideas... about how to make this more effective” (c21). The core team began to negotiate a new process midway through Year 1 (d0).

Informal debriefing. The team added a verbal ‘debrief’ at the end of each round of data collection, while they were still on site (c21, c23, d0). The benefits of the in-person debrief were described as four-fold. It helped program members to do some initial processing of emerging findings (c21); it provided feedback faster, closer to ‘real time’ (d0); its informality helped to reduce anxiety related to the more formal memo: “when the learning memo landed in the inbox it wasn’t a shock” (c21), noting that “when something is in writing it feels a little more high-stakes” (c21); it also provided the evaluators with more context. It: “gave the team a chance to sort of assess or say ‘is there something hot button here’ because they don’t know all the dynamics that are happening” (c21).

Selecting and framing feedback. The evaluators also became “more selective” about the information being delivered: “to make sure that we were not overwhelming them with data as we were at the beginning, and we were not suggesting way too many things at the same time, and being very selective in picking up and making sure that we were paying attention to things that were the most important” (c22, also c23). The memo was reformatted to be more structured around specific learning questions and to point more directly to what was working well, and what needed attention (d0, c21, c22). The memo was changed to a table format instead of in narrative form (c21). These changes were said to shift the memo from being “really overwhelming and really unexpected and almost a test sort of, to ‘here’s validation of some of the work being done, we’re pointing to some areas where you might want to continue to have work [and] because we’ve had the verbal debrief this isn’t a surprise’” (c21). This also provided more clarity around what needed immediate action and what did not (d0, c23). A participant noted that this restructuring helped change the dialogue “from a place of being defensive” and “attacking the data itself”, to “a ‘so what’ and ‘now what’ conversation” (c21).

Delineating priorities for action was described as a skill (d0). It was said to involve awareness and judgement on the part of the evaluators about what could be changed in the near term, what would involve a longer-term process of change, and what warranted monitoring but not action (c23, d0; v12). A participant compared data indicating a gap in awareness about the program, to data indicating attitudinal barriers. Awareness is more of a technical problem that can potentially be resolved quickly, but “changing... mindsets and cultural competency is a long term adaptive issue, so it’s not like we expect an action immediately... So, to set some boundaries and expectations about when we share things back. It’s not like we expect everything to be acted on right away. It’s just something to be aware of and over the next several months and years how do we start putting some things in place?” (c23).

As a further adjustment, the foundation and district team members began to hold a separate conversation on their own, after the teleconference with the evaluation team about the learning memo, to discuss possible next steps and priorities for action (c21).

This sequence of interactions following data collection (verbal debrief, learning memo, team teleconference, separate next steps discussion) are an example of distributed sensemaking (across tools, team members and time). The sequence was repeated every 2-3 months through the first two years, and helped the team to move through the ‘what’, ‘so what’, and ‘now what’ thinking process (i.e., from problem to action space, and from past through present, to future perspective). The sequence also iterated between informal and formal communications in a way that these communications could support one another (c22), providing another ‘runway’ for the use of findings, and reflecting the importance of “pacing and framing information” to fit how actors process and share information (d0).

Grounding the work in the data. As noted above, ‘grounding in data’ was described by participants as an overall guiding principle. Data were meant to be at the heart of the conversations with stakeholders. In the first year of data collection, the evaluation conducted three surveys and gathered interview and focus group data from more than 70 participants (d0). Data sharing agreements with the district provided additional data on student progress (e.g., attendance rates) (d23).

Actors describe using data in multiple ways. Data were used to bring more certainty to the initiative: “I think the role of data is to have more certainty as opposed to making assumptions. It’s illuminating assumptions and walking on a little bit more solid ground” (c22). Data were said to help discern trends amid the uncertainty of change: “when you are innovating or building something new, sometimes it’s really hard [to see] what are the early wins, the early outcomes of what you’re doing... some of what we do with all the data collection is trying to understand what the progress is... at the different levels” (c22). The data helped the evaluation to provide ‘on-track/off-track’ feedback, for example in terms of students’ academic progress and intentions to seek post-secondary education (d0). Data were also used to validate and extend stakeholder impressions of what was going on, to ‘crystallize an underlying issue’ or ‘point to a solution’ (d0). Being able to ground the work in data was said to be especially important when findings were controversial: “going back to the facts and saying: ‘this is what we heard, and this is how many people we heard it from, here are some of the key messages that came through to us’. Going back to the facts, going back to the data” (c23). Data were used persuasively, for

example to support a reconfiguration of the program in Year 3: “we had to go back to some of our key donors and say we are going to change our model are you ok with that? And they were. We were able to point to the evaluation findings to say ‘look’ there’s data to support this shift... If you want to talk about a win, that was huge, huge” (c21). Data were reported regularly to the stakeholder groups and to the wider community in public communications (e.g., in online reports to the public from the foundation, in media interviews) (e.g., d3, d23, d31). Media spokespeople report the initiative using monitoring data to identify areas of focus for improvement (e.g. d28). The public reporting suggests a commitment to transparency and the utilization of data, and the importance of data to credibility (see e.g. d3).

Building capacity and ‘buy-in’. ‘Helping stakeholders have conversations’ also involved capacity building. This evaluation was said to have “identified ways to build capacity among... stakeholders to receive feedback and to use information to drive decision making” and through both “formal and informal learning processes” (d20, also v12). The primary stakeholder meetings provide an illustration of this (c22, c21). At the June meetings, participants were “really going through all of the findings together” (c21) and the results were used to inform the initiative and DE direction for next year: “Those were extremely useful meetings ... where the rubber met the road in terms of us bringing a summary or synthesis of our findings so far, and then really talking about what actions, and how does this change our plans, and what should we think about for next year” (c23). We “turned the year end report on its head. We put the stakeholders in a room and had a sensemaking day for the whole day, and the report was informed by the sensemaking” (v12). Another participant described:

bringing together both institutions and giving them a table with a lot of data where they can solve some of the things that were coming up with the program and just make the program better, so to see that meeting and see the reactions and changes in strategy, that was really meaningful (c22).

In addition to instrumental use (e.g., changes in strategy) the regular primary stakeholder meetings were said to help reassure stakeholders by increasing the transparency of the evaluation (c22): “all of the school principals and superintendents and foundation staff... we would review the data then with them and nothing would come out without everyone buying in and everyone

just asking questions about the findings and trying to get a little bit of a better understanding of why certain things were happening” (c22). Two other factors were said to make these meetings productive: good preparation, and routine:

a couple of times I was like - ‘I don’t know if this will go well’ but [it] ended up actually being quite productive, I don’t know why, it surprised me... [W]e did some good work setting those up in terms of expectations and materials ... and after the first year especially, people knew the routine a little bit, ‘ok here is the meeting we come to where these guys talk about some findings and we talk about implications’, it had gotten more habitual (c23).

This participant describes why it took time to develop: “just the act of looking at data and thinking about implications is not a natural thing for most people, right? It’s not how most systems work” (c23). Building actors’ capacity and ‘habit’ for using data was important to the progress of the DE.

Coaching about developmental evaluation. Capacity building also included developing stakeholder understanding of DE, through ongoing coaching (d0, c22, c23, v12). This included explaining the differences between DE and formative or summative evaluation, and consistently guiding actors away from expectations and assumptions that were more appropriate to these other forms, which were said to be more familiar to stakeholders (d0); an actor noted that some stakeholders were ‘conditioned to think evaluation has to be a particular way’ (v12). A participant estimated that actors began to understand DE toward the end of Year 1 (c23) and DE was described as ‘an adjustment’ for stakeholders that took time (c22).

Positioning and ‘giving cover’. The evaluators’ position relative to the intervention was also used to enable productive conversations. As noted above, the evaluators were understood to be ‘at the table’ with program implementers and ‘on their side’, while at the same time setting boundaries between themselves and the intervention in order to maintain scope and perspective.

The DE role was described as very ‘relational’ and requiring skill to navigate (c21). For example, a participant described the evaluators being able to “have difficult conversations and negotiate them and help move through when there’s discomfort, or even being comfortable being

in a place of discomfort. It's just much different from someone who's purely in the land of controlled trials, where it's just very scientific" (c21). Understanding of context and being an 'insider' underpinned the relational role: "having the cultural competency and the awareness of power structures ... Seeing me there at different community events and just having conversations that have nothing to do with the evaluation" means that 'I am no longer just the guy... coming from outside asking all these questions' (c22).

However, an 'outsider' position was also said to be sometimes useful: "when you don't live in that community... in some ways it's a good thing because you have the external status and you can play that card and you can ask the questions you want to ask" (c23). Both negotiating skill and an image of neutrality were said to make the evaluators able to act as brokers among the collaborating parties. The evaluation team was described as a 'neutral broker' or 'the outsider' in the room in situations where there was a need to negotiate conflict or act as a credible source of potentially unwelcome information (c21). The evaluation process itself was described this way, because it sometimes "validated things that we were thinking... crystallized that and got it onto the table. [The evaluation] provided a vehicle for those kinds of things to be put forward by a third party where we could say, 'yes'" (c21). A participant described an example of a problematic situation brought forward by the evaluators. Although the situation was already known to members of the intervention team, it was politically difficult to navigate without damaging relationships in the collaboration. In this situation, the evaluation process brought the issue forward and 'gave cover' for actors to address it (c21).

Keeping an ear to the ground. 'Keeping their ears to the ground' (d0, v12) was said to increase understanding of the local context and changes in the context, and support responsiveness of the evaluation. 'Really understanding the local context' was described as helping the evaluators to maintain the evaluation's relevance and to build trust (c21). A participant said: "we want to know some of the news and how is their thinking shifting, so we are sure that we are meeting them where they are. So we're not looking for data that might not be as relevant for them" (c22, also c23). Keeping an ear to the ground and understanding context was enabled by having a trusted advisor living in the local community, and being in frequent communication with them (v12). The evaluators also described spending time with stakeholders

while in the community, ‘eating tacos’, participating in community events, and following local media, Twitter feeds and organizations’ Facebook pages (c22; c23). A ‘check-in’ email to front-line staff every few months elicited answers to a few short questions (d0, v12).

Through frequent feedback cycles, the evaluation was said to relay any observed issues to actors in the initiative in time for them to respond: “every time something comes up we just show it to them. We don’t wait until the end of the year to show them the report” (c22). When asked whether there were sudden changes in context that required a shift in the evaluation’s direction, a participant replied: “That was all the time” (c22). For example, rising property values and an ‘investment boom’ in the area took place, which media attributed to the program itself (d11, d13, d32, v1). The effect of this on housing costs was raised as a concern by parents in evaluation interviews (c22), and potential consequences in terms of student participation became a priority for inquiry in Years 2 and 3 (d23, c22, v12). The evaluators extended data collection to include local universities, business developers and nonprofits in neighbourhoods (c22). As a result, the foundation funded an effort to support creation of low income housing, and to ‘keep communities where they are’: that was a big turning point, that was something we needed to solve that we weren’t thinking about... we provided the initial solid grounding for them to start making decisions (c22).

Walking a fine line. The stakeholder conversations needed skilled navigation at times. Participants described the need for careful communication in collaborative initiatives in general: “an important role an evaluator plays is a little bit like a mediator or – I think you have to be really aware of power dynamics – and you need to be really careful in the way in which things are communicated and how you are asking the different questions” (c22). Evaluative critique of “pieces that people hold dear” can result in defensiveness, and “sometimes people are more ready to hear some messages than others” (c23). When asked about challenges encountered in this evaluation, a participant said that navigating challenges:

had a lot to do with how the evaluator is able to convey the information without making it seem like one party is being put upon. So obviously, there was information that we brought back around ... tricky topics [and] the message you want to convey can get

caught up in peoples' mental models, defensive stances... some of those conversations were trickier than others, and we just kind of had to work through it (c23).

As noted above, the evaluation's scope did not extend to the partners' internal operations, and this constrained the topics that could be discussed. The need to bound the work in this way was challenging because the work of the partners and the work of the initiative were highly interconnected; as described by a participant, the boundaries were 'super blurry', making it difficult 'to talk about what is happening in the initiative without talking about what a partner is doing or not doing'; "we had to be very careful, we were always walking with a really fine line" (c22).

Most comments on this theme related to delivery of feedback, however the sensitivity implied in 'walking a fine line' was part of other activities as well. For example, in choosing what questions to ask: "a big role for DEs is to know what the right questions are at the right time" (c23). Another example is sensitivity to the burden of evaluation on the implementing organizations: "we have to be very vigilant, because they have already many people, all this testing they're already doing, and here we are with another survey" (c22).

Helping to see differently. Also part of supporting productive conversations, the case data indicate efforts by the evaluators to help actors see the intervention in different ways. This included broadening perspectives, surfacing implicit ideas and recognizing the unexpected.

Broadening perspectives. The evaluation was credited by a team member with helping to 'open our eyes' to how complex the problems were that they were trying to solve, and how they could approach problems differently (d0 - c21). Actors recognized the challenge stakeholders have to see beyond their own day to day work. Some of the primary stakeholders in this initiative were managers of organizations, and described as: "very much head-down, putting out fires day to day" (c21). The systems mapping was described as especially helpful: "I think it was an effective tool because it really helped to get folks, especially the folks putting out fires every day, to kind of get their head up and see how the work, what we're trying to pull off here... is not just about what's happening right in front of you right now" (c21, also d0).

The systems mapping was described as different from just having a group discussion. It can be a more interactive, multisensory activity with the map acting as a ‘medium for people to interact around as a group’ (c23). A participant said:

it gave us a chance to reflect how each other were seeing our system. Interacting with each other. The exercise of sticking post-its up and moving them around and drawing linkages... that exercise of physically interacting was helpful, more engaging and more visual. If you think about hitting your different sensory in order for learning to seep in, you sort of have the visual cues, you have the auditory cues, the kinesthetic cues to help embed that perspective in a different way (c21, also c23).

A participant noted: “I think it does matter for people to look at a picture and words on a picture together... it does create a different kind of understanding ... [it] provides coherence around the concept of what you’re trying to do” (c23). The mapping also opened up new ideas. It: “brought a level of understanding to what this larger system is that [the intervention] lives in. That then helped us ask some questions and think about some outcomes” (c23).

Helping people ‘get their head up’ was described as something that needed ongoing work: “sometimes clients [are] caught in the day to day operation of the initiative... what we’re trying to do is to continuously say: ‘hey, yes, you’re having this problem right now, but remember that this where we’re going, and this is all the other things you need to pay attention to’. Keep everybody thinking ‘ok are we getting to where we said we were going to?’, ‘is the model really shaping in the way we thought it was going to shape?’” (c22). The quick ‘check-in emails’ that front-line staff were asked to reply to were said to become useful to the staff themselves “who commented on how writing such an email allowed them to take a few minutes outside the “busyness” of their days to reflect” (d0). Noted above, the data collected through the evaluation helped identify patterns that are described as difficult to discern when an initiative is being actively developed. This was also said to help stakeholders see the intervention from multiple perspectives. For example: “you see data from students changing, you see how teachers are perceiving different aspects of the program, administrators – ... a multi-level understanding of how are different groups and different activities are getting to full implementation, to where we’re trying to go – or not” (c22).

Bringing a range of perspectives to the evaluation is an important part of a systems approach and has been mentioned above as a ‘meta-strategy’ guiding the evaluation. The annual advisory group meetings were described as providing helpful perspective: “these are basically a group of folks that are involved but just not involved in the day to day, so they asked good questions and gave us some ideas to think about” (c23). Another example is the extensive effort to elicit perspectives of parents who were more difficult to reach. Data collection activities were modified for this effort, and the foundation funded the development of a network outside of the school community to engage parents (c21, d23). “[W]e were really relying on connecting to families in schools, and that was just not effective ... that network [is] as a way to hear parents’ voices reflected more completely in our evaluation findings” (c21). The network represents an effort to build structural capacity for accessing different perspectives (d0). It demonstrates commitment to including diverse ideas, and long-term strategic thinking.

Surfacing and articulating implicit ideas. Year 1 evaluation findings pointed to differences among stakeholders (e.g., staff, parents, community members) in how they understood key constructs important to the initiative, such as ‘college culture’ (d0, c21). Beginning in Year 2, part of the evaluation activity shifted to look at these variations and develop a shared understanding that could improve alignment in activities around building a ‘college culture’ (d0). Articulating and clarifying constructs such as this one was described as important for multiple reasons. Most fundamentally, fuzzy concepts are difficult to ‘see’:

it’s just really hard to see it, right? How can you evaluate it if you can’t define it... when you go in you have to really ground a solid understanding what it is... but I wouldn’t say just for us [evaluators]. Because this is not a traditional evaluation. You know for a traditional evaluation, you just need to facilitate a conversation so **you** can have a better understanding of what they’re doing, but in this case ... they needed to have a better understanding of what that meant (c22).

This participant elaborated: “if we didn’t get to a common understanding it would have been impossible to actually implement this [part of the intervention]” (c22). Once a shared understanding started to take shape, the team could structure action around it: “when we started getting a more common understanding, then we started thinking about what are the supports that

schools should have in place or what are the external relationships [they] need in order to ... create that culture?" (c22). For example, recognition of the importance of 'critical mass' of student participants to the development of college culture was one of the reasons for a reconfiguration of the program in Year 3 (c21).

Surfacing the unanticipated. A focus on strategic learning in this evaluation was said to orient the evaluation to 'provide signals': "about the initiative's direction as it is being implemented... illuminate topics that were previously unclear, and spotlight issues that were off the radar" (d0). Given that the evaluation began at the same time as the start of program implementation, unanticipated issues could be surfaced early enough for actors to respond. For example, indications of low parent awareness of the program were raised in time to meet a critical program milestone (d, c21, c23):

[W]hen we had brought some data back on low levels of parent engagement, and that sort of opened some eyes to the extent and nature of the problem, and people got into action mode sort of immediately and got the troops mobilized and stuff sent out and they actually got parent engagement up ... pretty fast over the next few weeks (c23).

Another example related to relationships with universities, and evaluation data which surfaced questions about alignment of activities across the collaboration, "we only thought we were going to do a few interviews. We decided to dig a little deeper. That became an important – we were not thinking about this" (c22). A third example had to do with an equity issue related to the program design: "when we started surfacing the data... [it] became a big issue, the... foundation took action and developed a solution for students" (c22). In this instance, the evaluation was soliciting perspectives of key actors early on, while they were still upstream of program implementation. The issue was therefore brought forward in time for actors to respond with design changes (c22, c21).

Demonstrating a commitment to learning together. As a new approach, stakeholders first approached DE with uncertainty, and with varying degrees of interest, caution and skepticism. Developing productive working relationships across the collaboration were a focus of sustained effort. The responsiveness of the evaluation and its ability to support successful

adaptations to the initiative, and particularly the foundation and other donors' willingness to resource these adaptations, were said to have demonstrated a genuine commitment to learning together. The program reconfiguration in Year 3 was offered as one example: concerns of school staff were raised for leadership attention by the evaluation. Mechanisms provided by the evaluation enabled collaborative development of a solution, which was supported by the foundation and other donors with the help of evaluation data. A participant related this to "creating this sensibility... that we're about learning together, and that we're about co-creating and that we're about being transparent, and we're about being adaptive" (c21). Arguably, demonstrating commitment through taking action was a factor in keeping actors engaged with the evaluation process (c21, c22, c23).

End Stage

Shifting to formative

The evaluation shifted to a hybrid DE/formative approach beginning in Year 3, becoming mostly formative in Year 4. As noted above, the evaluation questions became more focused, with deeper inquiry into areas for improvement. Communications became less frequent (c21). "We still catch up with them and have conversations with them, but now that it's morphing into more of a formative evaluation, we wait 6 months to deliver findings" (c22). The data collection and feedback cycle, formerly distributed through the year, became more concentrated in 3-4 months at the end of the school year (c23). The system map from Year 1 was 'recalibrated' in Year 2 based on learning and changes that had occurred (v12, c23), but by Year 3: "the purpose had been served. I think people kind of got the coherence that we were looking for... and then we just didn't need it" (c23).

By the end of Year 4, a number of outcomes of the evaluation appear to have emerged. These included: development of norms in the team around data use; changes in perspectives about evaluation and around key components of the intervention; instrumental changes to the intervention based on evaluation findings; and sustained public support. Participants also describe learning acquired through the experience about a need for patience in the process of developing an intervention.

Habits and norms

A participant spoke about establishing norms in the group that helped to create space for conversations. Although some actors were thought to be still operating in ‘compliance mode’ in other areas of their professional lives, for this collaboration: "I think what we got to was at least setting a norm for the conversations we had with them... having a conversation about implications based on data in a productive non-defensive way, we managed to carve out a space to have that" (c23). This participant described ‘norm-setting’ around evaluative inquiry as part of the intent of DE and a reason for DE to be long term, over multiple years, because establishing norms is a ‘longer term endeavour’ (c23). Documentation points to the process of the evaluation (e.g. co-creation of the design, participatory interpretation of data) as responsible for strengthening working relationships and increasing actors’ capacity to use evaluation (d, also c21, v12).

Changes in perspective

Perspectives about evaluation, and about relationships. Participants mentioned process use helping to change team members’ perspective about evaluation: “the slow gradual shifting... moving from being really sort of fearful and apprehensive about this process to embracing it – that... was because of the process, that wasn’t because of the findings” (c21). This participant elaborated:

think of what would be an alternative? [A]n evaluation with less, I don’t know, more *doing to* than *doing with* - so through DE you’re doing with, right? Formative, summative may be done to. The relational part of that, that was supported by this approach, has also benefited our work together and our ability to work together (c21).

When asked if the evaluation was able to increase levels of trust, a participant was definite: “Oh yes, oh yes, oh yes... [although] it took a lot of time“ (c22). Another participant described increased trust related to how actors responded to the learning memos, which changed from an anxious response to the memos being used as “an actual tool to help them think and do better” (c23):

I think we did build increased trust over time. For instance... the learning memo that we did after every visit... I remember in Year 1 the initial conversations were like, ‘what is this thing really, like is this going to be in the newspaper tomorrow?’ To like ‘oh when is the next learning memo coming? I’m looking forward to it’. So yes, over time, this took 2 or 3 years, but over time [it was] ‘oh, when are you guys bringing your next thing? I’m looking forward to what the insights were from your visit’ (c23).

Perspectives about aspects of the intervention. Broadly, documentation describes changes to actors’ perspective on the nature of complex problems and how to address them, in particular their ‘multidimensionality’ (d0, d23). The evaluation was also said to have changed how actors conceptualize specific aspects of the intervention, for example by clarifying and ‘contextualizing’ how different stakeholder groups interpreted the meaning of ‘college culture’ (d0, c22). The evaluation was said to have influenced actors’ ‘strategic thinking’ about the program. Changes in strategic thinking in turn were said to have influenced implementation of the initiative (d0).

Instrumental use; strategy changes and program changes

Examples of instrumental use of evaluation findings have already been described. A participant said: “we’ve had incredible findings that have been harvested from the DE that have absolutely shifted how we’re doing our work” (c21). Examples include the rapid development and launch of an awareness campaign at a crucial moment in the intervention’s implementation (d, c21, c23); the program reconfiguration in Year 3 (d17, d19, d23, d28, c22, c21); investment in affordable housing initiatives (d26, c22); and planning for extra support to students whose progress data indicated a need (d23, d20, v4). A foundation document asserts: “With [the evaluators’] guidance, we know we’re asking the right questions, gathering the most timely and relevant results, taking ample time to reflect and reason; and making informed and insightful decisions about the future. It’s intentional, purposeful work that’s changing lives, not merely crunching data” (d34).

Sustaining public support

The evaluation work is highlighted in annual reports and other public documents about the intervention (d3, d9, d23) and on the foundation's website. Reports mention the partnership with the evaluation firm, information about the DE approach, and also include progress data. As an example, one foundation report says: "Unlike traditional evaluation, which is linear and fixed, developmental evaluation empowers us to be adaptive. It is a deeper dive into what makes a complex, dynamic initiative like [this] succeed and thrive. It analyzes and anticipates the 'what's next' to ensure student success" (d3); and: "You've put your funds and your faith in [this program]. We take our role as stewards of your support seriously, and we promise to be transparent with you. These five data points inform how, year after year, we will measure and share our success" (d3). Progress data are presented in media reports about the program (e.g., absenteeism data; d31, d8) and other documents from external organizations that mention the intervention (e.g., d35), indicating it was widely shared. The evaluation is sometimes called 'evaluation research' or 'developmental research' in public documents. Data were also presented to the advisory group representing community organizations, parent groups, donors, local businesses and others. These annual meetings were described as 'not just PR sessions':

our willingness to be transparent and our willingness to have that dialogue with a lot – there are 60 people in the room - ... it's a really diverse kind of group. And I think that's also not typical ... it's not like we're up there saying this is going great. There are a lot of parts that aren't working very well and we're still trying to figure that out (c21).

This participant provided an example of following up with a donor after one of these sessions, who communicated that they were 'so excited' by the advisory group session. The ability of the intervention team to garner public support is suggested by the sizeable funds raised for the program over time ('exceeding expectations'), the team's ability to make program changes that required partner support, and the consistently positive tone of media reports, even when data suggested slow progress on some indicators (v9).

Learning patience

Some constraints on evaluation activities and use of findings are suggested in the data, relating to issues of trust and highly political operating environments for some actors. Although the evaluation questions shifted over time, the focus remained largely outside of internal organizational operations, as intended, for example on parent engagement, student attendance and academic achievement, and emerging issues such as housing costs. Where evaluation findings suggested changes, some findings were actioned more than others, related in part to the sphere of operational control in which change was indicated, differences in capacity to act, and because of the ‘dynamics of the situation’ (c23).

The three interview participants expressed some regret that faster progress could not be made with the intervention: “I think developmental evaluators are like, ‘on it, let’s make changes, let’s do this’ – and it takes longer for people to implement. They have to deal with structures, they have to deal with budgets” (c22); “I feel we would have gone so much farther, you can only go as fast as – this is a huge leap for [some]” (c21). DE was described as sometimes being ‘intrusive’ (c22). Some actors were described as pushed ‘outside their comfort zone’ (c21), and it was recognized that the distance that different people need to travel to engage with DE differs (c23). “I think it takes exceptional leaders and different thinkers to break out of it, and it takes time to build trust” (c23). “[I]t teaches you a little about patience and just some appreciation for just giving yourself some grace, I guess, and to others, and a little bit of wisdom around when to push, what to push around and so on” (c23, also v12).

This program brought funding to area schools (e.g., for enrichment programming) and is credited with increasing enrollment in area schools due to aspects of its design. These were schools which were previously experiencing declining enrollment. Increased enrollment translated to increased government funding (v11). These benefits constitute strong motivations for various partners to support the initiative, but they do not entirely explain actors’ engagement in the evaluation or how their engagement was reported to have changed over time. A participant described the overall process of the evaluation as a ‘glue’ that brought about a gradual shift from ‘apprehension’ to ‘embracing the evaluation’ (c21). Arguably, the data suggest a broad pattern in

which the evaluators intentionally built capacity for data interpretation and use through conversations (d0, v12), stakeholders agreed on important actions to develop the initiative, and these actions were seen as successes. The successes were instrumental to strengthening the collaboration, in turn enabling more productive conversations. Evaluation continues to support the initiative into its 5th year.

Case 3

This summary traces the course of a developmental evaluation. It outlines the initial stage, middle stage and late stage of evaluation work carried out over a 3-year period. Themes derived from initial coding and analysis are outlined with each ‘stage’ to illustrate how the evaluation progressed. However, most themes are relevant to aspects of the whole case.

History / Starting Conditions

The Program

The program was one of three concurrent place-based initiatives led by a foundation. It was designed to encourage residents of a small city to engage with the arts, on the understanding that engagement with the arts can positively affect how people relate to each other and their city, and that this in turn can contribute to positive systemic change. The program was also intentionally designed to generate learning for the organization through experimentation.

The program was made up of a series of arts-related projects (for example, hosted conversations, installations, performances) over a period of 5 years. During this time, the foundation collaborated with local and non-local artists, non-profit organizations, and funding agencies.

This was part of a “phase 1” of the foundation’s program development, described as an experiment in their first mode of working, in which they became embedded in a few small cities in transition (c31). Through the program, the foundation aimed to connect with the city’s transition: “and somehow begin to pull on the threads of what those conditions are for how the arts could have some impact” (c31, also d0, c32). A participant described the exploratory intent, saying that the foundation wanted to know: “if [it does] these kinds of convening and artistic

projects and sort of idea development, does that help arts flourish or help a place know itself better, and what does that lead to?” (c32)

There were no specific targets for outcomes at the level of this intervention. As described in a document: “[r]ather than expectations of performance on a well-defined outcome, the expectation [for this program] is for further development on something innovative” (d0). This perspective was thought to keep stakes low, and ‘create space for experimentation and adaptation’ (d0). Although the initiative was place-based, learning was also expected to have relevance beyond the local, to be applied to other interventions in other places, and contribute to the long-term development of the organization’s overall approach (d31).

The Foundation

The foundation was very young (about 3 years old) at the time the initiative was launched. The foundation is described as well resourced for its work, with limited need to generate revenue or secure grants from other funders. The foundation’s structure includes four intersecting groups: staff, a board of directors, a group of founders and a group of advisors. The foundation is unusual because it acts both as a funder, and as a direct implementer of its own programming (c32, c31, d0).

Learning agenda. From the start, the foundation had an explicit statement of purpose and guiding principles, in which elements of a theory of change are discernable (d35, c31, c32, d0). In various documents, the organization’s purpose is described as “to make the arts more central and meaningful in peoples' lives, in our communities, and in our societies” (e.g. d19). The foundation aimed to engage in ongoing experimentation and learning “to figure out what this means”, referring to their ideas about the role of the arts for individual growth and social transformation (c31), and to “discover [the foundation’s] evolving role in part through doing” (d35). The experiential learning agenda of the foundation is a central theme of this case study. It connects to the foundation’s direct involvement in programming for this case, in which foundation staff took part as on-site participants. As described by a participant:

we needed to be working on the ground, we needed to have our own stories to tell, we needed to be able to articulate an idea of how we thought that the arts could be

transformative and how it could connect to interventions in problems in society before we could become more of a voice for that (c31).

The foundation described risk as a necessary part of this learning and development, stating that as an organization, it would need to be: “eclectic, exploratory, catalytic, adroit, brave and humble. It may very well make mistakes, take roads that lead nowhere, and expose itself to being mocked. But if it does not take such risks, it will fail” (d35). An orientation to complexity thinking is also evident in the foundation’s early documents (e.g., d35) and this appears to have informed the foundation’s work throughout the duration of this case (e.g., d31).

Since its beginning, the foundation has held annual 2-day retreats for staff, board, advisors and guests to undertake ‘deep analysis’ and reflection on the foundation’s work (c31). These retreats represent a formal mechanism to support learning within the organization that predate the evaluation (c31).

Deciding to Try DE

There was skepticism in the organization about evaluation in the context of arts initiatives (see below, c31, c32). As an organization, the foundation had no prior experience with evaluation, although individual members of the advisory group had experience with developmental evaluation (DE) (c32, c31). DE was suggested by these members as a way for the foundation to further its learning agenda (c32, c31), with respect to how the foundation’s ideas about its principles and purpose translated into or were manifested in its programming, and the implications of this for its role and strategy as an organization (c32), as well as to capture learning about how the arts are meaningful for people, how the arts can be transformative, and to ‘tell the story about how the initiative could have impact’ (c31). Actors were interested in developing and articulating their theory of change (c32), and DE’s iterative approach seemed to align with the experimental nature of the initiative (d0, c31). This initiative involved collaboration with other funders who had some evaluation requirements, which was also thought to act as ‘a bit of a catalyst’ to try evaluation (c32). A DE consultant was hired to work alongside the organization for this initiative.

Initial Stage

Assessing Needs for Evaluation

Planning for the evaluation began with interviews and discussions between the consultant, foundation staff and members of the board of directors about actors' interests and needs for evaluation, and about the foundation's theory of change (d0).

Internalizing the Evaluation

The evaluation was structured to enable various actors to take on evaluation functions (d0). The consultant was described as a 'coach', who helped with evaluation design, instrument development, data collection and facilitating data interpretation. Program actors, including staff and collaborating artists, were also involved in the evaluation design, data collection and data interpretation (d0; c32).

The evaluation structure was described as an 'internal-external hybrid', in the sense that an external consultant partnered with internal foundation staff and other program actors to carry out activities (d0). However, 'external' was not understood to mean outside of the initiative. Instead, the consultant was "embedded in the projects with the artists" (c31). The intent was for the evaluator to be "inside of the project" (c31) so that:

the evaluation was part of and not separate from the artistic experiences. So, it wasn't this kind of like, weight on the side, it was how do we creatively try and integrate collecting information and doing analysis with people, often in a very participatory way as part of the experience (c32).

The role of the consultant involved building relationships with the artists so that they could work collaboratively to make the evaluation a part of their project (c32). This was said to relate to a concern that "too much evaluation that didn't feel like part of the [artistic] experience would actually be compromising to the experience" (c32).

Experimenting with Projects

As noted above, the initiative consisted of multiple artistic projects, each of them purposefully unique. Their uniqueness was intentional - to provide a variety of experiences from

which the foundation might learn (d0); they were designed as “experiments into the transformative power of the arts” (d31, also d06), with the city itself seen as a ‘laboratory; a place to experiment’ (d30, also d06).

There was minimal repetition of activities at the project level over time. Although a series of hosted conversations launched the initiative and informed the projects that followed (d05), the projects were otherwise not staged sequentially for later project designs to be build on earlier projects; the focus was on variation not on iteration (c31). For this reason, although the DE was positioned in the projects, its purpose was not to support development of the projects themselves, as much as to help answer questions at higher levels: at the organizational level relating to its role, theory of change and approach, and at the level of knowledge generation about the arts and social transformation more broadly (c31, c32).

Experimenting with DE

The uniqueness of the projects, and the desire to integrate the evaluation into project design, resulted in a series of semi-independent evaluations that were tailored to each project. In addition, each project was intended to have: “a different DE methodology to create a kind of array of possibility” for using DE (c32). In effect, actors aimed to experiment with DE in tandem with the project experimentation; to explore a variety of ways in which DE might be applied (c32). It was thought that: “having a diverse set of experiences [with DE] would provide a bigger set of options and ideas”, for the foundation and other arts organizations (d0), and could contribute to exploring and potentially developing a methodology or set of tools for learning from artistic projects, as a form of research on evaluation. This purpose had roots in the interest of a collaborating funder to explore options for evaluation in arts contexts (c31).

Aligning evaluation with the arts. The foundation’s approach to evaluation was also guided by a perspective on how some forms of systematic inquiry align (or don’t align) well with artistic endeavours. In a document describing its core principles and purpose, the foundation asserts: “artistic creativity embodies values and attributes diametrically opposed to the narrower concepts of efficiency and rationality that have contributed to the modern crisis, concepts

reflected in the prominence of words like objectivity, calculation, measurable, predictable, quantifiable, replicable, efficient, cost-effective, profitable, rational, and linear” (d35). The implication is that arts outcomes defy measurement through approaches and assumptions common to the field of evaluation. As stated in another document: “the very prospect of evaluation can be alienating to artists, producers, curators, and participants, because the inherent value of the arts cannot be evaluated [by] using the mechanistic and quantitative measures that are so frequently used [in evaluation]” (d0). A participant from the foundation explained: “there’s been this struggle in the field, in the sector for a long time, how do you evaluate arts organizations and artistic projects? What we’re trying to say is it cannot be the superficial measures of time and money spent. It needs to be about how meaningful those experiences are as well” (c31).

In spite of skepticism about what evaluation could do in arts contexts, the foundation was willing to experiment with DE provided the evaluation could be sensitive to and aligned with artistic endeavours. The need for sensitivity and alignment was about more than the risk of alienating artists and participants. It was also so that the evaluation could generate meaningful and rich insight for the foundation as an organization with a well-developed vision of art and the role of art in society (c32). How the organization and the evaluation consultant tried to make this work is described below.

Middle Stage

Tailoring to Each Project

The evaluation was carried out across five arts projects over three years. The evaluator worked alongside the artists to integrate evaluation into the design of each project (c31). The variety of methods used across projects, and how they were applied, speaks to the tailoring effort. Data collection methods included interviews, focus group conversations, open-ended questionnaire, journaling and observation. In one project, as an example, the evaluator

collaborated with artists to plan a one-day event at several sites in the city. Opportunities were built into the event's structure to allow attendees to gather together and share their experience through interviews, drawing and storywriting (d0, d15). This was described as a way to "layer evaluation into what was happening in a way that would enrich the program experience for participants" (d0). Specifically, the evaluation for this project was seen as a way to facilitate the 'sensemaking experience' of participants as they engaged with the event, as well as being an opportunity to 'capture information and perspectives' for learning by the organization later (c31, c32, d0). In addition, video taken at the event for the evaluation was later used for a documentary film about the project, "so it was kind of win-win on all sides" (c31).

Linking Projects Together

Although the project evaluations were tailored, they shared some common features, including an emphasis on documenting the experience, directly engaging participants at the project level, and crossing levels of inquiry.

Documenting the experience. Across all the projects, the evaluation work was described as important for documenting the experience of community participants and artists, in order to learn from these experiences. For one project, this was described as 'intensively documenting the entire process' of creating and carrying out the project (d0) and 'documenting the sensemaking that is unfolding in hope that it will bring changes in perception' (d33). For another project, this was described as aiming 'to document the collaborative process as a way of understanding its features' (d0). This suggests an effort to illuminate the 'what' at the centre of inquiry.

Engaging participants. Also common to all the projects were efforts to enable direct and meaningful participation in the evaluation by staff, artists and community participants. For participants from the community, this was said to involve, for example: "creating the right questions and ... creating the right environment in which those people can feel comfortable contributing" (c31). For instance, providing opportunities for people to communicate their experience with a project through various modes: conversation, drawing maps, writing stories, etc. (c31, d0).

The engagement of the foundation's directors and advisors with the evaluation was less direct and more limited. Early on, the evaluation consultant introduced DE (what it is) to these stakeholders at an annual retreat, spoke to specific findings, and provided input based on the evaluation data, by "representing the data in the larger conversation and saying here is what we're seeing, and thinking about and learning as we do this" (c32). The annual retreat was also seen as an opportunity for the consultant to learn more about these stakeholders' information needs so that the evaluation could be adapted to respond (c32). In later years, however, reporting at the retreats was handled by staff leadership (c31). As a small organization, the foundation was said to be able to accommodate less formal communication flows than can some other organizations (d0). As noted above, the foundation had an pre-existing method and structure for reflecting on its work, which was already supported by staff briefings on program activities through the year. These briefings were said to keep board and advisory members "quite up to date over the course of the year... so by the time we get to the retreat there's a lot of analysis that they've done in their own mind" (c32). Other guest participants are also routinely invited to the retreats for the foundation to: "get different perspectives" (c32). In sum, data suggest foundation actors did not see a need for the evaluation to help them develop formal learning mechanisms (sometimes reported as a process use outcome of DE). Overall, the evaluation was just one of a number of inputs to the foundation's reflection process at director/advisor level, mediated largely by staff leadership (d0, d31).

There were two other funding agencies involved in the initiative, however, they were not described as participants in this evaluation. This DE was directed at supporting the foundation, and its scope was maintained there (c32). Although it was noted that one of the funding agencies "had very different evaluation needs" and priorities (c32), and a different understanding of the initiative (c31), the tensions resulting from these differences were negotiated and addressed by foundation staff (c32, c31).

Crossing levels of inquiry. Although tailoring made for unique sets of activities which 'stood alone' at the project level, efforts were made to bring lessons forward to inform "an ongoing conversation about the overall concept and approach" (d0). For example, staff leadership were interviewed multiple times over the 3 years, as a way to facilitate cross-project

reflection and ‘sensemaking’ (c31). The one-to-one interviews were described as promoting ‘dialogic thinking’, which brought out new ideas for programming (c31). At this level of ‘overall concept and approach’ the evaluation was pursuing multiple purposes, from operational (i.e., program, organizational applications) to more abstract or theoretical (i.e., about the role of the arts in society).

Project to programming (operational). An example of a program level application supported by project level evaluation can be found in a shift the foundation made in its approach to collaboration. The foundation had started work in the city by connecting with local organizations that were said to be ‘key’ to a city-wide, arts-based initiative. However, levels of engagement were not as high as the foundation had hoped, and efforts to raise engagement were ‘burning up a lot of energy’ (c32). The foundation experimented by opening up invitations to a broader range of actors, and proceeded to work with those who came forward with sufficient interest and willingness to engage (c32, also d0, d33). The evaluation, already underway, was adapted to provide feedback on interest levels and efficacy of these relationships, to help with decisions about whether the foundation should approach selecting and fostering collaborations this way in the future and elsewhere (c32, also d33). The shift to a more emergent approach to collaboration eventually became part of the foundation’s general approach.

Practice to theory. In addition, the individual project evaluations shared an objective to develop a conceptual understanding of “how socially engaged artistic practices can be transformative in a community” (d0). This second objective was aimed at ‘bigger questions’ about the role of arts in society.

While the two objectives (operational/organizational and theory development) were said to be present for all three years (d0), the balance between these two appears to have shifted over time. Evaluation work for early projects was oriented more to program development. For example, the earliest work looked at the design and outcomes of an event (for example, reflections on design decisions, and whether / how the event influenced community participants’ later activities). Later, evaluation work was applied to understanding “effective collaboration and how this could be supported” in the foundation’s program approach (d0). Midway through the 3 years, the focus begins to shift toward the second objective, for example how participants

described their experiences: “to gain a better understanding of how the arts can be meaningful for people” (d0) and turning to the creative process of the artists, seeking, for example: “deeper exploration of themes and ideas relating to an artistic project” (d0), and “the contemplative experience of creation for the artist” (d0). In this sense, the evaluation was moving across what is sometimes considered the domain of evaluation, with its usual direct focus on program or organizational applications, into the domain of research (c31, c32). Notably, evaluation activities were scaled back in the final two projects (d0). This was attributed to the experimentation with different DE designs (d0).

Stretching the boundary of DE. Participants recognized that questions of interest to the foundation were leading the evaluation toward research. This was described as ‘stretching the boundary of developmental evaluation’ (c31, also c32), but this was not without awareness. Documents and participant comments suggest reflection about where a boundary lies. For example, a participant noted that: “DE is evaluation: ...critical thinking, systematic use of evidence, data, information, but in circumstances of an innovation in high complexity” (c32). In a document, it states that the intent of integrating the evaluation work with the projects was to be unobtrusive, and yet also to ensure that “evidence and critical perspective” and a “critical and constructive voice” would be brought to the program (d0). Working with the artists to make DE part of their project was described as reflecting: ‘principles applied in evaluation all the time about users and use’ but adapted to the arts context (c32). Another participant described “being able to bring [evaluation and research] in relation” as one of the successes of the project (c31).

‘Stretching the boundary’ helped the foundation pursue its questions of interest. It also reflected the interests of the artists. Asking questions that were of interest to the artists themselves made the work more palatable and was important to their participation (c31). This meant that they were “very keen to be involved [and] there was no difficulty engaging them” (c31). As described by a participant:

“[the artists] were quite engaged with this because they understood it as a body of research. If they understood it as us bringing an evaluator in so that we could rethink our organization or redesign our organization they would probably have been less interested in it. They were interested in the idea that it was research into how people will make meaning

of their work that they were experiencing, so they were very open to having an evaluator participating in the conversations... It wasn't that we were trying to evaluate whether their project was 'good' or 'bad' or should have been done differently, it was more research into the meaningfulness of that project, what it means to participate in that work" (c31).

The "rhetoric of research" was said to have been used more than 'evaluation' (c31) and this is evident in foundation documents in which the evaluation work was referred to as research or 'evaluation research' and the consultant at times referred to as "our researcher" or "researcher, listener, observer" (e.g., d15).

Although important for engagement and for answering questions of mutual interest to artists and the foundation, two challenges emerged from stretching the boundary of DE.

A tension in trying to do both. The foundation was said to be "closer to the research end of the spectrum in their interests and needs" (c32), which was echoed by another participant who described the foundation being primarily interested in "the big questions" about the meaningfulness of the arts experience (c31). A foundation document produced after the evaluation was over described the 'meaningfulness' question as the "central question that we aim to answer with our artistic experiments" (d31, also d30). However, the evaluation tended to lean toward the organizational level questions, such as: "how this was working for [the foundation] as an organization" (c31). The tendency for the evaluation to focus at the organizational level was attributed in part to the 'methodology' of DE, designed to provide rapid feedback to inform operational decisions, and to organizational development being the usual 'mode' of working for developmental evaluators and evaluation (c31). The foundation's desire to "think and work at more of a cultural or systems level", and to answer these larger questions, was said to be challenging for evaluation (c32, also c31). The difference was described as a push and pull in areas of interest. A participant noted that: "there was a tension between trying to do both, and that I think is what limited us in some ways" (c31).

Linking it back. The other challenge (ironically) came with trying to link what was learned back to applications for the organization (c31). The 'linking back' was described as having been done by foundation staff themselves, more than through a 'final analysis by the

evaluator of how the organization could use [the information]' (c31). It was suggested that there was difficulty 'bringing it all together in a final analysis', which may have been because the project stretched so far outside the core strengths of evaluation and became "unwieldy" for evaluation to manage (c31). However, it was also noted that the 'dialogic thinking' and reflective work that took place via interviews with the consultant during the evaluation process were important to that 'linking back' by staff (c31). This was described as a: "process of questioning" that brought up ideas for staff; that 'forced them into a frame of thinking' that led to developments in programming (c31). Although the 'link' to organizational applications was not straightforward or direct, this indirect pathway supported staff to generate ideas and draw their own conclusions about potential applications.

Capturing the nuances of organizational thinking. Over the course of the evaluation, the foundation was reported to have developed its theory of change and its understanding of its role as an organization. Actors were described as 'really keen' about doing this, from the start of the evaluation work and throughout the 3 years (c32). Early attempts to create a diagram (as TOCs are often constructed) were not successful, however. Although most organizations tend to prefer concise (single page) visual TOC presentations, for the foundation, "visual diagrams failed to capture the nuances of the organization's thinking and experiences" (d0, also d32). The foundation was said to "really struggle with this format" (c32). The effort turned to expressing the TOC in narrative form. A long-form (essay-style) format was chosen. This was described as "very unique" for a TOC, but illustrative of how effective communication depends on an understanding of "an organization's culture and how it thinks and what it likes" (c32).

Negotiating differences. Negotiating differences in perspectives, interests, and organizational cultures across a community is difficult work. A number of the 'turning points', challenges and opportunities mentioned by participants over the course of the evaluation relate to challenges in collaborative work among organizations involved. This is important for an organization that describes itself as a 'convenor' in its founding documents, intending to "take a transdisciplinary, cross-boundary, cross-sectoral and inter-generational approach to its work" (d35, also e.g. d27); and as a 'broker' by Year 3, seeing itself as "neither inside the system nor outside", which gives it "capacity to be mutable and to build bridges" (d30).

Collaboration was described as easiest and most successful with the artists involved in the initiative (c31, c32). However, it was more difficult elsewhere. For example, the foundation was an outsider to the city in which this initiative was based, and found it difficult to gain traction there among local community organizations and residents. This was thought to relate to aspects of local culture and the initial collaboration approach (c31, d0, d33, d07). The foundation struggled to find common ground with a partnering funding agency on directions the initiative should take, in part because the foundation's approach to programming is unconventional (e.g. d07), and the complexity of the work is challenging to communicate to those working in other contexts and with different underlying assumptions (c31). This foundation is described as a 'very unique' organization, with a firm commitment to honoring its purpose, principles and world view in its work, which makes for "negotiable and nonnegotiable aspects to the design of [its] projects" (d30, also c31, c32, d25). As noted above, collaboration strategies were an emergent topic of inquiry for the evaluation. The Year 3 document states: "the Big Question is whether or not our communities and collaborators are equipped and motivated to work with us and to continue after we leave" (d30). Ultimately, it was said that the foundation realized that it was important to engage with people who are committed to its mandate and "way of working": "Which is why we now work a little differently and allow the communities themselves with the resources we bring to it to determine themselves how they want to engage" (c31).

Within the evaluation, much of the collaborative work was aimed at preserving the integrity of the 'non-negotiables' by adapting evaluative work to the artistic projects. It was also aimed at finding common ground between evaluation's strength in serving operational decision-making and the foundation's other interest in answering more abstract or theoretical questions. Some success was achieved here. Ultimately however, the foundation moved away from developmental evaluation to research so that it could pursue its core questions of interest in greater depth, as discussed below (c31, c32).

End Stage

The evaluation ended after Year 3. The foundation also ended work in this city about a year later. Although the foundation chose not to continue with DE or in this city, the evaluation was said to contribute to ongoing work in other cities and with other types of projects.

Giving Insight, Prompting Thinking

As noted above, the results of the one-to-one interviewing with staff over the course of the evaluation were described as a valuable contribution. These interviews were said to have used a ‘depth of questioning’ that brought out ideas and ‘prompting a particular way of thinking’ which was described as one of the ‘intangible things’ that the evaluation added (c31).

The evaluation also contributed to changes in approach, for example by informing the foundation’s decision to adopt a less ‘centralized’ way of decision making and ‘driving’ the projects, and moving to “building a local capacity” in their work in other cities (c31).

we did that shift from more centralized approaches to decision making – which artists do we want to work with, how do we want to work there and such – to more of a platform based approach in which we create the container in these cities in which the local communities, local cultural organizations and artists can kind of determine themselves how they would like our program to be manifesting in their city. That is a result of this evaluation largely. [It] showed us that when people take more ownership and have more of a role in determining what is happening in their place, it has more meaning, it has more power for them (c31).

Learning and Articulating what it Does

Developing its theories of change. The foundation was interested in developing its theory of change from the beginning of the evaluation. The evolution of its theory is articulated in annual reports from the beginning of Years 2 and 3, as noted above. After Year 3, the evaluation is referenced in a report describing what the foundation believes about art and ‘the Meaning of Art for Individuals’ (d31); and in a document about the evaluation which states that the evaluation contributed to: “an improved conceptual understanding and an increasingly clarified change model” (d0). These were said to have evolved out of “[t]he critical thinking, engagement and careful observation of its activities during the developmental evaluation” (d0). The evaluation was also described by a participant as having been part of the foundation’s “story making” (c32). Data suggest that two interrelated theories of change were developing: the foundation’s theory about how society may change through the arts, and an internal theory of

change about the organization's own 'way of knowing' and becoming as an organization (c31, d30).

Theory of approach to social change. Foundation documents at the start of Year 2 describe the organization having: "a clearer understanding of how these concepts [articulated in their guiding principles] can be manifested" in their social interventions. A number of cause/effect and relational propositions are presented that "we are now ready to say [about our] approach to change" (d33). A document from the start of Year 3 states: "now we are in a position to tell our story to the world" (d30). Included in this document are statements about 'why we exist' (e.g., "to be a catalyst for renewed optimism"), how the arts can be transformative, what would inform an 'exit strategy' from an initiative, and considerations about time and trajectory of change. With the respect to the latter, the document notes that: "change comes suddenly at times and in unexpected ways. [The foundation] should acknowledge this, and perhaps even relax a bit; a too-intense search for change can stifle the art" (d30, also d15). These documents were outputs of the staff, board and advisory annual retreats, at which the foundation reflected on its work. As noted above, evaluation results served as one of the inputs to these discussions.

In a document describing the evaluation, the evaluation was said have helped the foundation to establish guiding values for its work. Seven guiding values are listed. For example, the evaluation provided: "[c]onfirmation that art instills in people the possibility of change, of hope; specifically, it sparks in people the possibilities of living with uncertainty, imagining something beyond present conditions, transforming, and starting again" (d0). This same statement is repeated in other documents from the foundation, including one from the beginning of Year 3 (d30) and a publication from the year after the evaluation ended (d15). In the former, the statement is included in the answer to the question: "What is [our] narrative?". This report states: "We have touched on this during previous retreats but have never really formulated an answer... this element of our narrative [that art instills the possibility of change] parallels the goal to make the arts more central and meaningful" (d30). The shared language between the documents about the evaluation and the foundation draws the link between the evaluation work and the foundation's developing narrative of purpose and change.

Theory of learning and development. The foundation also appeared to advance a theory about its internal development (d32). Its commitment to being a learning organization was evident before the evaluation began (e.g., d35), and is articulated as part of its identity after the evaluation ended. Although its commitment to learning was pre-existing, the evaluation was said to have influenced the organization and the trajectory of its work at this level. A ‘shift in the organization’ was said to have occurred following the evaluation, to “rethinking its own way of knowing” (c32). This links back to its approach to social change. Describing itself as an “emergent organization”, the foundation states: “Our work is a continual response to what we experience and to what we learn, like a snowball gathering layers with its momentum” (d32).

The foundation was said to have ‘internalized developmental evaluation’ following this DE (d0); other accounts suggest that the foundation continued with multiple forms of inquiry, not DE per se. The foundation had hoped the DE would create: “an aggregate of data on how meaningful peoples’ experiences are with art, so that we would be able to have that deeper analysis, in order to tell the story then of how the arts could be transformative” (c31). In the end, a ‘story the DE told’ was that the foundation needed to engage for a longer time and more deeply with participants, in order to help participants to develop their capacity for sensemaking, and to elicit deeper insight from their experience (c31, also d0, c32, d30). Described as a “logical development” from the DE, actors recognized that pedagogical tools would be needed (d30). In Year 3, the foundation writes about plans to create a toolkit as part of its ‘pedagogical work’. Part of this toolkit would be a set of “workshop designs, sense-making exercises, discussion questions, essays on the value of the arts, and interviews exploring the collaborative process” (d30).

After Year 3, the foundation transitioned from DE to a program of action research in other cities. The research was described as a new experiment in how to learn (d32, d0, also d27), as a direct outcome of the evaluation work (d32, d30, also d20). In particular, a participant said, “the experience of doing the evaluation taught us that we needed to take people on a much, much longer trip with us before they could articulate more clearly” about their experience (c31). Short interviews with participants during project evaluations had generated interesting data, but “the [arts] experience had just happened, people hadn’t processed it yet and they didn’t have

necessarily the powers of articulation that would be needed to do a thorough rhetorical analysis” of their responses (c31). The evaluation was thought to have:

revealed to [the foundation] how challenging it is for people to make sense of an artistic experience, and that that is a learned practice and there’s capacity building that can be done around that. I think that discovery led them, because some people were really strong in how they were able to make sense of it, and this was not about they’re drawing the wrong conclusions or right conclusions, I don’t think there was an expectation that there were right or wrong conclusions, but I think the depth of the analysis or the ability to be thoughtful about those experiences really ranged in how people were able to do that. That led them to develop a program out of that insight where they really worked with people over time to help them discover that and develop that in themselves (c32).

The projects in this new program were designed to extend over several weeks (or months). The first iteration:

moved progressively from just observing the city to engaging more critically with the city ... The point was that what evaluation showed us was that we needed to have people, if we really wanted to understand how they made sense of an aggregate of experiences, we needed to have them embedded in a concentrated way with a small group, a small sample of people that we could much more meaningfully engage with our researcher... To me that wasn’t evaluation any more that was really pure research (c31).

Participants and documents report the foundation learning and developing through the evaluation work (c31, c32, d0), but ultimately not able to go deeply enough into their questions about meaning and sensemaking of arts experiences by community members (c31). There was less interest in ‘purposeful iteration and experimentation’ at an operational level – where strengths of DE lie (c32, 31).

Continuing the Journey

Two years after the evaluation ended, a foundation document notes:

At times we glimpse evidence of the transformative power of the arts—we find great stories to point to—but measurement as a form of evaluation is a social construct. We need to create an alternate vocabulary and value system with which to discuss the immeasurable impact of the arts... We must talk about the arts in ways that aren't reductive... The value of the arts is not in answering questions, but in inviting people to ask them (d32, also d27).

Around this time the foundation transitioned to another programming stage, in order to “expand the ways that we’re considering the transformative power of the arts”. This expansion suggested a need to move beyond single city programs (c31) and to document and research the many and various ‘authentic experiences’ of art (d31). The outcomes remain open, in the foundation’s words: “we strive to create the conditions for everyone to experience the possible” (d15).

Case 4

This case summary covers the first 9 years of an ongoing developmental evaluation. The summary is organized in rough chronological form corresponding to broad stages of the evaluation effort: its starting conditions, early stage, maturing stage and current stage. Themes derived through initial coding and analysis are briefly elaborated with each ‘stage’ to illustrate how the evaluation unfolded, however many themes are relevant across the whole case.

1. Starting Conditions/History

Description of the Program

The initiative is a well-established international research and development program led by a North American foundation. The program provides funding and capacity building support to research projects in developing countries, with a focus on local agricultural innovation. The initiative operates at three levels: the overall program (global), regional communities of practice (multinational), and individual projects (local).

The program has been in existence for over 20 years. It is described as having a “participatory and bottom-up nature” (d05). It has an established practice of cultivating long-

term relationships with grantees, and providing sustained funding to research projects, with a view that impacts may take a substantial amount of time to accrue (c41, d05). The investment made in relationships by the foundation is said to be an important part of its approach and influential to its success (d05).

The lead foundation is described as open to experimentation, with “innovation in its DNA” (c42, d0). The foundation is also said to have a long-standing commitment to learning from people who are “working on the ground” (d0, also d13, c42). It is said to have a “dialogic, consensus-oriented, inclusive decision-making process” (d0) with a commitment to accumulating, over time, a knowledge base and systems to support its work (d13, c41).

Evaluation history. Prior to the period covered by this case, evaluation efforts were limited to some annual documentation and periodic, retrospective reviews (c41, d29, d0, d05): “there were limited frameworks, little reflective consciousness, and few formal protocols for sharing findings” (d0, also d05). Past evaluation studies had recommended increasing the program’s focus on “authentic farmer involvement” as a means to increase the context-sensitivity, relevance, use and impact of research outputs (d05). It was also recommended that the foundation seek to bring “more coherence and structure to the program” (d05, d29). Although the foundation’s leadership had considered these recommendations, major changes had not been made (d29).

Limiting factors. The data suggest at least three factors may have limited evaluation efforts. First, the foundation was concerned about making evaluation requirements onerous for grantees. A participant described this as a dilemma that is shared with other foundations, who “don’t want to be a burden to their grantees ... and yet they know that they need something, and so they struggle with what that should look like” (c41). Also, some of the program leadership were thought to be uncomfortable with the idea of evaluation and/or have had negative past experiences with evaluation (d29). Finally, although aggregating learning across projects was thought necessary for the program to have greater impact (c41), the program had to find a way to balance pursuit of ‘global’ learning with a strong commitment to “place-based knowledge” and “decentralized innovation” (d0, also d41, d05, d29).

Inflection Point.

The first year covered by this case (Year 1) was a time of a significant transformation, described as “an inflection point within the foundation and the program” (c42, also d0). An influx of grant funds from another foundation led to a program expansion, doubling the program’s scope (d0, d29, c42, c41). The new funding allowed for investment in new activities, including evaluation (d05, d31). The new grant also introduced new reporting commitments. These commitments involved sharing quality data on program activities and results (d29; c42), which may have encouraged the foundation to go more “data deep” (c42, d35).

The foundation’s president had come to the organization about two years prior. They initiated a restructuring of program leadership around Year 1, including hiring a new program director, expanding regional grantee support teams, and transitioning an existing oversight body into an advisory committee with new members (d05, c41). The new people and program restructuring in Year 1 introduced new perspectives and new ideas about the program, described as: “[an] increase in the diversity in understanding of what the program should be about and how it should be implemented” (d0). This was also the first time the foundation had itself assumed the role of ‘grantee’, and there were questions about how any differences in vision or perspective between the two funders would be successfully navigated (d05). At this same time, the program was beginning to focus more on developing regional communities of practice (d02, c41), which was intensifying the program’s interdisciplinary nature (d05).

The new program director was chosen in part for having an evaluation background (d05, d41, d42), ensuring (by intention) a prominent role for evaluation in the program (d29). A priority for the new director was to establish an evaluation system (d05). Within months the director was recruiting a team to work on design (d0).

2. Early stage

Designing an Evaluation System.

Bringing a frame with them and having It reinforced. The program director brought experience and interest in participatory, utilization-focused and systems approaches to evaluation (c42, d29). Conditions within the program reinforced their view that a systems-oriented approach would be appropriate, and in particular, one that would draw on complexity theory. Conditions included the “complex environment” around the program, the multiple concurrent changes underway within the program, and the need to build program coherence (d05). Prior to joining the foundation, the director had studied human systems dynamics (d0), and been thinking about systems approaches for complex situations (c42). Then:

[a]s I settled into the new job at the foundation, I clearly saw that all the transition in [the program] called for a systems-based, emergence-oriented approach. The program was changing even as we were trying to design the evaluation; it was a bit like building a bicycle while you are riding it. I could see that we needed to introduce more coherence and order - not through heightened bureaucracy, but through policies and processes that allowed us to remain flexible, reflective and responsive to change (d0, see also d05, c41).

To design the evaluation, the director recruited consultants who shared a background in complex systems approaches (c42, c401, d0). The intent was to design an evaluation system to support the program over a long term with evaluative thinking and data, not to design a single, time-bound evaluation study (d0). In addition to systems and complexity thinking, commitments to utilization and collaboration underpinned the design.

Prioritizing utilization. A member of the foundation’s leadership commented that the evaluation framework “began with the key notion that the only evaluation system that is worthwhile is one that is actually used”; and “we wanted it to be useful and to obtain information that we knew how to use” (d30); use was understood to include use of the evaluation process as well as its results (d29).

The position taken to data collection reflects a utilization focus: documents assert that data collection should be focused on “priority questions and uses” (d07); having fewer data, if

ready in time and in a useful form, was described as a better use of resources than collecting more, higher quality data but being unable to analyze the data in a timely way to foster use (d0, d30). The description of rigour in a program document also reflects this focus (d07). Rigour is described as relating to clarity, transparency of purpose and methods, context-sensitivity, and ‘meticulousness’: “[r]igor is not about using a specific method, but rather choosing the appropriate methods to support the claims that are to be made, and implementing those methods with fidelity to the situation” (d07).

Users and uses. Although external uses for the evaluation were noted (e.g., reporting to the partner foundation; potential for ‘global’ learning of value to the field (d29)), the primary emphasis was on internal use (i.e., by program actors) for learning and internal accountability. A foundation member spoke to the internal focus of accountability: “even as our strategies and tactics shift, our core intent will not change. To make the most of the private resources we administer for public good... we have a responsibility to set goals, to measure our results and impact over time, and to hold ourselves accountable for our work” (d30, also d29).

Getting there collectively. A commitment to multi-stakeholder collaboration in the evaluation was present from the beginning. The program director’s background in participatory evaluation aligned with the program leadership’s interest in collaborative processes (d29), and was consistent with the foundation’s approach to decision-making and the program’s way of operating (above). The rationale for stakeholder participation in the evaluation was multi-faceted, including both methodological, relational and ethical aspects.

Methodology. Collaboration was said to be: “essential to the success of [the evaluation] methods” (d07). The program was viewed as complex. From a complexity perspective, multiple viewpoints from different parts of a system are important to increase understanding (d07, c42). Evaluation documents note that participation by stakeholders increases ownership of the evaluation, which is important to encourage its use (d07, d30), to increase the quality of data (d30), and its interpretation. It: “gives you deeper - it creates more authentic meaning” (c42, also d07).

Interdependency and ethics. An interest in reciprocity and a sense of interdependency with grantees are referenced in documents (d29, d07). As noted by a member of the foundation:

The evaluation framework... begins with what we call the evaluative relationship... If the relationship [with grantees] is good, we'll have a clear and better understanding of how our missions align and how the work is going. If the relationship is bad, we'll have less certainty and trust in the information we receive... We're dependent on the groups that we support" (d30).

Added to this, a document asserts that: "[p]articipation in this [evaluation] process... has an ethical dimension, because people have the right to be consulted and involved in activities conducted on their behalf" (d07). A participant explained:

we *are* committed to co-creation, and that means we all bring something to the table to help understand better... for me it's kind of a validity / reliability thing but it's also a commitment to honouring all different forms of knowledge" (c42).

These commitments took shape for example in various efforts to make evaluation activities participatory (see below), and to share results with project grantees as well as 'upward' with program staff and leadership (d29).

Adopting an adaptive action framework. An adaptive action framework was chosen as a complexity-informed and participatory approach to learning for program development (d05, c42, d07, d29). It was felt that the framework would help build program coherence and "instill a data-informed decision-making process" that could respond to a dynamic context (d0). The team proposed six-month cycles of data collection, analysis and interpretation. Each cycle would incorporate data collection about what was happening ('what?' questions), analyzing and interpreting these data in terms of their implications (asking 'so what?') and systematically establishing next steps ('now what?') (d0, d07, d05, c42, d29). The framework was to operate at a project, regional and program levels (d0, d29), "with lower level results feeding into upper level conversations" (d0). Regular structured reflective practice sessions of the various teams, called 'generative dialogues', were to be a key mechanism for operationalizing the framework in the initial design (d29, d0).

Anchoring the Evaluation System.

Bringing the envisioned evaluation system to life was challenging for many reasons. The ability to sustain the work when challenges arose was aided by some starting conditions, including the foundation's 'innovative culture' and willingness to commit time and resources to develop long term solutions (d13, also c42). Also important at the start were some characteristics of the team and their position within the organization.

Team characteristics. The program director had a prior relationship with the consultants hired to design the evaluation. This gave the group a relational foundation on which to build a team (d0, c41). They shared a common language with respect to systems theory and had technical expertise in systems approaches, which they could combine to give them a flexible 'toolbox'. They had strong motivation for the work because of a shared interest in advancing systems approaches to evaluation (c42). For example, the director and the lead evaluation consultant: "shared similar dreams" (d0) in this respect. Two members of the team were recognized experts in systems thinking, one had also performed evaluation work in the past for the foundation, and was known to the foundation leadership (d36). These two members brought significant profile and credibility.

Position in the program. The director's position in the organization gave the team an internal champion to help anchor the evaluation. As described by the lead evaluator: "not only was the program a perfect fit for a systems-based evaluation, but having a program director who didn't need convincing and was fully on board felt like a gift" (d0). The importance of having internal "allies" (d0) for the evaluation will be revisited below. Structurally, the evaluation team was made part of the program leadership team (along with the program advisory committee, regional teams, and research methods team) (d0, d22, c42), which helped anchor evaluation in the program.

Implementing the Evaluation

Exploring. The team began with substantial exploratory work to understand the program's purpose and theory (d0). A participant described interviewing the members of the advisory board, the program staff and regional team members, and attending regional meetings. The findings reinforced the need for more coherence:

we'd ask people...: 'what is this program about?' And ... everybody had a different point of view on what was really going on, so really taking time to figure out what the focus of the program was, and how it all fit together, and in the end what was their theory of change, how do they think change happens? Getting really, really clear about that (c41).

In an exploratory phase like this, a participant stressed the importance of 'holding ideas loosely', described as: "listening and watching and holding any kind of structure, whether it's the set of goals and objectives or whether it's a program mission statement or vision, holding it loosely knowing it may not actually be what's happening" (c41). This appears to have been particularly important given the changes underway in the program and arrival of new stakeholders.

Encountering tensions. The team chose the evaluation approach based on their assessment of the complexity of the program and the evaluation's priorities, such as need for coherence, but the approach was unfamiliar to many program actors. These actors brought together a range of perspectives and inquiry cultures, especially from disciplines involved in agricultural research (d0, c41, c42, d05). In addition, at the leadership level, there were persisting differences of opinion about the most appropriate form and function of the proposed evaluation (d0). These were influenced by some of the tensions noted above about how evaluation (generally) might fit with the foundation's philosophy and approach (d29). Consequently, the evaluation's implementation was met with resistance. The team sought to navigate this by experimenting and adapting to stakeholder needs.

Experimenting. *Finding a balance.* In the early phase, a participant described experimenting with different methods and tools, trying to find a balance in the evaluation between structure and flexibility, so that the evaluation work could accommodate the heterogeneity in the program: "trying to work at the overarching level to really pin down some kind of framework [for the evaluation] that was loose enough to incorporate all of the differences [in the program] but tight enough that we could do something with it" (c41). This involved: "[trying] out these little pieces to see how they fit" (c41). For example, at the regional level, they

experimented with tools to understand how capacity building was progressing, and how people were networking. One tool was abandoned, because it did not work well for the people involved. Another worked “fairly well” and was developed further (c41). “Looking back, I appreciate the process, and I see that the trail-and-error approach was necessary for maneuvering the complexity of the process, and the diversity of people involved” (d0). The need to balance structure and diversity was a running theme throughout the data for this case.

The introduction of the ‘generative dialogues’ provides an illustration of how the team experimented with and adapted methods over time.

Piloting and pushback. The idea of ‘generative dialogues’ was piloted at a regional meeting in the fall of Year 1, only a few months after the evaluation team first came together. Meeting participants did not respond well. The approach did not meet their expectations for evaluation (d0, d41) and the language that the team was using lacked credibility: “people heard ‘generative dialogue’ as empty jargon”, and responded with: “*This* is how you are going to evaluate [the program]?” (d0). The pushback was attributed to a distance between the proposed approach and the existing inquiry culture, combined with a lack of trust in the evaluation team (as newcomers) that might otherwise have helped to bridge that distance (d0). Moreover, the lead evaluator had been hired for expertise in evaluation and systems approaches, but did not have a background in agriculture. As a result: “those who had lived, eaten, and breathed agriculture over the course of long careers were skeptical” (d0, also c41).

Trying again. The team developed a detailed guide to generative dialogues for a leadership meeting a few months later, early in Year 2 (d29). A cool response was again attributed to language that was unfamiliar and lacked currency, and discordance between the approach and expectations for evaluation: “it was difficult for some to buy into the idea that [generative dialogues] was supposed to be a core tool for program evaluation” (d29).

The team came to this meeting having also completed a cross-project analysis. With respect to this analysis, meeting participants voiced concerns about inadequate consultation (d0). Moreover, the assessment was aimed at building coherence and alignment among levels and with

program objectives, and although the team's exploratory work had reinforced this as a need (d08, d290), building coherence was not universally accepted as a priority. A document account states:

[w]hile the foundation felt a need for greater coherence, the members of the regional teams who had been working on the ground for the past few years did not feel the same need. For them, the program was working fine the way it was; why were we stirring the pot...what made us think we knew better than they? (d0).

Regional autonomy within the program was highly valued, and an effort at building coherence across regions was a concern if it would limit that flexibility (c41). In short, the team was encountering resistance to their approach, their methods, and even the underlying purpose (d0). The team had been working intensively for months, and the response was unexpected: "I was taken aback by the negative reactions " (d0). The experience was described as a "demoralizing" setback (d0). Although the members of the evaluation team had a shared background and 'common language' among themselves (d0), at the outset it may have positioned them as outsiders to others in the program. The speed of implementation may also have been a factor in actors' responses, to be revisited below.

Restructuring, changing tactics. The emphasis on utilization and collaboration meant that the evaluation had to happen 'with', not 'to' program stakeholders. Although the program director was a "powerful ally", accounts of this case stress the importance of mutual trust and credibility "across the program" in order for an approach like this to succeed (d0). Therefore, although the team held to their conviction that the approach fit the program's complexity (d0) they began to adapt it (d0, d29), while also seeking ways to build underlying support for its essential elements. The team believed that developing such support would require: "a paradigm shift at both the project and program levels" (d0). To enable this shift and establish the legitimacy of the approach, they increased their focus on building capacity for systems thinking, providing opportunities for program actors to try out and experience elements of the approach, and building trust and credibility. These three levels of activity were intertwined.

Enabling a Paradigm Shift.

Promoting systems thinking. The team worked to help actors ‘expand their frame’ of thinking about the program at all three levels: program, regions and projects. At project and regional levels, for example, actors were encouraged to extend ideas about the scope of research studies, for instance to consider the context in which research findings were intended to be used as part of the research process itself rather than as a post-research step (d0). A participant noted: “we spent a lot of the first few years really pushing this idea that they had to be thinking more systemically” in ways that were thought to be linked to achieving program objectives (c41, also d07, d30).

Anchoring and focusing with tools. There is a possible tension between disrupting established thinking and building program coherence. The team developed frameworks and other tools that could be used in common by various program actors to improve internal alignment, and which, through their use, could anchor systems thinking in the program: "it became increasingly clear that one of the most effective tools we had for building coherence across the program was using common conceptual frameworks for understanding and implementing the work" (d0). To be successful, it was important that new tools and frameworks not seem alien to stakeholders. Co-development with stakeholders and close attention to communications, especially language, helped: “this has meant many hours of writing, revising, analyzing, and designing products in collaboration with colleagues and stakeholders" (d0)

how these [shared conceptual] frameworks were communicated could either 'make or break' their effectiveness. Naming something can be empowering; however, it can be alienating if names used don't resonate with people. Introducing new language to communicate a concept (e.g. generative dialogue) frequently led people to dismiss the language as jargon, and then 'throw the baby out with the bathwater' (d0, also c42).

Two illustrative examples follow.

A common tool to support research quality. One of the “big ah-hahs” that came about through the team’s exploratory work concerned a need to support research quality (c41). Multiple stakeholders voiced a need to increase research quality at the project level and the evaluation team brought this forward to leadership (c41). The team worked to help actors at multiple levels agree on what quality research meant for the program, and at the same time to think about how quality might be broadened beyond research methods to include consideration of the system in which the research was meant to be used. That the definition of quality might include more than research methods generated some pushback (c41). A participant described the conversations: “we talked about it with the leadership team, we talked about it with individuals, ... [with the] research methods group that was doing capacity building with projects... it was a real back and forth thing” (c41).

The team developed a tool (a checklist) to help define and consistently assess quality across projects. Development took about two years and did not happen smoothly. A turning point came when, through collaboration with the research methods support team, an existing framework was identified that included the desired components but used familiar terms, in other words: “the way we talk about it in the research world” (c41). The evaluation team adapted the framework, working with the research methods group to develop a rubric, “and everybody kind of liked it” (c41); it “helped bring people around to the idea” in general (c41). However, resistance was encountered again at the point of implementation. Concerns were raised because individual projects were diverse and some actors did not feel that the tool was a fit with their own project (c41). A second turning point came when someone with a high level of credibility intervened: “I was trying to be gentle about the whole thing and he just came in and said, ‘what’s this about, are you serious’?... he came in with credibility because of who he is” (c41).

“Buy in” came slowly, but it was significant for the program: “it helped [them] define what they really mean by quality research and gave a consistent way of analyzing that across the projects... that was a big step. Basically, it was everybody saying, ok this is what we mean when we’re talking about research quality” (c41). This example describes the importance both of characteristics of the tool itself (e.g., language) and the approach to develop and introduce it.

Bringing data forward on the issue, encouraging discussion, developing a solution collaboratively, intervention by a credible champion, and time were all important.

Building theories of change. Theories of change are a second example of the use of tools to promote systems thinking and coherence.

Building a program theory. After introducing generative dialogues in Year 2, the team regrouped and decided to try developing a program theory of change (TOC) with the leadership team. TOC development was seen as a mechanism to build leadership engagement with the evaluation, and promote “shared meaning making”, for example by clarifying assumptions and connections within the program (d24, also d11, d29). It would provide an opportunity to think about the program with a systems lens (d29), and ultimately to contribute to the coherence objective (d0). It was also aimed at helping to drive collective action, and enable testing of theories about the program against measured results to promote learning (d29). The program TOC development was called: “a transformative moment, when, for the first time, the entire leadership team wrestled with the what, why and how of [the program]” (d0, also d29).

Program TOC development was difficult. Multiple perspectives needed to be negotiated while new relationships were being formed. A document describes it as “a conversation that took time, inspired passionate discussion, and even caused some disagreement” in part because of the multiple “scientific disciplines” brought together in the group (d05). This was echoed by a participant, who described resistance in part because of diverse perspectives and low levels of trust: “there were new people, I was new, people new to philanthropy... and you move at the speed of trust they say. We were just in those early stages of building trust” (c42).

Ultimately, the program TOC was described as a breakthrough because of both the collaborative process and the result: “really getting people engaged, even when there was sometimes pushback... that process and then having something in hand, give a lot of traction” (c41). Theory of change became one of the central activities in the evaluation (d29); TOC

development was the most referenced activity in the dataset for this case (e.g., c41, c42, d0, d14, d23, d24, d25, d33).

The program theory of change continued to be developed iteratively after Year 2 (d08, d29). Iterative development was important because the TOC was agreed to be a ‘best guess’ that would need to be tested and refined, and because it was accepted that conditions surrounding the program, and the program itself, would continue to change (c41). Commitment to the TOC increased over time:

Now [in Year 9] people love the theory of change... at this last leadership team meeting there was this huge pushback [to making changes] saying ‘we *love* our theory of change’ you know. It’s interesting because there *was* a lot of pushback initially on it. So the theory of change has been really important to ground [work] in at all levels (c41).

It was described as the biggest single initiative giving the evaluation traction (c41), and “the most helpful tool for focusing what we do” (c42).

Building project-level TOCs. The team introduced the idea of project-level TOCs at a community of practice meeting at the end of Year 2. The response was described as ‘polite confusion’, and at times “stormy” (d0). Although a whole day was dedicated to this at the meeting (d29), it was not enough time for project teams see the value, or the difference between a TOC and the logframes they already had: “initially for most people it was a new way of thinking, and culturally in some places its harder to do... [also] they have a lot of funders and every funder is going to want something different” (c42). Some projects were already far along in their implementation; they were least receptive. Projects that were at an earlier stage responded better (d0).

The team persisted. They worked directly with the regional teams to train them to facilitate TOC development with projects, for example. The program-level TOC was discussed at regional and project meetings as a way to increase shared understanding of the program theory at all levels (e.g. d24, c42, d22). The team also encouraged the foundation to mandate compliance. New grantees were required to develop a TOC at the start of their project, to refine it over time (d11, d07, d03), and to develop an M&E plan linked to their TOC (d07, d03). This was based in

an expectation that project members would come to value TOCs with practice: “I feel like a lot of just forging ahead and believing in it until people could grab on... In the beginning people just did it because they had to, and now I think they actually see value in it” (c42). After Year 3, most projects had TOCs, although improvements were still needed (d0).

Using the TOCs. “Building the habit” of using the program TOC as an input to decision making took years (d0). Use of the project-level theories of change also came slowly: “It was really hard in the beginning, because nobody talked about it really, particularly internationally. Everything was logframes” (c42). More than two years later, in Year 5, an evaluation report from one region indicated ‘mixed’ use of project-level TOCs there, though evidence of ‘systemic thinking’ (d01). The challenge to foster use of TOC at the project level is attributed in part to the ‘institutional environment’ within which many of the grantees work. This included a norm of accountability-focused evaluation required by other donors. A leadership team member is quoted as saying: “Evaluation is not [usually] something that’s really about reflecting ... It’s more about, “Oh, I said I was going to reach 5,000 farmers, can I check that box?”, because the focus is on “doing what they said they were going to do” (d29). There was said to have been a focus on the ‘doing’ of research without “investing our time and resources in learning how they are doing, and what the results they’re getting mean” (d29).

Although there was an intention to have TOCs at all three levels (project, region, program) (d07), TOCs have been embraced the most at the program level. Regional-level TOCs are less developed in Year 9 (c42, see also d29). At the project level, a participant noted that “many of the projects value it” (c42); “it’s really working well and they’ve really grabbed on to the theory of change and find it useful” (c42).

How the TOCs are used varies. Three applications have been described as especially important: as a base for M&E activities, to foster collective understanding and critical thinking, and to support focus and action at the program level.

Focusing M&E. The TOCs have been described as “building blocks” on which evaluation plans are built, and as a focusing tool for evaluation activities (d29). An evidence framework was created to guide testing of the program theory (d29), and program-level evaluation questions are depicted on the program TOC diagram to make a clear link between

focus questions and related aspects of the program theory (d11). A participant explained: “it’s how we know what to gather information about, it helps guide us... There are only so many questions you can really go deep in and engage with” (c42).

Collective understanding, critical thinking. At the project level, although the TOCs are used to varying degrees and in different ways, they are thought to be important for project actors to reach a collective understanding of their project:

it is not an uncommon experience... for project members to show up at the inception meeting without knowing what is written in the proposal... [I]t is important for collaborative partners to be on the same page with what the project is going to do and intend to achieve. The theory of change can be an important part of... [helping] key project stakeholders build coherence (d29).

Project members are encouraged to dig into assumptions as they develop their TOCs (d07). Creating their own project TOC is said to be key to making it useful to project members, because the TOC is largely about “opportunities to think about how can we actually make change here, and how might that work, and how do we see ourselves in that” (c42). An aim is to inform project planning, reach agreement on evaluation, and a mutual understanding of how the pieces of the project fit together (d07, d03). The project level TOCs were also said to be used for comparing actual project activities over time against their theory of change, to assess and guide alignment between the intentions and actions (d25), and to guide testing of aspects of the project theory through evaluation (e.g., d14).

The foundation’s regional support teams are also said to use the project TOCs to help grantees “identify neglected assumptions” and “to attend to aspects of the design that had not received much thought”, referred to by some as the ‘miracles’ in the project’s theory (d29, also d07). This includes thinking “more critically about systems their research worked within” (d29, see also c41, d0, d29, d07). An example describes a technical innovation (e.g., a more nutritious crop variety) that has an unintended effect of increasing milling costs, which undermine the envisioned outcomes of the project (d29). It was noted that researchers were not usually trained to consider the social and systemic context of their research (d29, also c41).

Focusing the program. The program TOC, together with reflective practice in the overall DE process, is said to have helped to support agreement on program scope and providing a basis for testing new ideas. This was described as important because of the diversity of actors involved, and the foundation's commitment to a highly participatory leadership process. For example, the TOC was said to help "bound the creative energy" of the leadership team, so that ideas introduced by its members "remained somewhat within the agreed-upon scope" of the program (d0). It is important to emphasize that this process surrounding the TOC is key: "there were times and places where the scope and direction of the work were threatened because the goals and practices were overly ambitious and/or diffuse, and the work thus seemed to be losing its center. Developmental evaluation has been successful at helping to pull in the boundaries and focus the work by asking questions about purpose, methods, context and capacity" (d0). The TOC has acted as a centering tool in that process.

Action. The program TOC was described as a tool for enabling action as well (c42). For example, in Year 4, a new framework to guide the program's research priorities was "formalized" by embedding it in an updated program TOC (d05). The program's criteria for reviewing grant proposals now include how a new project aligns with the program TOC, which reflects these priorities (d10). An advisory committee member is quoted as saying: "[t]he refined Theory of Change helped to translate the now defined framework into an action plan" (d05).

Translating intentions to action was not always easy. To continue the above example, the decision to focus grants on the new priorities for research had the implication of ending funding to some long-term grantees not working in those areas. This decision seemed to pull against the foundation's commitment to long-term relationships and long-term funding as part of their approach to grantmaking (c41). This was said to have been uncomfortable for people at times, taking a long time to work through. The evaluation team supported program actors to navigate this by asking questions and helping to clarify the rationale behind funding decisions and the process for making decisions. Some project commitments were changed to better align the program with its new priorities (c41).

The development and use of common tools and frameworks were a central feature of the early stage of the evaluation. They were described as a way for the evaluation team to engage

program actors in the evaluation and build their capacity, to help build program coherence, and to give some structure to ideas drawn from systems thinking. Such tools were carried across various parts of the program as ‘boundary objects’, for multilevel and cross-group communication, and for supporting action.

Maturing Stage

Between Years 3 and 5, the evaluation system began to mature. Indications of stability are notable by Year 5 (d29, see also c41), when the evaluation system was said to be, finally, ‘no longer a debatable thing’ (d29).

Year 5 was also significant because the grant that had enabled the program to expand in Year 1 was ending. The granting foundation was asking about ‘scaling up’ program results (d0); at the same time, it renewed its commitment with another 5-year grant (d28; d04), based on a proposal prepared by members of the evaluation team (d33).

Over these years and those that followed, the evaluation was integrated deeply into program activities and structures, and activities were distributed across actors at the three levels, to make the evaluation part of the work of the program.

Making it Part of the Work

Rationale. Integration was intended from early on, as noted by a member of the leadership team in Year 2: “we wanted to create an evaluation framework that was practical... because we’d have to incorporate it into our daily lives” (d30). The foundation considered it to be important that staff “see quality improvement, knowledge management - and, indeed, learning - as the responsibility of all” (d34). Program documents, including annual newsletters, depict a blurry line between ‘evaluation’ activities and other program activities, to the extent that it was difficult at times to know how to code the document data for this study. Members of the evaluation team were not always involved in activities that were distinctly ‘evaluative’. A participant provided an explanation:

people would say, shouldn’t the evaluators do that, and I would say ‘no because we don’t really have ‘evaluators’ - we’re integrating this remember?’ So, every meeting was an

evaluative meeting. It was an evaluation moment... It's this idea that it has to be integrated into the work that they do, and they have to become facile in the way that they do that (c41).

Another participant stated: "some people are going to be more adept at evaluation skills, and that's fine there's a place for that, but for people to hold more of a stance of inquiry and be able to make meaning and have that inform next steps is a skill everybody should have" (c42).

Evaluation as an activity has been built into the program TOC and the program's guiding principles (d11; d06). A study published in Year 8, which compared organizational structures at several foundations, identified this foundation as unique for the way that evaluation and learning functions had been combined there (d34).

Locating the evaluator in an integrated system. The evaluator role was positioned as facilitative: "[w]e walk alongside our program colleagues, and facilitate their transformation of information into knowledge, and knowledge into wisdom" (d0). 'Evaluation' activities were shared by actors in the program, with members of the evaluation team providing support when expertise is needed. A document quotes the lead evaluation consultant: "I have learned when to sit back and let the process unfold, and when to step in and offer expertise or challenge the direction something is headed" (d0). Otherwise, the evaluation team was described as responsible for the process for answering questions, including managing data collection and storage systems and "mak[ing] data available to the responsible teams for developing evaluative conclusions" (d29).

Specialized evaluators, including external consultants, look deeply into specific questions from time to time when directed by the foundation (c42 see also d05, d17, d20, d30, d34), particularly for summative questions, where impartiality of the embedded evaluators and collaborating program actors might be an issue. The blurring of roles in DE for both evaluators and program stakeholders is recognized in a document, which explains that DE evaluators collaborate closely as part of the program development team, and that program stakeholders collaborate in DE (d07). Although the lead evaluator was not a foundation employee and therefore not officially an 'internal evaluator', their close work with the program made them very "involved in helping to shape how things are happening" (c41). As explained by a participant: "it

just feels better to have someone come in from the outside and look to see how well those things are working” (c41). This was described as a combination of ensuring impartiality in ‘real’ terms, and in the minds of stakeholders: “sometimes... you want it to come from outside so there’s not pushback”; “it’s the idea of being able to recognize your biases. There’s the reality of that, and then there’s the perception of that, and you have to take both into consideration” (c41).

However, the use of specialized evaluation consultants did not work well when they were hired to provide ongoing support to grantees. Local consultants were hired to provide culturally and context-aware, on-the-ground support to project members (d31, d0). However, these consultants were viewed by actors as ‘separate’ (d29) which served to contribute to the perception of evaluation as a separate activity - the opposite of what was wanted for this purpose (d0, c41). Moreover, the consultants were accustomed to conventional evaluation approaches. Retraining them in an alternative approach required more time and resources than expected (c41, d0), and outputs that they provided were not well used by the regions or the projects (d29). Eventually, evaluation support was transitioned to the research methods support teams (d0). Although sometimes it was found that researchers struggled to “think evaluatively” (c41), it was observed that, in this case, “it was actually easier to teach non-evaluators to integrate and provide support for evaluation than it was to teach evaluators, who came from a really different perspective, how to do evaluation differently” (c41). The use of specialized consultants for internal support was ultimately abandoned, in large part because it was inhibiting rather than promoting integration of evaluation into the work.

Adding structures. In addition to technical support and facilitation, the evaluation team helped build structural capacity and embed processes to enable evaluative work to be carried out. Examples include ways in which meetings and meeting structures were developed and used. For example, at the program level, a strategic discussion was added to the program’s yearly activities. This was described by a leadership team member: “[w]e try to understand what we’re doing to achieve our goals and then look at what’s not there. This may sound very simple, but it’s the kind of reflection that, if you don’t take the time to do it, it’s not going to happen” (d30). This individual noted that previously the organization had no mechanism to help staff have such conversations (d30).

Meeting structures were also established at the project level. Project team members were often geographically dispersed. The evaluation team encouraged the foundation to provide resources for project members to “come together as a project” (c41). Providing support for project meetings was described as a way to contribute to project success: “to structure things to allow change to happen” at the project level (c41). The meeting was meant to include evaluative activities, but in a way that it: “wasn’t having this external evaluation thing being embedded in, but it was much more the team would come together to do stuff” (c41).

Tailoring evaluative activities. The team integrated evaluative activities into meetings at multiple levels (also d31, d29). To avoid repeating the experience of the generative dialogues, they tailored activities closely to context and capacity. As described in a document:

As we moved forward, we worked with reframing the concepts of complex systems into language that did not feel foreign, exclusionary, or overly precious. Sometimes, we learned, it is not important to name things; it is sometimes enough simply to integrate the concepts into the work, without calling attention to them (d0).

Although the term ‘generative dialogue’ was dropped, the team looked for other opportunities for actors to practice the adaptive action framework (d0). In the new project meetings, project teams were encouraged to use the ‘what, so what, now what’ questions in their discussions; and to develop, review and revise project-level theories of change, using them “as a springboard for conversations” (d07, also d22, d26). Project meetings are described as a “formal moment where insights can be consolidated, discussed, acted upon, documented, and shared” (d07), supported with data contributed through the evaluation effort.

Good quality, professional facilitation at meetings was identified as critical to results. The team worked with facilitators to plan and execute meetings. A team member explained:

we could talk about what we wanted to happen in the meetings and what we wanted to bring to the meetings and also what other kind of evaluative stuff we wanted to come out of the meetings, and we had trained facilitators who could make that happen (c41, also c42).

Facilitating interpretation of data. The team provides data to meetings at all levels. For example, to leadership team meetings: “[w]e always bring data... ‘what’ data from what we’ve done...and in that meeting we make meaning from it” (c42, also d07). For the regions, “[w]e do data collection and we do... some analysis and give it back to the regions, we do some interpretation, but we also expect the regional teams to do the bulk of the regional interpretation” (c42). The team experimented with providing raw data to program actors for interpretation, “[a]nd that didn’t really work well” (c42). The team instead typically provides partially analyzed or synthesized data for groups of program actors to interpret and discuss (c42). A participant described a particular example: some pre-meeting preparatory data analysis was done by a small team, and regional teams were given a question to answer (e.g., “How much progress are we making? Progress toward what?”) along with data. Each group prepared a presentation to deliver their findings. Then, “we took their thinking along with other thinking and then had a small group actually lead a discussion session that took us deeper into what that means” (c42). Having a series of steps for small group analysis and collaborative interpretation reflects a form of pacing, or a ‘runway’, in other words a process people can walk through in stages, to give time for interpretation, reflection and for the integration of multiple perspectives (c23).

Mandating integration. At times, ‘making it part of the work’ was mandated as well as enabled. Like the project-level TOCs (above), attendance at regional community of practice meetings was mandated. These meetings were described as a considerable time commitment for researchers: “people would really complain, but once they’d been to one, people would fight about who got to go to the next one” (c41, also c42).

Another example relates to development of a database. By Year 3 the team was “swimming in data”. Sifting through data was consuming so much time as to be unsustainable (d0), and providing the detail required for reporting to the partner foundation was a challenge (d35). The team looked for a streamlined way to collect and process data. In this situation, as with examples above, the team struggled with the desire to ‘embrace the heterogeneity in the program while building coherence’ (d0). The challenge was described as ‘systematizing a dynamic process without losing its heart and soul’ (d0). Development of a database system was supported by leadership. However, it was resisted at the point of implementation: “there was an

outrage of dissent” (d0). Concerns included that the database was too complicated, and being restricted to entering data predetermined by the database (“people were used to having an infinite canvas on which to create” (d0)), and it “demanded that we collect accurate data against the theories of change being articulated” (d0). Ultimately, the use of the system was mandated by the foundation, for a trial period, as part of grantee contracts (d0). As actors gained practice using the system, and had time to see its benefits, their willingness to use it was said to have increased: “Over time, people began to see what [it] could do” (d0). A participant noted: “[B]y the second year, people were saying ‘gosh this sure saved me a lot of time when I had to do this or do that or the other thing’. But had we asked if they wanted to do it, or tried to get consensus, it wouldn’t have happened” (c41). This was one of the times when the foundation’s willingness to “pull rank” and mandate use of a tool was important to advance the evaluation (c41).

Building capacity. Concerted effort to develop capacity at all levels is described in case documents (e.g., d0), including development of tools and structures as noted above. Skills development was also described as integral to integrating evaluation in the work:

to embed evaluative thinking, and that for us includes, if you break it down, just a way of working... In order to do that you have to have skills, right? You can have simple reflective skills, but you have to know how much data to get, which data, how to make meaning with it. It’s really the notion of embedding the what, so what, now what at all levels, and what are the skills and tools to do that well? (c42).

For example, as noted above, meetings at various levels were used as opportunities to introduce and discuss monitoring & evaluation (M&E) concepts (d0, d29). Other examples include development of a detailed ‘evaluation handbook’ (d07), inviting “key strategic members” to the American Evaluation Association conference (d0), and direct training of regional support teams in M&E (c42, d33). The foundation also invested in building staff capacity (through consultants) to identify and ask good evaluative questions, as part of building a ‘culture of inquiry’. Having well-constructed priority questions was noted as key to avoiding ‘information overload’ (d07). It was also noted that how questions are framed can change the perspective of the work in important ways, for example, shifting from “Did we make the mark?” to ‘How can we learn in

order to adapt, so we can make the difference that is needed?’ (d34) can lead to different inquiries.

Making space for DE. Challenges faced by the team included establishing a new approach to evaluation among a highly diverse and committed set of actors. As noted in a document: “[i]mplementing an evaluation that is very different from what people are used to can provoke doubt, resistance, pushback and even (occasionally) hostility. We experienced all of these at different points” (d0). Successfully integrating DE as an embedded system took years. The team needed to do more than sustain and facilitate the work through that time, they had to actively create space for it. They needed to create space to experiment, fail, recover, adapt, build capacity, establish legitimacy, and earn trust.

The *opportunity* to make space for DE is attributed to factors such as the foundation’s culture and its willingness to invest in long-term solutions (above). The innovative culture of the foundation was reflected by leaders in the foundation who were “trying to get things right”. A participant elaborated: “we had support in that there was such an openness, in the beginning there was also a lot of ‘what’s happening?’ but because there was so much openness, they were ‘ok let’s see’” (c42). This was identified as key among the factors that “kept the door open” for the developmental approach (c42, also d13). Growing interest in systems and complexity thinking in foundation circles over this period was also identified as important (c42, d34).

The *doing* of it is reflected in several of the examples above. These examples demonstrate a high degree of flexibility on the part of the evaluation team to respond to stakeholder needs, as well as patience, persistence and opportunism, discussed below. The strategic use of outside support is also noted below.

Flexibility, responsiveness. There was an intent and willingness from the beginning to fit the approach to the needs and context of the program and its stakeholders, as noted above (c41, d0). This stance placed some limits on commitment to particular methods or frameworks. For example, team members reported being unsure about the extent to which DE might be useful to the program, or what components of the program might need DE and which might be better served with other approaches (d0). It meant remaining flexible:

We let the evaluation evolve into what it needed to be. What was working, what wasn't working, as the foundation became more clear about what they needed and we became more clear about what was helpful, we let the evaluation become what it needed to be. And we realized yes, I guess this is developmental evaluation, sure (c41, also c42).

Patience, persistence. Patience and persistence are recurring themes in the data. The team felt that program actors would learn to value the approach if they had opportunity to practice it and experience results. The trust that 'belief would follow practice' led the team to create these opportunities via various means, as noted above, including building capacity, distributing the work, enabling participation, and mandating compliance. "Over time people could see... the case for this working, and started understanding and buying in" (c42). For example, over the early years the team consistently introduced the use of the 'what', 'so what', 'now what' sequence of questions. Eventually: "[l]ike magic, the questions stuck. The leadership team members began to use the framework explicitly in [program] work, and the regional teams used it with their projects" (d0). As program actors began to see this as effective, it also helped establish the credibility of the evaluation team (d0). The need for patience and persistence was attributed to resistance to the new approach, and to evaluation itself:

my experience with evaluation is that very frequently ... people hate the evaluation until they get the data and they have something they can sink their teeth into and they realize 'ooooh, this is very interesting, this is telling me a lot about my program, and this makes me feel like I'm accomplishing something'... I tell people, 'let's just get through this first round and get some data in your hands, and then you can tell me if you think it's a bad approach (c41).

A participant asserted that important to this persistence was having trust in "what we were trying to create - a complex and adaptive global grant-making program that used evaluative thinking at its core" (d0): a vision that could fuel years of effort.

Opportunism. Persistence worked together with opportunism. As described in an account: "It sometimes felt like for the first 3 or 4 years we learned more about what worked with the evaluation by initially getting it wrong and by crawling through windows of opportunity" (d0). As an example, the evaluation team encouraged the foundation to provide support for grantees to

come together for project meetings (see above). Support for project meetings was new to the foundation, and there were administrative barriers to overcome: “At first there was pushback... they said well, we can’t really do that because of this and because of this” (c41). When a review of budget structures came about, it was described as one of the “windows of opportunity” for the team (c41): “It was a big thing in terms of how the foundation did funding and what that meant. It was a big shift administratively, but we came in at a time when they were looking at those things, so there was an opening for it” (c41). Part of bringing about this change involved preparing the ground: “you can do little things that kind of plant things... It’s not like you just hold the idea and then you jump. You’re suggesting things and bringing things up and doing things and then when there’s an opportunity to make a shift... everybody’s sort of ready” (c41). Deliberately pacing change is also suggested by another participant who explained: “we have to be careful about innovating too fast” (c42).

Drawing from outside. The team also drew in outside support. Examples above include the use of trained facilitators. Outside support also included experts in systems and complexity thinking, whose experience and profile added weight to concepts being introduced (c42, c41): “Their expertise, key advice, and timely contributions helped us refine, refocus and inspire this work, particularly in the tough spots” (d0). Drawing in outside evaluation support was also a way to hybridize the evaluation. It allowed them to access the advantage of an internal position, with deep understanding of context, as well as external perspectives at greater ‘distance’ from the program when that was warranted (e.g., for summative work). This made the evaluation effort overall more robust and enhanced credibility with program stakeholders.

Current Stage

Use and Outcomes

This case covers a period of 9 years in which a new evaluation ‘system’ was designed, implemented and integrated into a program. The data suggest the evaluation has had impact at multiple levels. Use of various kinds is mentioned above, and include conceptual use (e.g., thinking more systemically, changing perspectives on scope of research quality) and instrumental uses (e.g., new procedures for project selection). They also include use of the evaluation process

(e.g., adopting the ‘habit’ of referencing TOCs in decision making, improving skills to interpret data) as well as use of findings. Some broader impacts noted in the data are touched on below.

Foundation level. This evaluation effort is said to have, over time, changed the way that the foundation thinks about evaluation. Having started with a sense that evaluation was needed, but with concerns about what it might involve (e.g., burden on grantees; “having these outsiders come in and judge what we’re doing” (c41)), the evaluation effort has “made them more committed to thinking about evaluation, by thinking about it in a more developmental way. So, they’re not suddenly having requirements for intensive data collection and analysis, but they’re wanting programs to really use evaluation to promote more reflective practice” (c41). This change is thought to have come about through experiencing evaluation that “was rigorous, [but] that wasn’t antithetical to that mindset” (c41). The approach was described as innovative by a representative of the partner foundation providing grant funds to the program, because it: “helps teams understand the implications of results and map out change. It impresses me as a process that is cutting new ground” (d33). The adaptive action framework is described as having been “incubated” by the foundation in this program. In Year 6, the foundation is reported to have adopted the framework “across the entire foundation” (d05).

Program level. There are indications of adaptation having become embedded at the program level as a valued purpose and outcome of evaluative work. In Year 8, a study reported perspectives of members of the program’s leadership team on the evaluation. These express a commitment to supporting grantees to adapt their projects based on the information they gather over time, and they contrast this with approaches by other donors who are said to sometimes penalize change (d29). Evaluation, including the adaptive action framework, has been embedded in the program’s guiding principles. For example, the program’s operational principles include: ‘Integrate M&E’; ‘Reflective practice, using the adaptive cycle’ and ‘Ensure use of the Theory of Change’ (d06).

Key findings from earlier evaluations (e.g., ‘authentic farmer involvement’; greater program cohesiveness), which were not implemented prior to Year 1, have been central to the work since. The evaluation system has also been used to help resolve a tension between stakeholder interest in ‘scaling up’ program results and the program’s commitment to place-

based grantmaking, for which scaling by “replication of best practices” would not be appropriate (d0). Instead, the evaluation system provides support to a contextualized scaling approach (d07), which includes ‘adaptive capacity’ as a measure of success (d0) and uses the program’s operating principles as ‘fidelity measures’ for planning and evaluating the scaling process (d0).

By defining "scale" as more than just aggregated numbers of adoption, [the evaluation system] considers the multiple ways in which an intervention has been adapted by others, how it has served as an inspiration for other interventions, and/or how it has contributed more generally to the broader public good for example, by influencing policy across contexts. That’s why most evaluation questions in the [the program] ask: For whom? How? (d07, also d29).

Project level. By Year 9, the effort to encourage actors to incorporate the context of application as part of the research process was said to be taking hold. Progress in this was attributed to more diverse collaboration at the project level and to the evaluation work. As reported by a participant:

I think pushing the projects to be thinking about development outcomes. What is your indication that this may be having an effect on nutrition, on livelihood, on productivity on the ground? In order to ask those questions, they have to be thinking about how to make that shift... you know they always say what you measure is what you end up doing, right? (c41).

This ‘shift’ was also attributed to capacity building, for example the foundation’s support for grantee training events, the foundation’s ongoing communications and feedback to grantees, and expectations for rigour and quality in the research process, which was supported by evaluation activities (above) (d01). A Year 6 independent survey reported that grantees ranked this foundation’s reporting and evaluation process high relative to other foundations for strengthening their organizations and their projects (d17). Research projects were reported to have become more “iterative and reflective” through exposure to “iterative and reflective” evaluation (d0).

Moving Forward

Recent directions suggest that the need to support program coherence and focus, balanced with respect for diversity, has not gone away. For example, the program has been establishing its guiding principles with an aim to “make sense out of and bring coherence and alignment to what is a very complex and dynamic set of initiatives within the overall umbrella of [the program]” (v01; also d29). The program has been extending its process for identifying priority questions by using an emergent learning framework. The learning framework helps the program focus on the coming year’s process for learning: “how are we going to pay attention, how are we going to make meaning from it, how is it going to inform what we do?”” It supports integrating learning in the planning process, and communicating it clearly to others at the foundation (c42). After almost a decade, the program remains committed to developmental evaluation as a means to carry the work forward.