Navigating autonomy, agency, and accountability:
A study of school self-evaluation in New Zealand secondary schools

Lisa Dyson

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Education, the University of Auckland, 2021.

Abstract

In New Zealand, schools are expected to carry out regular cycles of using and reflecting on data to evaluate their teaching and learning approaches, identify areas for improvement and refine their practices. This practice of school self-evaluation is increasingly prevalent globally, but the New Zealand context is unique in the fact that it is a core mechanism of internal and external accountability within an education system where schools have autonomy. Despite the New Zealand education system's reliance on school self-evaluation, there has been little recent New Zealand empirical research on these self-review processes. The small amount of literature shows that despite the expectation to engage in self-evaluation, many schools find these processes difficult. This thesis explores the practice of self-evaluation and data use in New Zealand secondary schools and the perspectives of the teachers and school leaders engaged in these activities. This study also investigates how these practices shape organisational change, changes in practice, and changes to teachers' work.

A realist approach to research is used to better understand the ways that lived experience interacts with processes of change, within particular contexts. The articles that comprise this thesis show that the outcomes for teachers and schools are not uniform and that they are shaped by the organisational contexts in which they work and the wider social and political context, particularly the accountability context. The schools in this study use data practices to improve their teaching and improve outcomes for students, however, external accountability pressures can negatively affect their data use processes and school self-evaluation. Navigating the new expectations and the competing influences in the wider context can cause tensions for teachers. This thesis helps us to better understand that the process of data use is more complex, contextual, and less rational than current theoretical understandings might suggest.

Preface: A Reading Guide

This is a thesis with publication. My research findings are presented in four articles, three of which have been published and one which has been submitted for publication. The chapters preceding the articles provide the background, context, theory, methodology and methods that inform my research. Each article is also preceded by a brief introduction. Following the four articles is a conclusion that draws out common themes emerging from the work. This thesis does not have an extensive initial literature review chapter, but instead, it has a brief review to set the stage about my overall topic of data use practices and school self-evaluation activities in New Zealand secondary schools. This is because my articles, while all built on this same overall topic, each draw on distinct bodies of literature, including inquiry, accountability, and evaluative thinking. I have attempted to avoid duplication of concepts in the various literature reviews, but the nature of a thesis as a series of discreet articles means that some duplication is inevitable.

The four included articles are:

Dyson, L. (2020). Walking on a tightrope: Agency and accountability in practitioner inquiry in New Zealand secondary schools. *Teaching and Teacher Education*, *93*, https://doi.org/10.1016/j.tate.2020.103075

Dyson, L. (2020). Data use in New Zealand secondary schools: Tracking, traffic lights, and triage. *Assessment Matters, 14*, 89–111. https://doi.org/10.18296/am.0043

Dyson, L. (2018). Drawing on theoretical knowledge to build evaluation capacity. *Evaluation Matters—He Take Tō Te Aromatawai*, *4*. https://doi.org/10.18296/em.0032

Dyson, L. (under review). Beyond the rational model of decision making: A conversation with evaluative thinking.

Acknowledgments

Table of Contents

Glossary

This section defines key terminology used in this thesis.

Decile

A school's decile rating refers to the socioeconomic level of the community surrounding a school and is used to provide funding to schools.

Education Review Office (ERO)

ERO, charged with the quality assurance of schooling, conducts periodic reviews of each school in New Zealand.

Māori

Māori are the indigenous people of Aotearoa/New Zealand.

Ministry of Education

The Ministry of Education is responsible for providing strategic and operational advice to government, ensuring that government policy is implemented and direct services are provided to students, and is responsible for the *New Zealand Curriculum.*

National Certificate of Educational Achievement (NCEA)

The National Certificate of Educational Achievement is the main qualification for secondary students in New Zealand. Students are assessed against a range of standards using both internal and external assessment. When students achieve in a standard, they gain a number of credits. Students must achieve a specific number of credits to gain an NCEA certificate.

New Zealand Qualifications Authority (NZQA)

NZQA is responsible for the integrity of the New Zealand Qualifications Framework, which includes NCEA. NZQA is responsible for the assessment of the standards, through preparing and marking examinations, the oversight of assessment conditions and moderation of internally assessed standards.

Organisation for Economic Co-operation and Development (OECD) countries

The OECD is a forum in which the governments of wealthy countries work together on economic, social and educational matters. OECD publishes the Programme for International Student Assessment (PISA) and annual Education at a Glance reports.

Pacific

A term used to describe people who have migrated from Pacific Island nations to live in New Zealand and their descendants.

Whānau

Māori-language term for family and extended family group.

Chapter One Introduction

My Journey into this Research

My interest in studying school self-evaluation in New Zealand secondary schools initially stemmed from my views of what it was *not*. I was an American with a professional background in education and programme evaluation living in New Zealand. I was working as an evaluator with a large social service agency in Auckland that was respected by the industry and government for their well-developed self-evaluation processes. My team was tasked with doing an assessment of the organisation's evaluation processes and when we spoke to the social workers within the agency, we found that managers and practitioners did not value the evaluation processes or use the evaluation results for improvement. The field of evaluation values the utilisation of both the evaluation findings and the process itself to inform decisions within an organisation (Patton, 2008). While the organisation's leadership touted their well-developed processes, I did not see them as truly a learning organisation. The exercise was more about the performative appearance that internal evaluation was being conducted than it was about meaningful work. I came to think of it as "evaluation theatre."

In conducting self-evaluation, this organisation was part of an international trend. Self-evaluation is an increasingly pervasive form of evaluation in governments, universities, for-profit and not-for-profit organisations, and in fact is the most common form of evaluation being practiced throughout the world today (Fetterman, Kaftarian, & Wandersman, 2015). As an evaluator who had worked with many organisations, I believed in the power and potential of internal evaluation, but knew that it could be difficult. I was disappointed in the approach of this social service agency because I was intrigued by the possibility of authentic self-evaluation.

I became aware that self-evaluation is at the centre of school evaluation in New Zealand (Nusche, Laveault, MacBeath, & Santiago, 2012), that schools had a longer history of these practices, and that a concept of teacher inquiry was embedded in the culture of teaching here. I wanted to learn more about this model that seemed to sit in contrast to the evaluation theatre I had encountered in the social service sector. School evaluation appeared *not* to be that. I wanted to learn more about the processes in this sector that appeared to hold the promise of authentic self-evaluation.

I was also interested in this system for what it was *not* in relation to school evaluation contexts in many other countries, especially the other English-speaking countries of the US, England, and Australia. I was aware of the overseas top-down accountability school evaluation models, such as No Child Left Behind in the US that were centred on standardised state tests, external evaluation, and rewards and sanctions for schools and teachers. I knew those models were changing the nature of teachers' work and often had negative unintended consequences. In contrast to those models, in my research I read that New Zealand's external evaluation agency, Education Review Office (ERO), had moved from a compliance, accountability function to a participatory, complementary function (Mutch, 2012) and I was intrigued by this model. I read books on the school self-evaluation context in New Zealand noting the lack of strong external accountability measures found overseas (Lai & Kushner, 2013; Lai & Schildkamp, 2016; Schildkamp, Lai, & Earl, 2013). I was attracted to the New Zealand approach to school evaluation that I viewed as a bottom-up, participatory, improvement model in contrast to the top-down, standardised state test-driven model of school evaluation that I saw in the US. The theory behind New Zealand's participatory approach sounded promising and I wanted to learn what happens in practice and how teachers and school leaders experience this system.

I came to learn that while New Zealand does not experience the same external accountability as some other countries, the accountability pressures are more complicated than I first envisioned. My research explores the ways in which the accountability context shapes school evaluation activities. In addition to the context of accountability, I also became aware that the way self-evaluation occurs is influenced by features of the New Zealand education context, such as school management, curriculum, and assessment.

Identifying and Addressing the Thesis Focus and Aims

Internationally, the way that schools are evaluated has changed dramatically over the last few decades and nearly every education system in the developed world has created or reformed their school evaluation policies and procedures (McNamara & O'Hara, 2008a). An international trend is an increase in school *self*-evaluation, which is evaluation processes conducted by the school itself. Interest in these processes have increased over the last 20 years, with self-evaluation "now seen as a matter of priority in most economically advanced countries in the world" (MacBeath, 2003, p. 2).

New Zealand is an interesting context in which to study this international trend, since school self-evaluation is the centrepiece of school evaluation in this country (Nusche et al., 2012). While it does have an external evaluation component, the links between external and internal review are some of the strongest in the world, with the OECD noting that, "New Zealand has probably gone furthest among countries internationally towards a collaborative school evaluation model" (Nusche et al., 2012, p. 95). In this system that foregrounds school self-evaluation, schools are expected to carry out an ongoing process of evaluation for improvement and are expected to use data for evaluation and decision making (ERO, 2015). Because of these expectations, teachers' practices of data use are more embedded in New Zealand than in many other countries (Nusche et al., 2012; Sinnema & Aitken, 2011).

Despite the New Zealand education system's reliance on school self-evaluation, there has been little recent New Zealand empirical research on these self-review processes (Mutch, 2012; Starkey & Eppel, 2019; Timperley, 2013). Most research on data use in New Zealand schools has occurred within the context of large-scale Ministry of Education funded interventions such as Te Kotahitanga, The Learning Schools Model, and Literacy Professional Development Project (see Bishop, Berryman, Wearmouth, & Peter, 2012; Lai & McNaughton, 2016; Timperley & Parr, 2008). All of these initiatives used quasi-experimental studies to provide evidence focused on whether the interventions worked. In contrast to these studies, this thesis does not focus on *whether* school evaluation processes work, but instead uses qualitative methods focused on *why* and *how* these processes occur. This thesis also does not involve an external intervention, such as those mentioned above, but instead examines everyday practice and focuses on the experience for teachers and school leaders. In New Zealand, we know little about how people in schools are interacting with and responding to data (Starkey & Eppel, 2019). Therefore, the exploratory study in this thesis explores what actually happens when people engage with data in the course of their ongoing everyday work and how teachers experience these processes.

The small amount of literature on the topic shows that despite the expectation to engage in self-evaluation, many schools find these processes difficult, self-review is highly variable across New Zealand schools, and only a minority of schools have strong processes in place (ERO, 2017; Mutch, 2012; Nusche et al., 2012; Timperley, 2013). This thesis identifies a selection of New Zealand secondary schools that are effective at school self-evaluation and attempts to gain greater understanding of how their evaluation processes work.

In addition to contributing to the knowledge base around school self-evaluation in New Zealand, this thesis also seeks to contribute to the larger international conversation around data use in school self-evaluation by focusing on several of the less-studied aspects of the topic. This thesis addresses two of the least-studied outcomes of school data use, change in teacher practice and organisational change (Coburn & Turner, 2011). This thesis also gives voice to the experiences and interpretations of teachers as it highlights how they interact with and respond to school self-evaluation processes. Both internationally and in New Zealand, teachers' work has been changed by the rise of school self-evaluation, but most research in this area has examined 'what works' in large-scale interventions rather than exploring teachers' perspectives and how they make sense of these new expectations.

Thesis Aims

The overarching aim of this thesis is to explore the actual practice of data use and self-evaluation activities, the perspectives of those embedded in these activities, and the ways these activities fit within wider contexts. From this broad aim, further questions are addressed in the various articles. Among other questions, this thesis asks:

- What underlying theories have informed teachers' data use, both in the literature and empirically?

- How do teachers experience the changing nature of their work as a result of the new expectations for data use in school self-evaluation?

- How have teachers and school leaders acquired the capacity to carry out these new expectations and new ways of working?

- Has school self-evaluation led to organisational change, and if so, how?

- How has the accountability context influenced the processes of data use and school self-evaluation?

- What aspects of the overall educational context have influenced these processes?

Opening the "Black Box" to Understand Change Processes

Much of the data use literature has focused on variables that lead to favourable outcomes with less attention on the actual change processes of school self-evaluation (Coburn & Turner, 2011; Mausethagen, Prøitz, & Skedsmo, 2018; Ozga, 2017). In contrast, this study explores how school self-evaluation processes shape organisational change, changes in practice, changes in perception, and changes to teachers' work. This research seeks to "open the black box" to understand the processes of school self-evaluation. The concept of the "black box" refers to research that focuses more on inputs and outputs while viewing the intervening process as an unobservable "black box". Close-up studies opening the black box of school self-evaluation offer a way to better understand teachers' work in today's schools and highlight the broader context in which they occur. Zooming in the lens to study "work processes and relationships is a crucial means for grasping macro-level developments and for shaping the ways in which we theorise about" the changing nature of educators' work (Little, 2012, p. 145). While close-up studies of a relatively small sample could be dismissed by some, particularly those with positivist inclinations, learning *how* the world works is what makes it more generalisable.

Opening the black box to explore data use and self-evaluation activities and the perspectives of those educators embedded in these activities calls for a research approach that will enable this, both conceptually and methodologically. A realist way of thinking proves to be effective to the exploration in this thesis, since this is an approach that specifically focuses on how and why things work in different contexts (Emmel, Greenhalgh, Manzano, Monaghan, & Dalkin, 2018; Pawson, 2013). Using a realist approach, which emphasises surfacing underlying influences, allows me to examine the implicit theories of my participants and of the literature I interrogate. A realist approach also emphasises human agency and highlights that teachers are active agents who interpret and make sense of new expectations on them (Maxwell, 2012). Realists believe that understanding the ways that lived experience interacts with – and in turn influences – processes of change can lead to deeper levels of explanation and understanding (McEvoy & Richards, 2006). Using a realist approach, this thesis examines the pathways or mechanisms of these change processes. Briefly, mechanisms are the underlying processes or social structures that operate in particular contexts to bring about change in the reasoning and behaviour of participants. Studying these mechanisms of change creates increased understanding of the ways these processes shape the work of teachers (McFadden & Williams, 2020). In this thesis,

qualitative methods focus on exposing these underlying change processes and foregrounding teachers' perspectives. A realist approach also highlights the context and its important role in shaping change processes. Context does not simply mean the setting; in complex social systems a range of contextual differences influence outcomes, including individuals' characteristics and capacities, interpersonal relationships, the organisational setting, and the broader policy environment (Astbury & Leeuw, 2010). In sum, using a realist approach allows me to explore how data use and school self-evaluation bring about organisational change and changes to teachers' work, highlights that the outcomes are not uniform, and argues that the context shapes these outcomes.

Reflexive Stance

Consistent with a realist approach, I have taken a reflexive stance when combining the individual articles into a thesis. A reflexive stance requires a recognition that the standpoint of the researcher is inevitably involved in the research process. The need for explicit reflexivity and openness about my decision-making during the research process was important because of the unique position of this thesis straddling two fields of study, evaluation studies and critical education studies. Because the research cuts across these two fields, the articles draw on different bodies of literature and different theoretical lenses, and are directed at different audiences, sometimes an evaluation audience and sometimes an education audience. In this thesis, I've used some of the sections that surround the individual articles for reflecting on and being transparent about the processes I followed, some of the decisions I made, and some things I struggled with, particularly in my data analysis activities. Qualitative researchers explain that allowing our decisions, knowledge construction, and analysis processes to be open to inspection enhances the credibility and trustworthiness of the research (Maxwell, 2013; Miles, Huberman, & Saldaña, 2014; Patton, 2015). This stance also impacts the writing style of various sections of the thesis; for example, the more formal academic articles are preceded by a less-formal introduction where I explain some of my thinking using a reflexive first-person active voice.

Thesis Overview

There are nine chapters following this one. The remainder of the thesis is organised as follows:

Contextualising the Thesis: A brief introduction to school self-evaluation provides a background of the most pertinent literature on school self-evaluation and data use. It presents an overview of the way school self-evaluation manifests differently in different countries and how this is connected to accountability agendas. It also introduces the multiple agendas at play in New Zealand school self-evaluation.

Context: The New Zealand Education System provides a summary of the salient aspects of the New Zealand context, focusing on elements of the context that impact school evaluation. This chapter shows that multiple competing agendas influence how and why school self-evaluation unfolds the way it does and shows that New Zealand's education system simultaneously encompasses two conflicting cultures of data use. The four articles in this thesis examine the complexities of teachers navigating the multiple competing agendas at work in the New Zealand context.

The Role of Theory in the Thesis is a brief chapter explaining the meaning and role of theory in this thesis. I draw on multiple theoretical lenses to account for the nuances and multiple stories in my data.

Methodology and Method outlines and explains the methodology and methods used in this thesis. It articulates the underlying assumptions of the research paradigm I draw on, realism, and how these inform the methods employed to explore the multifaceted stories of school self-evaluation in New Zealand secondary schools. I discuss the unique realist approach to thinking about change processes and causality, and the importance of the concepts of context, mechanism, and outcome. After describing the realist approach to qualitative methods, I explain the three phases of my study.

Article One *Walking on a tightrope: Agency and accountability in practitioner inquiry in New Zealand secondary schools* explores how inquiry is conceived in the literature and how it is enacted and experienced by New Zealand secondary school teachers. I find that the literature on inquiry conveys conflicting advice, competing demands, and contradictory purposes. Using interviews with educators, this paper suggests that inquiry is a central, yet contested, concept; teachers feel tension when enacting inquiry; and their inquiry processes occasionally slip into accountability.

Article Two *Data use in New Zealand secondary schools: Tracking, traffic lights, and triage* explores the perspectives of teachers and school leaders whose schools had recently

implemented a tracking and monitoring data use tool. The results highlight that data use processes led to changes in practice in teachers' work and contributed to structural changes in these schools. This study shows that data use can be a powerful force, with the potential for good, but it also raises some concerns about the unintended consequences of the use of assessment data.

Article Three *Drawing on theoretical knowledge to build evaluation capacity* explores how teachers and school leaders are engaging in evaluation capacity building (ECB) within their own organisations. In contrast with many other ECB efforts in education, these schools emphasise building capacity for sense making and data interpretation. The findings highlight that educators appear to draw on learning theory to build organisational evaluation capacity. This article explores how learning theories can be used as a framework for ECB, and how the field of evaluation could benefit from theories about learning from the field of education.

Article Four *Beyond the rational model of data based decision making: a conversation with evaluative thinking* is a more conceptual piece than the other three. In this article I trace the history of the Data Based Decision Making field and then problematise this literature to examine what theories are embedded in it. I find that it predominantly starts from a linear, rational model of decision-making. Recent research on the way that teachers actually engage with data in their everyday work reveals that this dominant discourse has often not captured the reality that teachers experience with data use and suggests that these processes are more complex, contextual, and less rational than the linear model might suggest. This article seeks to extend this understanding and proposes drawing on some of the conversations about the construct of Evaluative Thinking to inform our conception of data-based decision making (DBDM).

Discussion and Conclusion provides a summary and analysis of some of the key themes of thesis. It also provides a review of some of the content, methodological, and theoretical contributions and offers implications for further consideration.

Chapter Two

Contextualising the Thesis: A Brief Introduction to School Self-Evaluation

In order to further orient the reader to the focus of this thesis, this chapter provides a brief background of the most salient literature on school self-evaluation and data use. It presents an overview of the way school self-evaluation manifests differently in different countries and how this is connected to accountability agendas. A theme of this thesis is that these accountability agendas influence the way that school self-evaluation unfolds in New Zealand.

The International Rise of School Self-Evaluation

Internationally, the way that schools are evaluated has changed dramatically over the last few decades. Nearly every education system in the developed world has been busy creating or reforming their school evaluation policies and procedures (McNamara & O'Hara, 2008a). A major trend is the increase in school *self*-evaluation, with schools increasingly becoming self-evaluating organisations (MacBeath, 2006; McNamara & O'Hara, 2008; Schildkamp & Vissher, 2010; Vanhoof, Van Petegem, & De Maeyer, 2009). School self-evaluation is now a major force in the discourse on school evaluation and "the direction of school and teacher evaluation in many education systems is undoubtedly towards internal, school- based self-evaluation" (McNamara & O'Hara, 2008b, p. 43). A move toward the use of school self-evaluation reflects several wider trends, such as an emphasis internationally on the use of evidence as a basis for school improvement processes and the trend of turning over financial responsibility, decision-making power, and accountability for student success to individual schools (MacBeath, 2006) as well as grassroots movements encouraging schools to become learning organisations and teachers to be reflective practitioners (Cochran-Smith & Lytle, 2009).

School self-evaluation is an internal process to support decision making, ensure quality, and improve the teaching–learning process. In a nutshell, school self-evaluation is evaluation conducted by the school itself. A useful definition of self-evaluation has been offered by Schildkamp and Vissher (2010):

> a procedure involving systematic information gathering that is initiated by the school itself and intends to assess the functioning of the school and the attainment of its educational goals for purposes of supporting decision making and learning and for

fostering school improvement as a whole. (p. 1)

Mutch (2012) echoes that an important purpose of school self-evaluation is a focus on continuous internal improvement and strategic decision making. Teachers are increasingly expected to reflect on their own practice and to participate in collecting, analysing and interpreting data on department and school-level functions (ERO & MOE, 2015). As an improvement cycle, school self-evaluation typically involves gathering data that allows the school: to judge how it is doing in relation to a specific area; to identify strengths and areas for improvement and to develop and implement actions that lead to improvement (O'Brien, McNamara, O'Hara, & Brown, 2019). A number of terms are used synonymously with school self-evaluation in the literature and these include: internal evaluation, DBDM, and self-review (O'Brien et al., 2019).

Data Use as Part of Self-Evaluation

A key part of school self-evaluation is school personnel's use of data. Data use has become an important part of educational reform in various countries around the world. Schools are expected to use data from school self-evaluation processes that relate to school priorities and professional development at the individual and school levels (Clinton & Dawson, 2018). Internationally, teachers' work is changing, as they are increasingly expected to interact with various sources of data in the course of their practice. "Teachers' work is transforming as the kinds of available data multiply and expectations for their use begin to shift" (Little, 2012, p. 163). Using data is now part of the professional work of being a teacher, as teachers are expected to be reflective practitioners (MOE, 2007). The international trend is that "research and evaluation capabilities are no longer optional extras for the professional work of teachers, but are required skills" (McFadden & Williams, 2020, p. 13).Schools collect a wealth of data, including student assessment and achievement data, attendance data, student demographic information, classroom observations, student work samples, and survey results from students, teachers, and parents. While there was an international trend to define "data" in a narrow way to signify quantitative standardised assessment data, definitions of "data" and "data use" can be intentionally conceptualised very broadly so that they encompass all relevant quantitative and qualitative information about pupils, teachers, parents and schools. Data can be thought of broadly as any information that is systematically collected and organised from relevant stakeholders (e.g., parents, teachers, and students) to examine particular aspect of the school (Lai & Schildkamp, 2013). All of the articles in this thesis

address data use in some way and specifically Article Four traces the evolution of the field of DBDM and the ongoing changes in defining and theorising data and data use.

The literature extols many benefits of data use in education. A New Zealand seminal synthesis of the literature found that schools that gather data about their students and their practices and use these data to inquire into the effectiveness of their teaching and school practices made significant improvements in student achievement (Timperley, Wilson, Barrar, & Fung, 2007). Research on school improvement and school effectiveness suggests that data use is central to the school improvement process (e.g., Daly, 2012; Schaffer, Reynolds, & Stringfield, 2012). Using data can enable teachers and school leaders to tailor their teaching, school practices, and curriculum to address student-learning needs. Some scholars highlight the equity goal of data use and argue that expanded data use practices allow schools to operate more equitably (Bertrand & Marsh, 2015; Datnow, Greene, & Gannon-Slater, 2017; Selwyn, 2014).

It could be argued that since data use practices are so widespread and have taken on such importance, contemporary education cannot be fully understood without focusing attention on the particulars of data use in schools and on educational research that makes data use practices visible (Coburn & Turner, 2011; Selwyn, 2014). Scholars call for educational research to develop nuanced approaches to understanding what it means to work within what Selwyn (2014) refers to as "data deluged conditions" in educational contexts. Despite widespread attempts to encourage data use, the field is a relatively new area of scholarship that remains undertheorized (Mausethagen et al., 2018).

School Self-Evaluation Manifests Differently in Different Countries

School self-evaluation and data use practices have developed in different ways in different countries, often due to their unique social, political, educational, and historical context. These differences include the legal requirements for schools to conduct self-evaluation; how self-evaluation is conceptualised; how it links with external evaluation; the frequency of the process; its uses, outputs, outcomes; and the consequences for engagement and non-engagement by schools (OECD, 2013). This manifests differently in different countries and involves a mix of school self-evaluation/external evaluation, top-down/bottom-up, internal/external accountability, high-stakes/low-stakes. In some countries, strong external forms of school and teacher evaluation have developed and many governments, such as the US and Australia give rewards or sanctions in response to attaining targets on standardised

assessments. For example, in the US, the system is based on teacher evaluation and student performance data from standardised large-scale state assessments, while in many other countries, particularly in Europe and the UK, external evaluation usually relies on school inspection. In some European countries, such as in Germany, Ireland, and Scandinavia, there has been little emphasis on the use of data for internal and external evaluation (McNamara & O'Hara 2008b). In other parts of continental Europe, a high degree of school autonomy means a high use of data for internal evaluation processes (McNamara & O'Hara, 2008b). In England, a country that has had strong external evaluation based on school inspection, school self-evaluation is a highly-charged process, centrally regulated and associated with quality assurance (Hall & Noyes, 2009). Sahlberg (2011) describes recent educational policies as 'GERM'—the Global Education Reform Movement and lists five reform strategies that have 'infected' such countries such as Australia, the US and England: standardisation; a focus on core curriculum subjects at the expense of areas such as creative arts; risk-avoidance; corporate management models, and test-based accountancy policies. These various examples of school evaluation in different countries show that the context influences how school evaluation occurs and hints at the fact that one of the main reasons these processes manifest differently is that they stem from different accountability agendas.

Two Competing Agendas

It has been argued that the push for school self-evaluation and data use come from two competing agendas—one for external accountability and one for internal accountability (Cochran-Smith & Lytle, 2009; Gannon-Slater et al., 2017; Hargreaves & Braun, 2013; Hofer, Holzberger, & Reiss, 2020; Lai & Schildkamp, 2016; Schildkamp et al., 2013). In an external accountability framework, teachers and schools are held externally accountable for student educational outcomes, primarily through large-scale summative assessment data. These high-stakes testing data are used to show the state, region, and/or school inspection office how well a school is performing (Booher-Jennings, 2005; Diamond & Spillane, 2004; Lai & Schildkamp, 2013; Wohlstetter, Datnow, & Park, 2008). A culture of external accountability tends to be reactive and is motivated by the need to raise assessment scores as an end in itself (Lai & Schildkamp, 2016). This approach is generally associated with top-down (government or district driven) accountability and can come in the form of: school inspection; centrally-administered standardised tests; explicit formal measurable standards of performance and success; externally-imposed targets; and/or possible rewards or sanctions in response to attaining targets. This reflects the macro international trend that individual

schools are increasingly held accountable for the academic success of their students (OECD, 2013). While there are different approaches in different countries, the ultimate aim is the same, that schools need to ensure that all students achieve some sort of minimum standard or benchmark.

On the other hand, a culture of internal (teacher or school) accountability emphasises a more developmental, improvement focus. This is a 'bottom-up' approach that emphasises teacher inquiry, teacher expertise, and teachers as knowledge generators (Cochran-Smith & Lytle, 2009; Earl & Katz, 2006; Lai & Schildkamp, 2013; Sinnema & Aitken, 2011; Wayman, Midgley, & Stringfield, 2006). Fullan and colleagues (2015) describe internal accountability as when individuals and groups "willingly take on personal, professional and collective responsibility for continuous improvement and success for all students" (p. 4). This approach also includes the use of assessment data for decision-making, but may emphasise formative assessment, and positions assessment data as just one among multiple sources of a wider range of data including other academic and non-academic information to inform instructional decision making. As an internally-referenced approach, educators may be freer to orient their data use efforts around a variety of problems in their local context (Datnow et al., 2017). Internal accountability also emphasises a particular mindset, a means of viewing the world, a way of thinking recently termed Evaluative Thinking (Bennett & Jessani, 2011; Clinton & Dawson, 2018; Cochran-Smith & Lytle, 2009; Earl & Timperley, 2015; McFadden & Williams, 2020). Since our understanding of the construct of Evaluative Thinking is evolving, the theory around it is explored in Article Four.

The literature notes that high-stakes external accountability systems are more likely to put pressure on teachers to use data in certain ways that may lead to the misuse and abuse of data and demonstrates the negative outcomes of accountability policies on teachers' pedagogical practices and on students (Booher-Jennings, 2005; Farrell & Marsh, 2016; Lingard & Sellar, 2013; Mandinach & Schildkamp, 2020; Mausethagen et al., 2018). External accountability can work at cross-purposes to improvement efforts and can result in distraction from core purposes and gaming of the system, especially when test-based accountability means punitive consequences for failing to meet numerical targets (Hargreaves & Braun, 2013). For example, in the US, an external-accountability context of the policies of No Child Left Behind Act and Race to the Top has resulted in increasingly negative consequences on public education, because the prime drivers of educational change have been high-stakes and high-threat accountability, rather than improvement (Hargreaves & Braun, 2013). One of the

practices that may occur when data use is inextricably connected with high-stakes accountability systems is that educators tend to focus remediation efforts on a small group of students on the cusp of proficiency (Booher-Jennings, 2005). I show that some of these practices occur in New Zealand and discuss the ramifications of this topic in more depth in Article Two.

These two different agendas of internal and external accountability influence how data are perceived and used as part of decision-making (Lai & Schildkamp, 2016). Teachers' data use is influenced by the wider accountability context in which they teach. Although external accountability is meant to contribute to improvement, there are often direct conflicts between the twin purposes of improvement and accountability (Downey & Kelly, 2013; Hargreaves & Braun, 2013). The literature conveys that data use can lead to positive outcomes and greater quality *or* to negative outcomes and reduced equity depending on the accountability context (e.g., Datnow et al., 2017). The education systems in many countries have "struggled to balance the at times apparently irreconcilable demands for public accountability and oversight alongside those of professional development and autonomy" (McNamara & O'Hara, 2008b, p. 139).

Two Agendas, but More Than a Dichotomy

The literature conveys that school self-evaluation and data use come from two competing agendas—one for external accountability and one for internal accountability—and that these different accountability agendas shape how assessment data are used (Cochran-Smith & Lytle, 2009; Gannon-Slater et al., 2017; Hargreaves & Braun, 2013; Hofer et al., 2020; Lai & Schildkamp, 2016; Schildkamp et al., 2013). While this concept is certainly true, this thesis will argue that the situation is more complex than a simple dichotomy. In fact, school leaders and teachers often have to navigate both agendas simultaneously (Lai & Schildkamp, 2016). Research on data use is often conceptualised in a dichotomous way, but this thesis will show that in New Zealand secondary schools internal and external accountability co-exist and that teachers have to negotiate between these competing discourses (Mausethagen, 2013).

The school self-evaluation situation is more complex and ambiguous than a simple dichotomy. While much of the literature characterises education systems as being *either* high-stakes external accountability *or* low-stakes internal accountability, governments and schools have a range of objectives they hope to achieve in undertaking school self-evaluation

(McNamara & O'Hara, 2008a). These include both external accountability on one hand and increased local decision-making, teacher professional development, and school improvement on the other. "In an ideal world, school evaluation might be able to deliver on all of these goals simultaneously" (McNamara & O'Hara, 2008a, p. 174).

Negotiating Competing Discourses of Accountability

Furthermore, explorations of the impact on teachers of the tension between the competing accountability agendas are less common in the literature (Mausethagen et al., 2018). When this tension is addressed, it is more commonly through examples of fairly simplistic statements of survey results showing that teachers believe data in their school is collected for external accountability reasons but they thought it should be used for internal accountability reasons and this demonstrates that a tension exists (e.g., Downey & Kelly, 2013). In the data use literature, the tension between internal and external accountability is often not addressed, yet it is important for understanding what happens when educators engage with data in their everyday work (Mausethagen et al., 2018). There is a dominating perspective in the data use literature addressing implementation and less attention is given to the contexts in which they occur and the constraints that are often placed on teachers in interactions around data use within the broader context of accountability (Mausethagen et al., 2018). In contrast, this thesis centres on these overlooked themes of context, constraints, and accountability as it attempts to provide nuance to the dichotomous view of accountability, show the range of perspectives that accountability can take, and explore how teachers negotiate these competing discourses.

One of the themes this study explores is the way that educators enact these processes, and the way they make sense of the conflicting accountability discourses at play in the New Zealand context. The thesis seeks to gain more understanding about what actually happens when people engage with data in the course of their ongoing everyday work and how that relates to changes in practice, organisational change, and the way these processes shape the work of teachers. This thesis explores the way that teachers navigate these new expectations and negotiate different discourses on data use and accountability. Each educational system is a mix of internal and external accountability and the following chapter on the context will show that New Zealand's education system simultaneously encompasses two conflicting cultures of data use. That chapter will discuss other aspects of the context that influence how school self-evaluation plays out in New Zealand, because accountability approaches closely

intersect with other characteristics of the system. Because there are multiple competing agendas at play in New Zealand school self-evaluation, the outcome is not uniform for teachers and schools. Therefore, an examination of school self-evaluation in New Zealand is not a singular story, but instead it is multiple, multifaceted stories. The articles in this thesis provide some of these stories.

Chapter Three

Context: The New Zealand Education System

This chapter explores some of the features of the New Zealand education context that are relevant to the study. Educational data use is interwoven with the larger policy, political, and accountability context of the country, so it is important to consider this macro context (Coburn & Turner, 2011). The curriculum, the assessment system and the way in which schools are governed impact these processes (Datnow et al., 2017). Mausethagen and colleagues (2018) argue that the data use literature has emphasised the design and implementation of data use practices and has given less attention to the contexts in which they occur including the broader context of accountability. The foregrounding of the context is also consistent with the realist approach, which informs this thesis.

This chapter will explain the New Zealand context of school self-management, equity issues, the curriculum and assessment structures, the unique system of complementary evaluation, some of the challenges of school self-evaluation, and efforts to build capacity in this area. It will also familiarise the overseas reader with some of the concepts and vocabulary used in the New Zealand education system. The New Zealand schooling system has several characteristics that make it an interesting context in which to study school self-evaluation.

Self-Managing Schools

An important feature of the New Zealand education system is that each school is self-managing. Education is largely funded by the State and largely provided by state (public) schools. In 1989 a policy called *Tomorrow's Schools* created a radical change when all state schools became self-managing, abolishing all layers of district administration, giving more responsibility to schools, and decreasing the former Department of Education to a smaller MOE (Education Act, 1989). These reforms meant greater independence, variety and flexibility for schools and increased decision-making responsibility for school leaders (Wylie, 2012).

In giving more responsibility to schools, the goals of the Tomorrow's Schools reforms were to: reduce bureaucracy, move educational decision-making closer to the community, improve educational outcomes for disadvantaged groups, and make schools more accountable to their communities (Timperley, 2013; Wylie, 2012). Wylie (2020) argues that the policy was premised on three underlying theories about improving school quality: that

increasing parental involvement in schools' governance would make schools more responsive; that giving parents school choice and basing school funding on the size of the student roll would give schools an incentive to compete; and that separating policy from operations would make both more efficient, and that accountability mechanisms such as regular reviews would improve quality. Court and O'Neill (2011) argue that the *Tomorrow's Schools* policy was presented to the public using a mix of: a) a social democratic argument for more parental participation in education in order to meet locally-determined educational needs and b) a neoliberal argument for the system to become more efficient and effective through enabling consumer choice and therefore increasing market-like accountability. From the beginning, this policy was a confluence of neoliberal discourses and progressive discourses that were similar to rhetoric around both internal and external accountability.

Each school in New Zealand is now an autonomous unit governed by its own board of parent-elected trustees and each school manages its own budget with funding from the MOE. The board of trustees has discretion to govern as it sees fit, subject only to the law on financial, employment, health and safety, and property matters. Perhaps because they are self-managing, New Zealand high schools have possibly more layers of school leadership than in other countries, including roles such as deputy principals, assistant principals, heads of department, and tutor group teacher (similar to a home room teacher in other countries). The responsibility for academic rigour and student achievement also lies with individual schools. A recent independent taskforce illustrates that "our current model has created almost 2,500 self-governing 'islands' that work for the most part independently of each other" (Tomorrow's Schools Independent Taskforce, 2019, p. 4).

The extent of this school self-management is unique to New Zealand. The OECD has noted that it has one of the most devolved education systems in the world (Nusche et al., 2012). As a critique of this system, Wylie (2012) argues that New Zealand is "the only country that has built its national school system on schools operating on their own" (p. 4) and that the freedom that New Zealand schools enjoy all too often means that schools need to reinvent the wheel. This governance structure means that schools design, implement and evaluate their own strategies and approaches. Wylie (2020) argues that New Zealand's 30-year experience with self-managing schools has meant that schools: are used to operating with considerable autonomy, have too much variability of quality, and need much stronger connections if the education system is to progress and better meet the needs of all its students.

In many other countries, when governments have transferred primary responsibility for student achievement from central authority to individual schools, governments have prioritised the dual goals of school autonomy and school accountability (McNamara & O'Hara, 2008a). Scholars argue that overseas, maintaining accountability in this situation has paradoxically led to *more* rather than less state control of education in the form of evaluation, inspection, testing, standards, and so on (Ball, 2003; Connell, 2009; McNamara & O'Hara, 2008a). Ball (2003) explains that these reform processes are not simply a strategy of de-regulation, they are processes of re-regulation; it is a much more hands-off, self-regulating regulation that he referred to as steering from a distance.

> Such developments are deeply paradoxical. On the one hand, they are frequently presented as a move away from 'low-trust', centralized, forms of employee control. Managerial responsibilities are delegated, initiative and problem-solving are highly valued. On the other hand, new forms of very immediate surveillance and self-monitoring are put in place; e.g. appraisal systems, target-setting, output comparisons (p. 219).

Similarly in New Zealand, some scholars argue that central government education agencies 'steer at a distance' (Court & O'Neill, 2011; Wylie, 2020). This particularly happens through what one Secretary for Education described as a 'tight-loose-tight' relationship between schools and central government agencies (Wylie, 2020). A central government agency sets the standards and the outcomes desired ('tight'); the schools are free to respond to these according to their specific context ('loose'); and the schools report their results in relation to outcomes ('tight'). Court and O'Neill (2011) argue that the *Tomorrow's Schools* reforms had been presented as the opportunity to meet *locally determined* educational needs and priorities, but resulted in market managerial steerage mechanisms through which schools, principals and teachers would all be held accountable in various ways for the school's progress toward meeting *centrally-mandated* national educational goals for student outcomes.

There are different views on whether the New Zealand education system is a high trust, internal accountability-focused system or a low trust system. Authors such as Timperley (2013) have called the New Zealand system of self-managing schools a high-trust system. While there is an appearance of a high-trust system, a recent national taskforce (2019) said they were "saddened and alarmed" by the absence of trust in the system. They argue that this

low-trust system is characterised by schools' need to monitor their compliance, increased unproductive workload and stress, often unnecessary performance reporting, and an undermining of professional commitments to ensure that all learners succeed. In other words, an emphasis on external accountability and an undercutting of internal accountability.

A characteristic of New Zealand's system of self-managing schools is that it is now structured around competition for student numbers, with students and parents becoming akin to schools' clients (Smardon & Charteris, 2017). Since parents are free to choose schools for their children, parents become like clients to whom schools need to market themselves (Court, 2004). Scholars argue that this feature of New Zealand's education system is underpinned by neoliberal principles of public policy (e.g., Court, 2004; Dale & Ozga, 1993; Smardon & Charteris, 2017; Thrupp, 2017). Neoliberalism is essentially an economic philosophy grounded in the belief that the free market uses resources most effectively and efficiently and that the responsibility of the state is to ensure that such a "free market" exists in the public spheres, including education (Harvey, 2005). The neoliberal view is that individuals should make their own choices of welfare services just as they make their own choices of groceries, cars and hairdressers (Dale & Ozga, 1993).

Authors argue that accountability becomes a market version of accountability (Court, 2004) or consumer accountability (Ranson, 2003) whose purpose is to strengthen consumer responsiveness through market competition, consumer choice, and assessment data. Each school can be thought of as "a firm competing with all others for students, marks and money, in markets where parents are consumers expected to exercise 'choice' between different firms/schools" (Connell, 2015, p. 183). In this system of competition, one example of the pervasiveness of the consumer accountability discourse is demonstrated when popular magazines such as *Metro* and *North and South* publish high schools' NCEA qualification results in their "Best Schools" issues. *Metro* promotes this issue as a way of helping parents to make "the right decision" about schools (Metro's best schools in Auckland, 2019). In a context where competition is fostered, these publications draw the public's attention to schools' NCEA results. I admit that as a parent, I carefully read each of these issues myself every time they were published. The implications of consumer accountability in New Zealand are particularly discussed in Articles One and Two.

Equity Issues

New Zealand's cultural, historical, and socioeconomic context also impacts its education system and highlight its student equity issues. New Zealand today is characterised by increasing levels of immigration, increasing cultural diversity, and increasing segregation by socioeconomic circumstances (Tomorrow's Schools Independent Taskforce, 2019). Its history is characterised as a colonised territory of the British Empire. New Zealand's foundational document, the Treaty of Waitangi between the British crown and the indigenous Māori, established New Zealand as a bicultural nation. The obligations in the Treaty still remain in education policies, with the education system expected to honour the Treaty and support Māori-Crown relationships. Therefore, it is a legislative mandate from the Treaty of Waitangi that any educational evaluation in New Zealand must consider the implications of any programming, projects, or other interventions for Māori. However, the rhetoric around a commitment to equity has not always led to the outcomes to which the education system aspires. Despite the status of Māori as a Treaty partner, the education system has not always produced equitable outcomes (ERO & MOE, 2015).

The New Zealand schooling system is characterised by inequality (Organisation for Economic Co-operation and Development., 2019). There is a strong relationship between ethnicity and school achievement, with Māori and Pacific Island students being over-represented among lower achieving students, while Pakeha and Asian students are over-represented in the higher achieving groups (Organisation for Economic Co-operation and Development., 2019). Māori students leave school with lower levels of achievement than their peers. There is also a strong relationship between socioeconomic background and achievement; the impact of socioeconomic background on Programme for International Student Assessment (PISA) scores is greater and has increased by more in New Zealand than in other wealthy countries (Organisation for Economic Co-operation and Development., 2019). In addition, while New Zealand as a whole performs above the OECD average on PISA exams, performance has slipped over the last 15 years on PISA reading, science, and maths exams, resulting in the lowest scores ever recorded last year. A recent national taskforce argued that this "legacy of system failure to invest the necessary resources in achieving equity…[is a] 'debt mountain' that has been amassed over generations at the expense of significant groups of learners" (Tomorrow's Schools Independent Taskforce, 2019, p. 13). In response to these challenges, the New Zealand MOE prioritises a national effort to reduce the achievement disparities within and across schools and improve education

provision and outcomes for all students, particularly priority learners (Māori, Pacific, and low-income students) (ERO & MOE, 2015). Improved Māori student achievement has been a key government priority in education over the last several decades.

Attaining more equitable outcomes for priority learners can be assisted by effective data use and school self-evaluation. Authors explain that expanded access to data allows schools to operate more equitably (Selwyn, 2014), the goal of data use is success for all students (Datnow & Park, 2018), and data use can help accomplish the equity goal of supporting all students to reach their full potential (Bertrand & Marsh, 2015; Datnow et al., 2017; ERO, 2015). Efforts to raise achievement for priority learners are linked to effective school self-evaluation practices in New Zealand, with an ERO report finding that effective secondary schools with equitable outcomes consistently have strong self-review systems (ERO, 2014). The University of Auckland's Starpath Project (Kiro et al., 2016) found that for Māori, Pacific and other students in low-decile schools to be successful, schools need data that enables student progress to be tracked over time. However, data is not always used effectively to support priority learners. As recently as a decade ago, ERO reported that despite the widespread information and support available, a substantial proportion of schools did not review their own performance in relation to Māori student achievement or make effective use of data to improve classroom programmes and school-wide systems to promote success for Māori (ERO, 2010).

Curriculum and Assessment

New Zealand's flexible curriculum and assessment structure are another important feature of the education system. The New Zealand education sector, including its curriculum and assessment, has a history of being quite progressive and many aspects continue to be (Mutch, 2013). For example, the national curriculum is not tightly prescriptive, but instead is adaptable, avoiding standardising teacher practice and allowing each school to select the curriculum emphasis and materials for its own teaching and learning programme (Starkey & Eppel, 2019). This high level of teacher agency in setting curriculum content builds on the New Zealand education system's history of teacher autonomy (Mutch, 2017). There are many advantages to this curriculum flexibility, including that schools and teachers can be responsive to local contexts and students' needs (MOE, 2007). "Schools have generally been enthusiastic about the NZ Curriculum, and regard it as a 'national treasure'" (Wylie, 2020, p. 365). However, some researchers have found a downside of this responsiveness; combined

with pressure for all students to succeed, this can result in students from lower socioeconomic schools receiving an, easier, watered-down curriculum with fewer opportunities to engage with challenging material (Wilson, Madjar, & McNaughton, 2016).

Another key component of the *New Zealand Curriculum* (NZC) is its inclusion of Teaching as Inquiry as an essential aspect of effective pedagogy (MOE, 2007). In the national curriculum, Teaching as Inquiry provides the expectation that teachers use data to reflect on their pedagogical reasoning and inform their teaching decisions. Teaching as Inquiry is described as a cyclical process in which educators pose questions, gather evidence, and make decisions about their practice. The New Zealand education system has a focus on viewing teachers as adaptive experts who have the ability and autonomy to make decisions based on data to solve problems at the local level (MOE, 2007). As will be discussed further in Article One, the initial intent of Teaching as Inquiry was an internal accountability that holds teachers accountable for holding an inquiry disposition (Sinnema & Aitken, 2011).

The NCEA is the secondary qualification system offered in the final three years of schooling in New Zealand. NCEA was designed by progressive educational policy makers who had equity in mind and were "rightly concerned" about young people leaving school without qualifications, especially those from low socioeconomic backgrounds (Hipkins, Johnston, & Sheehan, 2016, p. 106). In the NCEA system, students accrue credits when they demonstrate that they have reached required standards. It involves internal assessments made by students' own teachers and external assessments made by independent assessors. Internal assessments are moderated for accuracy and consistency of teachers' judgements against the standards (Wylie, 2020). Students are awarded a NCEA qualification at levels one, two, or three when they have completed a sufficient number of standards at the corresponding level. While there is flexibility in the system, the central government agency NZQA sets the standards, based on the New Zealand Curriculum. In 2009 the practice of reporting against accountability targets based on achievement measures was introduced. The structure of NCEA, the national curriculum, and decentralised decision-making policy have provided flexible conditions for New Zealand secondary schools. However, during the time of data collection for my study, the National Government (not the Education Review Office or the Ministry of Education) had set a Better Public Service target that 85% of students would achieve Level 2 of NCEA, which made this target essentially the indication of school quality. The impact of that target in relation to school self-evaluation and the NCEA system will be highlighted in Article Two.

Schools Self-Evaluate

Because New Zealand schools are self-managing, each school needs to develop self-evaluation systems (Timperley, 2013). New Zealand school evaluation sits in contrasts to school inspections in some other countries, since school self-evaluation is the centrepiece of school evaluation, emphasising good practice and school improvement more than compliance (Mutch, 2012). An OECD report noted that in New Zealand, "school self-review is at the heart of quality assurance and improvement processes. The basic premise is that schools are best placed to analyse their own contexts" (Nusche et al., 2012, p. 89). Schools are expected to carry out an "ongoing, cyclical process of evaluation and inquiry for improvement" (ERO, 2015, p. 6). This system has been called developmental, negotiated, localised, and collaborative (Lai & Kushner, 2013).

Each school is expected to: conduct ongoing self-evaluation of their policies, plans and programmes, including evaluation of information on student achievement; report to students and their parents on students' individual achievement; and report to the school community on the achievement of students as a whole and of groups of students, including specifically the achievement of Māori students to acknowledge New Zealand's bicultural heritage (ERO, 2015). Schools are required to use data about student achievement to evaluate the success of the schools' curriculum and teaching, inform strategic planning and school development, and set their own goals and objectives for student outcomes each year. Schools can select the approach and tools used for self-evaluation (ERO, 2015). Several years ago, ERO developed new research-based evaluation indicators that it encourages schools to use in their own self-review, and also provides evidence about good practice in schools through online reports and videos (ERO, 2016). This system of school self-evaluation is built on the theory that teachers and school leaders are agentic problem solvers. "By treating local educators as professionals who are able to perform a thorough self-evaluation, school personnel are able to take ownership of their schools and their school improvement plans" (Ryan & Timmer, 2013, p. 201).

External Evaluation

The New Zealand system is not solely built on school self-review, but is also integrated with periodic external evaluation performed by an independent government agency, ERO. ERO was established in 1989 and has undergone several shifts since that time. In the 1990s it emphasised external accountability and compliance, with statutory obligations and

accountability for the quality of education (Thrupp, 2017). In those days, the joke was that ERO are like wolves: they wear black and hunt in packs. Court and O'Neill (2011) argue that during the 1990s, a 'centralised decentralisation' allowed the state to steer at a distance local schools' implementation of centrally-mandated goals that were measured, recorded and reported by ERO. Since that time, ERO has evolved from a more external, compliance, accountability-oriented focus to a system that is more focused on improvement and is complementary, collaborative and negotiated (ERO, 2017; Lai & Kushner, 2013; Mutch, 2012). Timperley (2013) points out that this shift is in the opposite direction of many other aspects of the education system in recent years that have moved toward a stronger accountability agenda.

In the new system of *complementary evaluation*, the external evaluation of schools conducted by ERO is meant to complement schools' internal evaluation, which is now at the heart of school evaluation. "Participation in periodic external evaluation supports the school's cycle of continuous improvement through providing an external lens on the school's improvement journey" (ERO, 2015, p. 6). The OECD noted that, "New Zealand has probably gone furthest among countries internationally towards a collaborative school evaluation model" (Nusche et al., 2012, p. 95), with one of the strongest links between external and internal review. ERO reviews most schools every three years, and visits some schools more or less frequently than this depending on how well the school meets ERO review criteria. After the review, ERO gives schools a report that is publicly available online, and often on school websites. ERO teams have limited time in each school, and increasingly rely on documentation provided by the school (Wylie, 2020). It appears that in the 90's, ERO functioned as the loose-tight-loose model (Wylie, 2020) by having an external, compliance, accountability-oriented focus, but in recent years has had a model in which its external evaluation is collaborative, focused on improvement, and complements schools' internal evaluation, which is now at the heart of the system of school evaluation.

School Self-Evaluation is Difficult

While self-review is an expectation for New Zealand schools, many schools find these processes difficult. The small amount of literature on the topic shows that self-review is highly variable across New Zealand schools and only a minority of schools have strong processes in place (ERO, 2017; Mutch, 2012; Nusche et al., 2012; Timperley, 2013). Wylie (2012) characterises the variability this way: school autonomy provides the opportunity to

succeed but it also provides the opportunity to fail. Recently ERO stated that only 10% of schools it reviewed in 2016/17 were on a 'sustainable improvement path' (ERO, 2017). A national report indicates that schools are not widely using effective evaluation processes, with only one-quarter of schools in the sample judged to be effectively inquiring into their annual NCEA achievement data and then developing activities, innovations or approaches to improve achievement (ERO, 2014). Most of the remaining schools had invested considerable time analysing achievement information, but this process did not result in new approaches or innovations that improved achievement. Earl (2013) notes that New Zealand schools have the "*potential* for very productive continuation on a path of continuous improvement, inquiry-based and evidence-informed decisions, grounded professionalism and equity. These directions are not a given, however, nor are they firmly embedded in the culture of New Zealand schooling" (p. 186, emphasis added).

One of the challenges New Zealand schools face in conducting effective self-evaluation is their access to appropriate data management systems (e.g., computerised student information systems that include analysis and data warehousing functions). In a schooling improvement initiative, Lai (2013) found that the majority of schools, even with support, did not have robust enough data on which to build a self-review process. As mentioned previously, the Starpath Project found that a key barrier to tertiary study for underrepresented groups was the lack of long-term data on student participation and achievement (Starpath Project, n.d.). "Without this, it is very difficult for teachers and schools to identify when students fall off the pathway to tertiary education" (Starpath Project, n.d.). In addition, each school leader needs to decide which data management tool to use, how to use it and who will analyse data, a process that can be impeded by a lack of knowledge, time, or budget (Starkey & Eppel, 2019). Perhaps these data management systems are both more important and more complicated because New Zealand's NCEA assessment system itself is so complicated.

Another challenge of conducting effective school self-evaluation is the evaluative capacity of schools and teachers. Because New Zealand's school self-evaluation is so decentralised, its success is dependent on local educators and their capacity to carry out effective self-review. In an MOE initiative called Building Evaluative Capability in Schooling Improvement (BECSI), evaluative capability was believed to be crucial to, but lacking in, school self-review processes (Timperley, 2013). A recent report (Organisation for Economic Co-operation and Development, 2015) echoes the earlier OECD suggestion by Nusche et al. (2012) that learning in New Zealand "will be more effective if current efforts to ensure that

teachers and schools have the skills to collect, analyse and interpret data in order to support improved student outcomes are strengthened over time" (p. 46). But this is not easy. Ryan and Timmer (2013) remind us that evaluation capacity building in New Zealand is a "challenging and complicated process" (p. 198). A recent independent taskforce argued that this system in which each school is essentially a self-managing island has meant that schools are mainly left to learn on their own and that, "adoption of good practice is almost always referred to as patchy and the uptake of promising innovation is seen as slow to spread across the system" (Tomorrow's Schools Independent Taskforce, 2019, p. 16). In addition, the fragmented nature of the self-managing school model has meant that it is difficult to gauge who is doing what and with what tools (Starkey & Eppel, 2019).

Summary of the New Zealand Context

While the New Zealand education system has a progressive history, and elements such as the curriculum and pedagogy remain progressive, schools have been administered in a very neoliberal manner for more than 30 years, in a devolved education system of individual self-managing schools (Mutch, 2013; Thrupp, 2018). A system that includes school choice has led to a market version of accountability, which can lead to schools acting like businesses competing for parent clients in an education market (Court, 2004). Some aspects of New Zealand's progressive tradition have been under pressure in recent years and the climate in education has shifted to include a range of managerial-oriented government policies (Maguire, 2018; Smardon & Charteris, 2012; Thrupp, 2017). So teachers need to navigate progressive ideals constrained by neoliberal imperatives. Because the New Zealand education system is shaped by these competing theories, a progressive tradition with a neoliberal undercurrent, there are often conflicts in values and philosophical perspectives (Maguire, 2018). While some of the particulars of the context will be unique to New Zealand, the pressures of accountability are felt in most education sectors around the world.

This thesis explores school self-evaluation in a system marked by the loose-tight-loose arrangements of autonomy, agency, and accountability against a backdrop of school choice under neoliberal doctrine. These influences are reshaping the terrain of teachers' work and create tensions in the system for teachers. Since there has been little research on how teachers experience these school self-evaluation processes in New Zealand, this thesis attempts to "open the black box" to understand processes of data use for school self-evaluation and the way that educators enact these processes. This chapter highlights the

competing agendas in the New Zealand education sector that influence why and how school self-evaluation occurs. Teachers must navigate the competing discourses on internal and external accountability, which can create opportunities, challenges, tensions, and contradictions which are highlighted in the articles in this thesis.

Chapter Four

The Role of Theory in the Thesis

The Meaning of Theory

In order to explore these contexts and their impacts on school self-evaluation, I draw on multiple theoretical lenses to account for the nuances and multiple stories in the data. The purpose of this brief chapter is to explain the role and meaning of theory in this thesis. The role, and indeed even the meaning of the word *theory*, is more ambiguous and multifaceted in this thesis. The unique position of this thesis straddling two fields of study, evaluation and education, means there is a need to articulate the multiple meanings of the term theory.

Before explaining the role of theory in this thesis, it is necessary to explain what is meant by *theory* here, since the field of evaluation uses this word in multiple "fragmented" and "often confusing" ways (Leeuw & Donaldson, 2015, p. 478). There are at least four different ways this term is used in my field of evaluation. First, evaluators draw on "evaluation theories" which are equivalent to evaluation models or evaluation approaches. These are theories of practice for conducting evaluations, for example, empowerment evaluation, or participatory evaluation. Another category of theory is philosophy, which covers fundamental questions, such as ontology, which deals with the nature of reality, and epistemology, which is the nature of knowledge. Different philosophies also make different assumptions about causality, or what causes change. This thesis draws on a realist perspective and the methodology section will explain the ramifications of this for epistemology, ontology, and particularly causality.

Another use of the word theory, and one which is particularly relevant to realist evaluators is the concept of programme theory. Realist evaluation falls under the umbrella of *theory-based* evaluations. Programme theory explains how and why different outcomes are generated in different contexts. This concept has been influential in the field of evaluation, as many evaluators are familiar with a theory-of-change approach describing processes of change in an initiative by outlining linkages between the intervention and expected outcomes. Evaluators are accustomed to working with stakeholders to try to understand how those who implement programmes theorise the problem and how to bring about change. These evaluators may not realise that this is built on the realist assumption that every policy and every intervention contains underlying *theories* about what a particular problem is and

how to bring about change to improve the situation (Pawson, 2013). In this thesis, these *theories* are the ones held by the participants in my study.

The final category of theory is called alternately formal theory, substantive theory, explanatory theory, or scientific theory. These are the tested and robust theories that operate in different domains or disciplines, such as the social, behavioural, and policy sciences (Leeuw & Donaldson, 2015). These include theories such as behaviourism or rational choice theory. A substantive theory offers a framework for making sense of empirical data, functions as explanatory suggestions and provides a model or map clarifying and explaining something we want to understand (Maxwell, 2013). "A useful theory is one that tells an enlightening story about some phenomenon, one that gives you new insights and broadens your understanding of that phenomenon" (Maxwell, 2013, p. 49). Leeuw and Donaldson (2015) argue that the field of evaluation does not draw on these substantive theories as much as it should. This claim is central to my Article Three, which argues that the field of evaluation could benefit from attending to and drawing on explanatory theories.

My Relationship with Theory

As noted previously, this thesis is informed by a realist approach that encourages researcher reflexivity; therefore I will share my own prior experiences with theory. My experience in programme evaluation meant that I came to this PhD from a background as a decidedly *applied* researcher. I value research with real-world implications more than scholars whose work remains in the "ivory tower" and I viewed substantive theory as living in that tower. With a background in evaluation, a field that emphasises methods, I was not as accustomed to bringing in substantive theory. In evaluation, methods take centre stage, as I discuss in Article Four. The concept that *Method matters more than theory* is ingrained in us in evaluation. Evaluation "practitioners lament that [theorists'] ideas are far too theoretical, too impractical. Practitioners have to do the project work tomorrow, not jawbone fruitlessly forever" (House, 2003). This foregrounding of method over theory has perhaps occurred not just in evaluation, but may be reflected in much of the social sciences. Adams and Buetow (2014) explain there has been a deep suspicion of substantive theory and that the skills of doing theory have been neglected to a point where the nature and origins of an idea were not important so long as correct methods were being applied.

My Use of Theory in this Thesis

In this thesis I draw on multiple substantive theories and conceptual tools from different theoretical traditions. I explore theories across a range of fields, from adult learning theory to organisational change for different aspects of the study. Because I was not tied to a particular viewpoint, stance, theory, or "ideological hegemony" (Maxwell, 2013), I believe that I could look at the various issues with fresh eyes.

This thesis makes a contribution by exploring how theory in one field can inform the conversation in another field. Because this thesis is focused on school self-evaluation, this topic spans two bodies of scholarship that generally speaking, do not overlap: education and evaluation. Each field has its own journals and literature that do not usually cite each other, despite many overlaps in the processes that are researched in each field. I draw from both fields. An important aspect of my thesis is how theory in these different fields can contribute to the other. Article Three is primarily concerned with bringing a substantive theory from education into evaluation capacity building (Huffman, Thomas, & Lawrenz, 2008). Conversely, Article Four is primarily about bringing theory from evaluation into education and argues that the way the field of DBDM has been theorised is insufficient.

This thesis also brings in some critical theory, for example, by acknowledging the effect of neoliberal policies in schooling. Critical theory exposes and examines power relations that reproduce inequality (Gibson, 1986). Authors argue that many educational researchers do not engage with the underlying political and economic contexts in which schooling occurs (e.g., Connell, 2015; Ozga, 2017). This may be particularly true in the field of data use, as scholars who research data use in education often do not acknowledge or engage with the politics of the data work that they do, and generally start with the stance that data use results in improved performance (Ozga, 2017). The literature and policy predominantly treat data and data-driven decisions as neutral, objective, impartial, non-political, and scientific (Booher-Jennings, 2005; Selwyn, 2014; Spillane, 2012).

A realist approach and critical theory may not seem to complement each other, but they both focus on privilege, positionality, and experience; they both challenge ideals of rationality, objectivity, and knowledge hierarchies; and they both embrace complexity, contextualisation, and other theoretically and politically essential concepts. They also both aim to reveal the uninterrogated underlying influences that may not be immediately apparent or visible but that are nonetheless real and shape the actions which result. The following

chapter articulates the assumptions of a realist approach, the particularities of a realist approach to qualitative methods, and how these inform the methods employed to explore the multifaceted stories of school self-evaluation in New Zealand secondary schools.

Chapter Five

Methodology and Methods

The purpose of this chapter is to outline and explain the methodology and methods used in this study. I begin with a brief discussion of tensions in research paradigms in the social sciences before turning to the underlying assumptions of the research paradigm I draw on, realism. I discuss the unique realist approach to thinking about change processes and causality, and the importance of the concepts of context, mechanism, and outcome. After describing the realist approach to qualitative methods, I explain the three phases of my study.

Research Paradigm Tensions

Social science research has been described as experiencing a paradigm war between positivism on the one hand and interpretivism on the other. My field of evaluation has also been characterised by these two views, but has been heavily influenced by the positivist paradigm. Discourse in evaluation has been dominated for decades by underlying positivist epistemologies and mechanistic ontologies (Archibald, Sharrock, Buckley, & Young, 2018). Evaluation has had a dominant ethos of rationality, control, predictability, objectivity, accountability, and an emphasis on methods and procedures (Dahler-Larsen, 2012; Patton, 2018a; Schwandt, 2019). Evaluation's positivist underpinnings have been concerned with discovering "best practice" and answering questions about whether a programme (or any intervention) works, with a focus on such questions as "What works?" These questions focused on causal attribution are the classic aims of applied research and evaluation (more discussion of assumptions about causality will be presented later). In the positivist paradigm, the relationships between phenomena can be predicted through hypotheses and studied through measurement using quantifiable variables. On the other side of the paradigm war is interpretivism, which was developed as a critique to positivism and recognises the move from the concept of one observable reality, to the ontological concept that there are multiple realities constructed by the interpretations and meaning of people. Despite the gains of qualitative research in the late 20th century, many governmental and funding agencies have preferred to commission research and evaluation that is quantitative, experimental, and statistically generalisable (Tracy, 2010).

Realism and its Underlying Assumptions

More recently, methodologies have been understood not as either/or but as a continuum of possibilities between positivism and interpretivism/constructivism, with realism falling in between the two. This thesis is informed by a realist perspective, and particularly a realist approach to qualitative methodology. Realism is gaining increased attention as an alternative to both positivism and constructivism as a stance for research and evaluation in the social sciences. There are many forms of realism and ongoing philosophical debates among realist scholars. The realism I'm drawing from is espoused by philosopher of science Nancy Cartwright (Cartwright & Hardie, 2012); the evaluation researchers Pawson and Tilley (1997; Pawson, 2013), and the qualitative researchers Maxwell (2013) and Huberman and Miles (Miles, Huberman, & Saldaña, 2014). Various terms have been used for this version of realism, including empirical realism, subtle realism, emergent realism and scientific realism. While there are many forms of realist philosophy, Bryman (2016) notes that there are two major forms of realism: empirical (Emmel et al., 2018; Manzano, 2016; Pawson, 2013) and critical (Bhaskar, 1978; Danermark, Ekstrom, Jakobsen, & Karlsson, 2002).

Realist approaches combine a realist ontology (there is a real world that exists independently of our perceptions, theories, and constructions) with a constructivist epistemology (our understanding of this world is a construction based on our social experiences and interaction) (Maxwell, 2012). More specifically, they integrate an ontological realism – the belief that there is a real world out there in which human agency interacts with other aspects of context to bring about change – with epistemological constructivism – holding that our understanding of the world is socially constructed, that these constructions will inevitably influence any attempt to describe reality. Realism takes the view that entities really do exist independently of our theories about them, but denies that we can have any "objective" or certain knowledge of the world, and accepts that there are multiple valid perspectives of any phenomenon. "All theories about the world are seen as grounded in a particular perspective and worldview, and all knowledge is partial, incomplete, and fallible" (Maxwell, 2012, p. 5).

In realism, there is an underlying belief in a stratified nature of social reality, in that all human actions are embedded within a wider range of social processes (Pawson & Tilley, 1997). Much of the philosophical foundation of realism draws on the critical realist Roy Bhaskar who asserts that there are three domains of reality. The first domain is the empirical domain, the world as we experience it through our senses; whilst our experiences are a

critical part of the world, Bhaskar asserts that they cannot be used to define the world (Bhaskar, 1978). The second domain is the actual domain. This domain describes the actual events whether we experience them or not. The third domain is described as the real domain which refers to the underlying mechanisms that can produce events. Something is 'real' if it has a causal effect on behaviour. Realism treats human thoughts, feelings, values, and attitudes as just as real as physical objects even though they are not directly observable, which is a different stance than positivism (Maxwell, 2012; Patton, 2015).

Realist View of Causality

Realism is different from positivism in many of its premises and implications, one of the most significant being that realism offers a particular understanding of causality. Realists have been among the strongest critics of the view of causation that is typical of positivism and is still dominant in quantitative research. My understanding of and interest in the realist evaluation approach was spurred by my research into alternative ways to conceive of causality when researching a paper for a PhD-level class on qualitative inquiry that eventually became an article published in the *American Journal of Evaluation* (Gates & Dyson, 2017). This paper appears to have touched on a need in the field of evaluation for multiple ways to conceive of causality and has resonated with readers in the field; according to AJE, our article has been downloaded more than 1,400 times. I think the article resonated because of increased awareness of problems with experimental designs in complex contexts; the desire to gain greater understanding of the inner workings of interventions to increase the policy and practice relevance of evaluations; and increasing recognition of the importance of understanding context and navigating complexity, and thus methods that encompass these concepts.

Realism's distinct understanding of causal explanations stands in contrast to the traditional way of thinking about causality. For much of the history of social science research, a successionist framework has been the dominant way of thinking about and assessing causality. A successionist framework underlies two closely-related logics of causality, regularity and counterfactual. Regularity thinking is the basis for correlational studies and statistical techniques including survey research methodology and statistical analyses of data sets. Counterfactual logic is the basis for randomised controlled trials and quasi-experimental designs. The counterfactual view treats the actual process of causality as unobservable, a "black box," focusing simply on discovering whether there is a systematic relationship

between inputs and outputs, with little attention paid to how those effects are produced. Understanding causality as an unobservable black box denies that we can know anything about the hidden mechanisms behind these regularities (Maxwell, 2012).

In contrast to a successionist view of causality, the realist approach takes a generative view of causality. A generative way of thinking about causal relationships builds and verifies a theory-based explanation of how causal processes happen. The basic idea is that the things that we experience or can observe are caused by 'deeper', usually non-observable processes. In evaluation, this way of thinking is oriented towards understanding how, why, for whom, and under what conditions interventions work to produce specific results. This way of thinking assumes there are multiple possible causal pathways linking an intervention to an outcome. These alternative causal paths will only be true for certain people under certain conditions. Realists are interested in opening the "black box" between conditions and outcomes to identify and understand the "underlying entities, processes, or [social] structures which operate in particular contexts to generate outcomes of interest and change" (Astbury & Leeuw, 2010, p. 368).

Context, Mechanism, Outcome

One of the central arguments of generative causality, as explained by the realist evaluators Pawson and Tilley (1997), is that Context + Mechanism = Outcome (C + M = O). They advocate examining how mechanisms work within particular contexts to generate outcome patterns. *Outcomes* or outcome patterns are the consequences (anticipated and unanticipated) that result from the interaction of different programme mechanisms in different contexts. *Context* refers to relevant conditions that are likely to enable or hinder the activation of programme mechanisms. Context does not simply mean the setting; in complex social systems a range of contextual differences are likely to affect for whom and in what circumstances a particular programme or intervention works (or fails to work) (Astbury & Leeuw, 2010). These include differences in individuals' characteristics and capacities, interpersonal relationships, the organisational setting, the geographical and historical context, and the broader policy and institutional environment in which a programme is embedded. Or, put another way, context includes the characteristics of: i) the individual players, ii) their interrelationships, iii) the institutional location, and iv) the surrounding infrastructure (Pawson, 2013). This concept is more than an argument that causal relationships vary across contexts. Rather, it is an argument that the context within which a

causal process occurs is intrinsically involved in that process and an important factor in determining whether a program will work in a certain setting. The realist approach highlighting context sits in contrast to much evaluation and social science research in which the dominant approach to assessing causality has viewed *context* as "unwelcome noise, a confounding variable to be controlled for" (Pawson, 2013, p. 120).

Mechanisms describe what it is about a programme, policy or other intervention that makes it work. The definition of "mechanism" is contested and continues to evolve (Lemire et al., 2020; Schmitt, 2020; Westhorp, 2018). I draw on Pawson and Tilley's (1997) conceptualisation of mechanisms as a combination of resources offered by the social programme under study and stakeholders' reasoning in response. The resources offered may be material, social, emotional, political and so on. The interaction between these resources, and how participants interpret and act upon them, is the mechanism of change. Mechanisms are the processes, or interactions that generate change within an intervention and can include group and social phenomena as well as frequently occurring at the individual/psychological level, describing how people react when faced with an intervention (Punton & Vogel, 2020; Schmitt, 2020; Westhorp, 2018). They include the "choices, reasoning, and decisions that people make as a result of the resources provided by a program" (Punton, Vogel, & Lloyd, 2016, p. 2). In this conceptualisation, mechanisms are real, although usually invisible. Reasoning can be, but is not limited to, cognitive or emotional responses that are triggered by the resources of the programme. Importantly, reasoning is often hidden (Manzano, 2016). Thus, mechanisms are often invisible, so identifying and analysing mechanisms can be challenging.

Realists believe that understanding the ways that lived experience interacts with – and in turn influences – processes of change can lead to deeper levels of explanation and understanding (McEvoy & Richards, 2006). Research explicitly focused on mechanisms has been advocated for use in both of the fields that my study draws from: data use in schools (Coburn & Turner, 2011) and research *on* evaluation (Astbury & Leeuw, 2010). These scholars suggest that mechanism-based research can highlight the nature of change that occurs within organisations and illuminate how and why evaluation works to bring about desired change.

Evaluations that strictly adhere to a realist approach often try to create context-mechanism-outcome chains (Lemire et al., 2020). However, A frequent critique of realist evaluation is the difficulty in identifying these context-mechanism-outcome configurations (Lemire et al.,

2020; Marchal, van Belle, van Olmen, Hoerée, & Kegels, 2012). I experienced this first-hand when I attended a realist evaluation conference in Australia. There was much hand-wringing over whether something was a context, mechanism, or outcome in a particular situation. The lack of clarity around defining context and mechanisms in particular situations has raised dilemmas for evaluators (Astbury & Leeuw, 2010; Marchal et al., 2012). Fortunately, conceiving of and creating these chains is not the purpose of my research. I am not attempting to determine all steps of the school self-evaluation process. Rather, I explore the reasoning of teachers and school leaders involved in data use and school self-evaluation processes. Therefore, qualitative research methods are appropriate since they focus on understanding a situation from the perspective of the participants within it.

A Realist Approach to Qualitative Methods

Qualitative methods can be important in helping to understand mechanisms (Manzano, 2016; Maxwell, 2012; Pawson, 2013). Mechanisms that occur inside people's heads are not directly accessible to the observer. Pawson (2013) argued that close qualitative research is an ideal way of "mining" these mechanisms. "Mechanisms are embodied in the subjects' reasoning and they are best investigated therein" (Pawson, 2013, p. 21). I can recall in one of my early meetings with my supervisors, I discussed how I could get participants to specifically discuss mechanisms. I remember saying hypothetically to a participant, "yes, yes, but what I really want you to talk about is invisible mechanisms." How would I ask interviewees about mechanisms and get them to explain their ideas on this topic? Then I came across a particular qualitative method called realist interviewing (Manzano, 2016). Before describing my specific methods, it is important to understand some of the ways that qualitative methods are different using a realist approach, since qualitative research is more commonly associated with interpretivism and constructivism. Maxwell and Mittapalli (2010) claim that realist, process-oriented qualitative investigations deserve a more prominent place in social research. Maxwell (2012) asserts that the ontological underpinnings of realist research fit with the features of qualitative research and that he is 'puzzled' that realist research hasn't been more prominent within qualitative research. While realist research is not a specific method in the sense of a set of steps that can be followed, its methodological orientation (Pawson, 2013) does have specific implications for methods practice.

One of the implications of using a realist approach is the importance of theory and theory development. In realist approaches this means that it focuses on looking for and refining

explanations (Pawson & Tilley, 1997). Using a realist approach, the research process starts by theorising and then tests those theories; these are "refined and tested again and, in this iterative process, our understanding of the real world is also refined" (Manzano, 2016, p. 347). Pawson (2013) suggests the possibility of conducting qualitative work prior to fieldwork to elicit some hypotheses about why outcomes are varied. Then the fieldwork tests and refines these theories. Thus my study can be conceived of as occurring in three phases: 1) Theory gleaning, 2) Theory refinement, and 3) theory synthesis. In the first two phases, I am referring to participants' theories of what is occurring in their context and in the third phase I bring in substantive theory.

Phase One: Theory Gleaning

Part One: A novel approach combining realist and soft systems methodologies

The first phase of this study was what Manzano (2016) calls "theory gleaning." The first step involved reading New Zealand and international school self-evaluation literature and New Zealand policy documents. In New Zealand much of the empirical data on school data use has been in the context of large-scale schooling improvement interventions (e.g., Lai & McNaughton, 2016) and not in the context of everyday practice. Therefore, to learn about effective schools' normal practice, the next step I used was an open-ended qualitative activity with participants to develop initial ideas about the mechanisms at play in these processes, a technique recommended by Pawson (2013). Realist approaches are method-neutral (Marchal et al., 2012), which allows them to be combined with other approaches. This was done through a series of three focus groups with six heads of department and one deputy principal in one school. This school was chosen because it had previously been identified as successful at self-evaluation (more details will be provided in the sampling section).

The first focus group used a novel and creative method for this initial theory development, the use of Rich Pictures, a tool used in Soft Systems Methodology (SSM) (Checkland & Poulter, 2006). Rich pictures are a compilation of drawings, pictures, symbols and text that represent aspects of the particular situation or issue from the viewpoint of the person who drew them. Rich pictures offer a quick and efficient way to work with key stakeholders to better understand their 'problematical situation' (Checkland & Poulter, 2006). Rich pictures can show relationships, connections, influences, points of view and cause and effect, and can both record and evoke insight into a situation (Checkland & Poulter, 2006; Williams & van't

Hof, 2016). They are used in organisational development and action learning approaches as well as in counselling (Williams & van't Hof, 2016).

The primary use of SSM is in the analysis of complex situations where there are divergent views about a problem. SSM is "oriented towards tackling complex situations through the experiential learning of a group of participants" so that learning takes place through the iterative process of using and refining system concepts (Dalkin, Lhussier, Williams, Burton, & Rycroft-Malone, 2018, p. 87). Authors argue that is not necessary to use the whole Soft Systems Methodology to use Rich Pictures, and the literature confirms that researchers do just use some aspects of the methodology (Oakden, 2014; Williams & Hummelbrunner, 2011). While it is possible to merely use the method of Rich Pictures, Dalkin and colleagues (2018) argue that the epistemology of Soft Systems Methodology is compatible with realist approaches. Both SSM and realist approaches engage in developing hypothetical descriptions that reveal the underlying mechanisms of everyday life, and refine them over time through an iterative process (Dalkin et al., 2018).

In the first focus group, participants were split into two groups and asked to create a Rich Picture mapping the current situation of their school's self-evaluation. The question they were asked to consider was: How do your school's self-evaluation processes lead to school improvement? The instructions were to:

- Use drawings, pictures and symbols that represent the situation for you.

- Don't worry about the artistic quality of the drawings. It's ok if it looks messy or uses stick figures. We want good ideas, not perfect drawings.

- Only use words when you can't think of how to represent your ideas in another way.

- Draw everything that you think is significant or problematic.

Include:

- **Issues** – concerns, conflicts, agreements, goals, motivations and perceptions

- The **structures** – organisational structures, the people who are affected by the situation, the hierarchy of authority.

- The **processes** within the situation (ongoing activities & routines).

- The **resources** – people, skills, tools, particular data sources, etc.

- **Connections** you see between your various elements--relationships, influences, cause and effect. Relationships can be drawn as arrows to show the nature of the interrelationships (e.g., strong, weak, fast, slow, conflicted, collaborative, direct, indirect, etc).

While the two groups worked and discussed, I took field notes of their discussion. When they were finished creating their picture, each group described and explained their Rich Picture to the wider group and this explanation was recorded. The field notes of their discussion and transcript of their presentation, rather than the actual picture, was the meaningful data. An example of the data that came out of the Rich Pictures is: when describing a portion of their picture, a participant stated "here's the community, looking over the wall at what we're doing." She went on to describe a feeling that parents and members of the community were scrutinising the schools' results. This was the beginning of my understanding of the external accountability pressure that teachers feel.

Phase One, Part Two

The remaining two meetings with these participants were more standard focus groups involving semi-structured group interviews. The interview questions asked were mainly exploratory and the wording of those questions tried to ascertain how their self-evaluation processes work, for whom, and in what circumstance. They were behavioural questions asking participants to give specific examples of how the processes happen in their school. They also teased out some topics that were raised in the Rich Pictures exercise. Some of the questions included: Can you give me an example of where you think your school self-evaluation is working really well? Can you think of an example of a time that this type of process led to a structural change in the school? Do you think that school self-evaluation leads to school improvement in your school? The full list of questions can be found in Appendix C.

As with most semi-structured interviews, I also asked follow-up clarifying questions, such as "Do you remember how that decision was made?" As with many focus groups, there was also group interaction, with participants occasionally chiming in with additional information or asking clarifying questions. The focus group recordings were transcribed, analysed by

focusing on processes or intermediate outcomes of their self-evaluation, and used to inform the development of more precise questions in Phase Two, the realist interviews.

Phase Two: Theory Refinement

Realist interviews

The results of the Phase One focus groups were used to inform questions for one-hour-long individual "realist interviews" (Manzano, 2016; Pawson, 2013). Realist interviews are designed to test hypotheses about how processes work (or fail to work) in practice (Astbury & Leeuw, 2010; Manzano, 2016). In realist interviews, questions often focus on the "reasoning and resources" of participants and "the social and cultural conditions necessary to sustain change" (Pawson & Tilley, 1997, p. 154). Manzano (2016) advocates beginning interviews with general questions about interviewee role/experiences/views about an initiative and then asking for their stories about specific experiences or issues.

Sampling and participants

This study used purposeful sampling (Patton, 2015) to select participating schools. This sampling strategy is different than random sampling in quantitative methods and entails selecting information-rich cases that help illuminate the issues related to the purpose of the study. Studying fewer information-rich cases allows insights and in-depth understanding. One such strategy is selecting cases that are unusual or special in some way. In the literature, school self-evaluation in New Zealand has been identified as being of variable quality, so the purpose of this study is to learn from those schools that are engaged with higher-quality processes. Since the purpose of the study was to explore *why* and *how* self-evaluation works rather than *whether* it works, the purpose of the sampling was to select schools that had been identified as being effective at self-review.

Five high schools were chosen based on their previous identification as successful at self-evaluation, similar socioeconomic level, and student achievement levels that were above comparable schools. To identify schools, the ERO reports of all 37 mid-decile schools in two geographic areas of the North Island were read along with other school documents, school websites, and independent publications with quantitative data on school achievement levels. Within this fairly-prescribed initial sample (mid-decile, identified as effective), diversity was sought. Realist sampling should have variability in order to test theories about the processes in different contexts (Manzano, 2016). Therefore, diversity was sought among the schools in

terms of size, location, single-sex/co-educational, state/integrated, and length of tenure of the principal.

The four schools in Phase Two were each given a pseudonym based on a New Zealand native tree, and are described briefly below:

- Totara School is a very large, co-ed state school with a very diverse student body and an established principal. It has high University Entrance results for both boys and girls for its decile level.

- Kauri School is a large, co-ed, integrated (Catholic) school, with a less-diverse student population that is approximately two-thirds European. It has high NCEA Level 2 achievement rates, a fairly new principal, and new deputy principal.

- Matai School is a very large, co-ed state school with a diverse study body, a sizeable Māori population, and a fairly new principal. Māori and Pacific student achievement is above the national average.

- Rata School is a medium-sized, single-sex state school with a large Māori population and an established principal. Māori student achievement at Rata is significantly higher for NCEA at all levels compared to Māori students nationally.

Interview logistics

In Phase Two, a total of 13 educators at the four high schools were interviewed. Within each school, while only a small number of people were interviewed, a diversity of roles was sought because of their awareness of different aspects of the schools' processes. Within each school, three people were usually interviewed: a principal or deputy principal, and two staff who were not part of the school's senior leadership team. Many were heads of department, and all were knowledgeable about their school's self-evaluation processes. Interviews lasted approximately one hour, ranging from 58 minutes to 79 minutes. Documents from each school were also analysed to triangulate the data to gain a more complete understanding of the schools' processes. These included ERO reports, background information on the schools and tools used by the schools to build capacity, such as templates and handbooks. The tools

created by the schools often illuminated their underlying theories about what constitutes effective self-evaluation and how to effectively carry it out.

Some of the interview questions included: Describe how your school carries out ongoing self-review. What are some of the: structures, organisational routines, processes, tools and resources, specific data sources that are part of that process? How have teachers and school leaders in this school acquired the capacity to be able to carry out these processes? Can you give me an example of when your self-evaluation made a difference for targeted groups of students? A full list of interview questions can be found in Appendix D.

Intermediate outcomes questions

Most of the questions are fairly standard behavioural interview questions asking participants to give specific examples of how the processes happen in their school. However, one question warrants more explanation since it is a less-common aspect of qualitative interviews and is perhaps the most realist-informed aspect. Pawson and Tilley (1997) argue that in realist interviews the researcher explains their theory, or potential explanations for a specific set of actions, to the participant. Then interviewees are asked to confirm, clarify and improve the theory. This allows for a more nuanced, focused explanation of how processes operate in various settings. Pawson and Tilley (1997) suggest the possibility of giving a very short survey asking participants to rate to what extent the statement applies to them. The statements articulate *how* a programme operates in the interviewee's setting (example statement: "this course helps students to experience success"). I used the results of Phase One to develop a list of 10 intermediate outcomes of school self-evaluation leading to school improvement. Evaluation activities can be linked with intermediate outcomes that are precursors of the specific long-range social goals that motivate evaluation. The following are the instructions I used for this question:

> "The subtitle of this ERO document is *how to do and use internal evaluation for improvement*. So clearly they theorise that internal evaluation leads to school improvement and I tend to agree with them. So I'm interested in *how* that happens and what happens in between self-evaluation and improvement. There are obviously multiple ways that a school could go about that. In talking to other schools and looking at the literature I've come up with a list of about 10 different ways that a school could go about this. You might call them intermediate outcomes. I would like you to choose three of these that you think your school focuses on and does well. If

you could prioritise three of those and then I'll just ask you to walk me through how each of those plays out in your school."

The intermediate outcomes:

- Changed teacher pedagogy

- Increased pastoral care

- Changes to grouping of students

- Greater tracking and monitoring of student achievement

- Changes to class offerings to respond to student needs

- More mentoring and support of students

- Differentiated responses for students at risk of not achieving

- Increased information provided to students and whānau about their progress

- Increased use of data for planning, teaching and decision making

- Increased teamwork and collaboration across the school

Participants chose three items and talked about *how* the processes such as tracking and monitoring work in their school. Walking me through the process encouraged participants to provide concrete examples and stories related to the process of how evaluation processes worked in their school. One of these processes, *greater tracking and monitoring of student achievement* became the focus of Article Two.

Phase Three: Theory Synthesis

Phase three corresponds to my processes for data analysis. Data analysis in realist research and evaluation is not a specific method but a way of using theory to understand patterns in data. In a colourful description, Emmel (2013) compared realist qualitative analysis to the process of splitting open a chicken to study its inner parts:

Splitting a chicken down its breastbone and opening it up to reveal the details of its

thoracic and abdominal cavities. In a similar way, in research we will split these

things… open and lay bare their anatomy for scrutiny and explanation through

theorisation and empirical investigation. In the process of which we will be able to

better describe, interpret, and, ultimately, explain the sample. (p. 100)

In each of the articles in this study, the data analysis was slightly different and is described in each article, but they also had very similar processes. The page limit for journal articles does not allow for the detailed description of analysis that is conveyed here. In each paper the analysis eventually involved a back-and-forth conversation between the empirical data and a range of social and behavioural theories (Astbury & Leeuw, 2010). Only in one article, Article Three on evaluation capacity building, was the overall topic even determined a priori; in all of the other articles, the topics were driven by the data and I developed them inductively. Below I offer one detailed example of the iterative nature of the analysis to illustrate the process; this describes the process for the ECB article, Article Three.

Data analysis began with transcribing the audio recordings of the interviews and reading them to become familiar with the data. For this study, segments of interview data relating to ECB were pulled out from the wider interview data. This was merely the initial topic of interest, a "bin" for sorting data for further analysis, and not yet any interpretation of what was said about the topic. Maxwell (2013) calls this an organisational category. The segments of data were primarily participants' responses to the interview questions: "How have teachers and school leaders in this school acquired the capacity to be able to carry out these (school self-evaluation) processes?" "What resources have you drawn on to build capacity?" Any other instances in the interviews where participants discussed the concept of capacity building were also sought. I read and re-read these segments.

I initially decided to analyse each school as a separate case study and wrote a profile or vignette of each school's capacity building. In this, I focused on how the schools differed. After further data analysis and in-depth discussions with my supervisors, I noticed that there were meaningful recurring similarities among the schools' processes. I abandoned the case-based focus and adopted a focus on the similarity of their processes, using a thematic analysis (Miles, Huberman, & Saldaña, 2014). Thematic analysis involves identifying, coding, categorising, and labelling the patterns in the data to determine what's significant (Patton, 2015). Re-reading the data, I did an inductive analysis, noticing certain themes or patterns in the data. I identified themes such as resources and group interaction. I adopted a

process of open coding of the data (Patton, 2015) which involved reading the data and developing categories based on what data seemed meaningful, interesting, or important rather than codes identified ahead of time. Coding categories are a way to sort the descriptive data so that the material on a given topic can be separated from other data (Patton, 2015). I then re-read the data, assigned codes to segments of data, and grouped together similar data into themes.

After several rounds of data analysis, I developed a hunch that educators were describing using processes consistent with sociocultural learning theory and that this framework could help to understand how schools were building evaluative capacity. I turned to the literature to read more about the sociocultural learning theory. When I noticed the pattern of the prevalence of these sociocultural practices, I developed specific codes for these. Drawing on a realist approach to data qualitative data analysis, Maxwell (2012) calls these *theoretical categories* since they are derived from prior theory. I searched for additional examples among the interview data of codes such as "tools". Thus I did a deductive analysis using the new categories to test whether the data did in fact support the hunch about the sociocultural learning theory. The analysis confirmed that this theory was an appropriate framework for understanding the schools' capacity building activities. This process is called *abduction* to distinguish it from both purely data-based inductive analysis and theory-driven deductive analysis. Abduction combines the deductive and inductive models of proposition development and theory construction (Emmel et al., 2018). Realist philosophers and evaluators sometimes refer to this constant shuttling between theory and empirical data, using both inductive and deductive reasoning, as ''retroduction'' (Astbury & Leeuw, 2010; Bhaskar, 1978).

I have offered a detailed description of my data collection processes. This is intended to create transparency, demonstrate rigour, and establish confirmability (Miles et al., 2014). Another reason is to contribute to the field methodologically, since researchers and evaluators advocate the use of qualitative realist methods, but there have been few examples showing exactly how realist evaluation methods work in practice (Astbury & Leeuw, 2010; Manzano, 2016).

Chapter Six

Introduction to Article One

The topic of Article One, *inquiry* was not an initial focus of my study, but I identified the topic inductively, driven by participants' data. Participants were informed of the study's focus on department-level and school-level self-evaluation but, in talking about *how* school self-evaluation occurs in each of their settings, participants highlighted the importance of the theme of inquiry. This article demonstrates that the model of an inquiring teacher informs educators' views of their work and their professional identity. It also argues that theories about inquiry have heavily informed the theories of school-level self-evaluation in New Zealand. While inquiry processes are often focused on the individual teacher, educators explain that inquiry concepts have informed their department and school-level self-evaluation.

Throughout this thesis, I examine the literature in such a way as to ask: what theories and uninterrogated underlying assumptions are embedded in this work? In this article I problematise the ubiquitous (in New Zealand) term *inquiry* and dissect the competing theories of the concept. The term *inquiry* is used in contrasting ways in the literature which can be problematic. In scrutinising the theories of inquiry in the literature, I demonstrate that it often conveys conflicting advice, competing demands, and contradictory purposes. I discuss two versions of inquiry that have different traditions, roots, emphases, and purposes. These contrasting influences can make it difficult for educators to navigate these processes.

Some of those conflicting demands are reflected in the wider struggle between the two competing agendas of internal and external accountability that were introduced in Chapter Two. These agendas of internal and external accountability influence how data are perceived and used as part of decision-making (Lai & Schildkamp, 2016). This article demonstrates that internal and external accountability discourses are *both* embedded in the concept of inquiry and that teachers have to negotiate between these competing discourses (Mausethagen, 2013). The concept of data use in New Zealand has been framed as inquiry which teachers view as having an internal accountability orientation. In this framing, inquiry is a life-long disposition for rigorous and critical reflection about one's own practice. But, as enacted, it can be more problematic and less straightforward than envisioned. It is often not an internal accountability that holds teachers accountable for exemplifying an inquiry stance, which may have been its original intent (Sinnema & Aitken, 2011). Rather, external

accountability can seep in in several ways that are discussed in the article. This article explores how teachers experience tensions, especially when what is called inquiry slips from internal to external accountability.

I decided to publish this article in an international journal because I believed the tensions experienced by these educators reflect issues that teachers in many countries face in data-rich systems.

Article One

Walking on a Tightrope: Agency and Accountability in Practitioner Inquiry in New Zealand Secondary Schools

Introduction

In many different countries teachers are increasingly recognised as key to educational change. This recognition has been accompanied by an increased expectation that teachers will be reflective practitioners of their own practice and participants in collecting, analysing and interpreting data on department and school-level functions. Both of these processes, at the individual practitioner level and professional learning community level, have been labelled *inquiry*. Inquiry involves educators questioning, investigating, reflecting, sharing, and using data and evidence about their own practice. It views teachers as active, agentic decision-makers in their own contexts (Sinnema, Meyer, & Aitken, 2017). Inquiry approaches sit in contrast to models that treat teachers as passive participants who implement solutions generated by external "experts". As described in the literature, inquiry centres on the promise of teacher agency to pose their own questions and solve problems in their own context. Because of the increased importance of inquiry and teacher data-use processes, the purpose of this study was to explore how inquiry is conceived in the literature and how it was enacted and experienced by New Zealand high school teachers.

An examination of the literature on inquiry revealed that it often conveys conflicting advice, competing demands, and contradictory purposes. The tensions created from those contrasting influences can make it difficult for educators to navigate these processes. An examination of the literature in this field reveals three tensions in the way inquiry is currently conceived: 1. the term is so pervasive that it means all things to all people and the actual differences are rarely articulated; 2. enacting inquiry as illustrated in the literature requires walking a narrow tightrope of practices; and 3. the promise of agentic inquiry can be precarious when enacted in a context of neoliberal accountability. What perhaps has not been discussed in the literature is the way these tensions actually affect teachers and school leaders when they enact inquiry. This article will discuss the way these tensions are conceived in the literature and then explore the ways in which high school educators in New Zealand attempt to balance these tensions when they walk the tightrope that comprises enacted inquiry processes. Although their context may be different, many readers concerned with

practitioner inquiry, teacher data use, and school self-evaluation may have an interest in the examination of the way inquiry was experienced by these teachers.

Inquiry in the New Zealand Context

Before delineating the three current issues in the literature, it is valuable to consider why New Zealand is a constructive context in which to study inquiry. The emphasis on inquiry is more prominent and central in New Zealand education policy documents than in other countries (Sinnema & Aitken, 2011). The *NZC* presents Teaching as Inquiry as an essential aspect of effective pedagogy (MOE, 2007). In the NZC, Teaching as Inquiry is described as a cyclical process in which educators pose questions, gather evidence, and make decisions about their practice. According to Sinnema and Aitken (2011), the inclusion of this model of pedagogy in a national curriculum statement is unique to New Zealand. Other New Zealand policy documents that include concepts of inquiry include the new Standards for the Teaching Profession that specify the expectation for teachers to "inquire into and reflect on the effectiveness of practice in an ongoing way, using evidence from a range of sources" (Education Council of Aotearoa New Zealand, 2017, p. 18). A recent MOE good practice document for schools links the concepts of school self-evaluation and inquiry, and uses the word *inquiry* 43 times (ERO & MOE, 2015). Inquiry frameworks are also included in teacher preparation courses in New Zealand and are a component of the way students are assessed (Hill, 2016). While the model of inquiry therefore seems like a mantra in the New Zealand education sector, it is actually quite a slippery construct to nail down. The term inquiry is ubiquitous in New Zealand, but the concept is not universally understood nor consistently practiced (Benade, 2012; ERO, 2011; Sinnema & Aitken, 2011). In fact, practitioners often confuse the concept of teaching as inquiry with students' 'inquiry learning', an approach in which students learn by investigating and exploring topics of interest (Benade, 2015; ERO, 2011; Sinnema & Aitken, 2011).

New Zealand is also an interesting context in which to examine this issue because the education sector as a whole comprises competing influences and theories. The New Zealand education sector, including its curriculum and assessment, has a history of being quite progressive and many aspects continue to be (Mutch, 2013). New Zealand school evaluation is an area that also contrasts to school inspections in some other countries, as the ERO has made school self-evaluation the centrepiece of school evaluation, emphasising good practice and school improvement more than compliance (Mutch, 2012). Elsewhere I have written

about some of the positive responses of high school teachers to New Zealand schools' self-evaluation processes (Dyson, 2018). However, some aspects of New Zealand's progressive tradition have been under pressure in recent years and the climate in education has shifted to include a range of managerial-oriented government policies and elements of performativity (Maguire, 2018; Smardon & Charteris, 2012; Thrupp, 2017). While aspects of the New Zealand system have a progressive history, schools have been administered in a very neoliberal manner for more than 30 years, including a devolved education system of individual self-managing schools (Thrupp, 2018). A system that includes school choice has led to a market version of accountability, which can lead to schools acting like businesses competing for parent clients in an education market (Court, 2004). Because the New Zealand education system is shaped by these differing influences and theories, there are often conflicts in values and philosophical perspectives (Maguire, 2018). While some of the particulars of the context will be unique to New Zealand, the pressures of accountability are felt in most education sectors around the world.

Tensions in the Literature

Now that the New Zealand context has been briefly explained, the paper returns to an examination of the literature on inquiry and the way it reveals three tensions: 1. issues arising from the prevalence of the term; 2. the challenge of enacting inquiry as illustrated in the literature; and 3. the challenge of enacting inquiry in a context of neoliberal accountability.

Tension one: Inquiry as all things to all people

The first problem with the way inquiry is currently conceived in the literature is that terminology for these practices is often not clearly delineated. Despite the increased use of the term "inquiry", there has been little explicit recognition that different authors mean different things and draw on different definitions. When they are using the generic term, inquiry, most authors do not define what this means to them or explain the theoretical, political, and epistemological traditions they are drawing on in using it. It is therefore often difficult to discern very much about a specific approach to teaching merely from an author's espousal of the use of inquiry. (A notable exception is a recent text by Sinnema and colleagues (2017) that defines what they mean by inquiry and delineates which authors and traditions they are drawing from.) More specific terms such as practitioner inquiry, practitioner research, teacher research, and action research are often used interchangeably in

the literature and among teachers in practice (Benade, 2015). Generally speaking, practitioner inquiry can be understood as an umbrella term for an approach that can be conducted using such specific methods as action research, teacher research, problem-based methodology, or teaching as inquiry (Benade, 2015; Cochran-Smith & Lytle, 2009), but in the discussion of my data, these terms may be used interchangeably to reflect the way they were represented by my participants.

The two main traditions in practitioner inquiry could be classified as 'stance' and 'school improvement' models. Some literature on inquiry emphasises the idea of "inquiry as stance" (Cochran-Smith & Lytle, 2009) with a social justice and equity orientation, and origins in grassroots teacher groups, critical research movements, and wider social movements. When inquiry is thought of as a stance, it is an active and habitual way of being that continues over the course of a professional career (Cochran-Smith & Lytle, 2009). Teachers learn throughout their careers by generating knowledge through posing important questions, learning how to unpack deeply held assumptions, and finding new ways to work with colleagues to transform students' learning opportunities and outcomes (Cochran-Smith & Lytle, 2009). Reid (2004) echoes that rather than teachers 'doing' inquiry as an activity separate from their professional being, it should be seen as a *disposition* and way of strengthening professional practice where they are "always striving to develop and expand their capacity to inquire" (Reid, 2004, p. 5). Reid (2004) emphasises the agentic position of teachers in the inquiry process:

> Inquiry is a process of systematic, rigorous and critical reflection about professional practice, and the contexts in which it occurs, in ways that question taken-for-granted assumptions. Its purpose is to inform decision-making for action. Inquiry can be undertaken individually, but it is most powerful when it is collaborative. It involves educators pursuing their 'wonderings' (Hubbard & Power, 1993), seeking answers to questions or puzzles that come from real-world observations and dilemmas. (p. 3)

This position emphasises teacher agency and the concept of teachers being responsible for leading their own learning.

Other literature focuses less on inquiry as stance and more on inquiry as a vehicle for school reform (e.g., Bryk, Gomez, Grunow, & LeMahieu, 2015). This body of literature is distinct from, and rarely overlaps with, the inquiry as stance literature; it draws on backgrounds and traditions in school improvement, school effectiveness, and organisational theory. In this

model, teachers usually work in professional learning communities to examine student and school data, with an inquiry process or inquiry cycle as the central driver of school reform. Collaborative inquiry involves groups of educators working together to use data to examine and improve their own practice or overall school functioning. These professional learning communities are often focused on increasing students' performance on assessments, decreasing student absenteeism or producing other desired school outcomes (Cochran-Smith & Lytle, 2009). With an emphasis on assessment data, a strong outcomes-focus, and a more comfortable fit with accountability pressures, some authors argue that these groups can tend to concentrate on "quick fixes and short-term goals" (Cochran-Smith & Lytle, 2009, p. 55). Despite the differences with inquiry as stance, this version of inquiry also emphasises agency at the teacher and school levels, in the sense of educators' ability to solve their own problems, and the importance of knowledge within the school to generate solutions (McNamara & O'Hara, 2008b). Many school improvement advocates would therefore see inquiry as a "process of developing collective professional agency within a school" (Timperley, Kaser, & Halbert, 2014, p. 5).

These different approaches at their core share a promise of inquiry centred on teacher agency and ability to problem-solve for their own learning and for improved student outcomes. However, there are tensions between these two versions of inquiry, with their different traditions, roots, emphases, and purposes. The existence of different versions of inquiry and the tensions between them are often not acknowledged in the literature.

Tension two: Walking a fine line to enact inquiry

Other conversations in the literature on inquiry focus on areas that educators must attend to when enacting inquiry. Enacting inquiry as illustrated in the literature requires walking a tightrope of a narrow range of practices, since even behaviours suggested as good practice have the potential to slip into less desirable behaviour. When using inquiry processes, schools must perform a delicate balancing act to negotiate the following:

- Focus on student outcomes but not too narrowly

- Use data but do not be too data-driven

- Give teachers agency but encourage coherence within the school

Each of these points of negotiation is discussed in brief below.

Focus on student outcomes but not too narrowly

An area where inquiry walks a fine line is a focus on student outcomes. Timperley, Parr, and Bertanees (2009) explain their view of the balancing act. Inquiry efforts should focus on improving valued educational outcomes for students and should be judged on these outcomes. They argue that, if student outcomes are not foregrounded in teachers' inquiries, this can lead to a system that enhances teachers' collegiality but has little impact on students. However, over-relying on student outcomes risks the danger of "gaming the system", for example, by narrowing the curriculum or using narrow assessment measures to ensure student assessment results appear to be favourable. "The challenge, therefore, is to develop teachers' knowledge and skills in ways that respect their professionalism, while ensuring that students benefit" (Timperley et al., 2009, p. 229). Charteris and Smith (2017) also contend that a focus on student outcomes requires walking a fine line and they criticise an inquiry cycle that focuses purely on measurable student outcomes. Benade (2015) argues that New Zealand's system of teaching as inquiry was well-intentioned but, in its present form, has a singular focus on the attainment of student learning outcomes that significantly narrows its potential to deliver social justice outcomes.

Use data but do not be too data-driven

Another area where educators need to carefully balance their inquiry processes is in their reliance on data. On the one hand, educators are increasingly encouraged to use data when making instructional and organisational decisions. Advocates argue that using data is central to the school improvement process (Daly, 2012; Lai & Schildkamp, 2013; Schaffer et al., 2012). A New Zealand synthesis of literature found that schools that gather data about their students and their practices and use these data to inquire into the effectiveness of their teaching and school practices made significant improvements in student achievement (Timperley et al., 2007). Data become part of a cycle of reflective inquiry where teachers and school leaders monitor the impact of their practices on student learning, and adjust these practices when the data show them to be less effective than desired (Timperley et al., 2007). This type of data use enables teachers and school leaders to tailor their teaching, school practices, and curriculum to address student-learning needs.

On the other hand are concerns that inquiry can become overly data-driven and focused on narrow student assessment data. Charteris and Smith (2017) argue that, "'data driven' decision-making aligns with a technicist shift in education that undervalues 'teacher-driven' or 'student-driven' interpretations of the world" (p. 601). They explain that, while they do

value data, they object to its position as a driver of all practice. Teacher inquiry as a purely data-driven construct can undermine teacher agency and decision-making (Lingard, 2013). Educators often need to balance two competing agendas of data use, one focused on external accountability and one focused on bottom-up, internal school self-evaluation. These agendas influence how data are perceived and used (Lai & Schildkamp, 2016). Schwandt (2005) argues that practitioner inquiry is walking a fine line between teacher practical wisdom and data-driven scientific knowledge, with the latter often pushing out the former. One of the thorny issues is the definition of *data* since, internationally, there has been a tendency toward a narrow definition of what count as data in education, with most studies focusing on standardised test scores (Earl & Seashore Louis, 2013; Lai & Schildkamp, 2013). Broader definitions of data include any information that is collected and organised from relevant stakeholders (e.g., parents, teachers, and students) to represent some aspect of schools (Lai & Schildkamp, 2013).

Give teachers agency but encourage coherence within the school

An area of disagreement within conversations on inquiry is how much freedom to give teachers in choosing a focus for their inquiry. The debate often comes down to the competing demands of teacher choice on the one hand and school coherence on the other. Timperley and colleagues (2014) explain their ideal of inquiry wherein they encourage a collaborative inquiry process and coherence within a school but point out that leaders should not decide the focus of teachers' inquiries. Cochran-Smith and Lytle (2009) explain that, ideally in practitioner research, community is a key feature but "the individual does not disappear into the collective" (p. 55).

Some authors emphasise more teacher choice and some emphasise less. Hill and Colby (2016) advocate for greater teacher choice and agency in the inquiry process when they describe a New Zealand project in which teachers who were working alongside a university researcher to do an in-depth inquiry were able to become more critically reflective and modify their practice. The authors emphasise the importance of teachers' ability to select their own inquiry topic based on building on their strengths, in contrast to many inquiry projects, which focus on identifying teacher deficits. Arguing for more school coherence and a focus on areas for improvement, Bendikson (2018) claims that the cycle of inquiry is not about individual teacher inquiries and that teachers should not generate "a plethora of ad hoc topics" (n.p.). Instead, she believes that the cycle of inquiry "is a problem identification and problem-solving cycle and it is recorded in the annual improvement plan under a target"

(n.p.). These two viewpoints seem to illustrate the difference between the 'stance' and 'school improvement' models discussed in the previous section.

One practice that some authors find problematic is *requiring* teachers to engage in practitioner inquiry and holding them accountable for completing an inquiry. Smardon and Charteris (2012) note the tension in mandating inquiry and argue that agentic inquiry must be teacher initiated and owned. In an English context, Evans (2011) labels this requirement as "demanded professionalism" and argues this maintains a focus on teachers' behaviour rather than attitudes, knowledge, or understanding. University lecturers lament that many New Zealand teachers regard inquiry as mainly an activity they are required to do for teacher appraisal (Bendikson, 2018; Hill, 2016). "Inquiry is becoming a mandated activity within a performance frame" (Hill, 2016, p. 7). Their inquiries tend to be small pedagogical intervention projects specified by the teachers' school—"they have to 'do' two or three inquiries a year" (Bendikson, 2018, n.p.)—rather than viewing inquiry as a disposition. Cardno, Bassett, and Wood (2017) describe a situation where a school required practitioner inquiry and chose the focus. A teacher explained that the school's senior leaders pronounced, "'your inquiry is going to be about this; this is what we want the students to get out of it', rather than us having the professional judgement" (p. 18). Bendikson (2018) argues that "inquiry has become yet another compliance activity (i.e., a 'project' to be completed), and quite possibly another time-waster for teachers, as opposed to a genuine and automatic part of the teaching cycle that involves checking one's own effectiveness" (n.p.) This is in contrast to the concept of inquiry as a habit of mind and continuous attitude for teaching practice, not a series of steps for solving classroom problems.

Tension three: On a tightrope without a net

The previous sections have delineated some of the ways that enacting inquiry can require careful navigation to stay on a narrow tightrope of practice, but perhaps the biggest threat to authentic practitioner inquiry is external accountability. Some authors argue that this is particularly true in a neoliberal accountability context characterised by competition between schools, managerialism and performativity (Ball, 2003). Ranson (2003) argues that recent educational accountability has been more about regulation and performance than educational improvement, local capacity building, and the encouragement of democracy in schools. What has occurred is "performance accountability for marketability" (Ranson, 2003, p. 467).

Some authors worry that inquiry can slip into performativity and be a way to hold educators accountable for their own and students' performance. Smardon and Charteris (2017) argue that in the New Zealand MOE model, inquiry is linked with teacher performance and school leader accountability, making it merely "accountability in the guise of agency" (p. 180). Peters (2016) argues that practitioner reflection has become "bureaucratized" and "managerialised" by neoliberalism, which has "appropriated" and "corrupted" the concept (p. 5).

An emphasis on performativity can also result in teachers engaging in self-monitoring (Ball, 2003). Brady (2016) warns that the related concept of teacher self-evaluation may be espoused to mean that teachers evaluate themselves, for themselves, using their own criteria and beliefs about teaching, but in practice self-evaluation is often corrupted through neoliberal visions of education, using accountability procedures like monitoring and self-surveillance. This makes "what is supposed to be a spontaneous and self-generated process into a bureaucratic, externally produced, standard-driven procedure" (p. 527).

In short, the literature indicates that external accountability has the potential to jeopardise authentic practitioner inquiry, with a danger for inquiry to become *corrupted, managerialised, appropriated, contrived.* Smardon and Charteris (2017) warn that we may miss the nuances of this shift since "it seeps under the door, with change happening so gradually that practices become normalised" (p. 181). While some scholars warn of the potential for external accountability to impact authentic teacher inquiry, much less literature explores how teachers actually live with these tensions.

This brief examination of the literature on inquiry has highlighted three tensions. What perhaps has not been discussed in the literature is the way these tensions actually affect teachers and school leaders when they enact inquiry. When these tensions are manifested in teachers' practice, the impact on educators' work is often on the affective domain. This was revealed during a larger study on school self-evaluation, details of which are described in the methods section. In this study it became apparent that there was an emotional dimension to educators' practice of inquiry, and this study attempted to unpack the embodied tensions from inquiry and what it means to do inquiry as a teacher.

Methods

This study was part of larger research study exploring the way that school self-evaluation leads to improvement in New Zealand secondary schools. That study sought and received ethics approval from University of Auckland. A purposive sample of five high schools was chosen based on their previous identification as successful at self-evaluation, similar socioeconomic level, and student achievement levels that were above average among comparable schools. To identify schools, the ERO reports of all 37 schools in the middle socioeconomic range of two geographic areas of the North Island were read along with other school documents, school websites and independent publications on school achievement levels.

The study draws on a realist theoretical framework (Maxwell, 2012; Pawson & Tilley, 1997) that attends to the way behaviour is shaped by aspects of the context, including the broader policy and institutional environment. Because of a focus on participants' unobservable thought processes, individual interviews were emphasised. Data collection began at one high school where six heads of department were each interviewed three times in open-ended focus groups. The results of these were used to inform questions for one-hour-long individual "realist interviews" (Manzano, 2016; Pawson, 2013) with 13 educators at four high schools. In realist interviews, questions often focus on the "reasoning and resources" of participants and "the social and cultural conditions necessary to sustain change" (Pawson & Tilley, 1997, p. 154). Within each school, three people were usually interviewed: a principal or assistant principal, and two staff who were not part of the school's senior leadership team. Many were heads of department, and all were selected because they were knowledgeable about their school's self-evaluation processes. Documents from each school were also analysed to gain a more complete understanding of the schools' processes. These included ERO reports, school websites, internal strategy documents, and background information on the schools. The four schools were each given a pseudonym based on a New Zealand native tree.

Interviews were recorded, transcribed and imported into the software package, NVivo 11. Before coding began, interview recordings were listened to several times, and interviews were read and re-read as a whole to try to see the big picture and grasp basic themes in the data. The topic of inquiry was identified inductively and driven by the data. Inquiry was not an initial focus of the study and participants were informed of the study's focus on

department-level and school-level self-evaluation but, in talking about *how* school self-evaluation occurs in each of their settings, participants highlighted the importance of the theme of inquiry. Inquiry appeared to be an important theme at all schools involved and across all types of participants (e.g., senior leaders, heads of department and teachers). Therefore, all segments of text related to the topic of inquiry were then grouped together. Then similar codes were assembled together to create a Pattern Code in order to organise the data and try to search for explanations in the data (Saldaña, 2016). This understanding was checked by developing, refining, and clarifying codes in relation to these data. Further analysis revealed that participants were often using emotive language and discussing feeling conflicted when talking about enacting inquiry. The tensions in the participants' comments encouraged a re-examination of the literature on inquiry to identify places where tensions existed there as well. Extensive debriefing was utilised in this study; two colleagues acted as a "reality check" of the data analysis and interpretations, a separate group of colleagues acted as debriefers on an early draft of the article, and a content expert who researches and writes extensively about inquiry commented on a more complete draft. Bazeley (2009) argues that these conversations usually create "added awareness of dimensions in the data and prompt fresh ideas, with new questions to pursue" (p. 7).

Results and Discussion

The analysis of the data identified three themes: inquiry is a central, yet contested, concept in these schools; teachers feel a great deal of tension when enacting inquiry; and their inquiry processes occasionally slip into accountability.

[Note: in places I have added italics to participants' quotes for emphasis. This is my emphasis.]

Inquiry: A central but contested concept in schools

In New Zealand, school-level self-evaluation appears to be heavily informed by the concept of inquiry. The model of an inquiring teacher informs educators' views of their work and their professional identity. Educators explained that inquiry concepts had informed their department and school-level self-evaluation. A deputy principal explained:

> I think the reason why we're good at self-review here is because it's a spin off of our professional learning programme in terms of teaching as inquiry. From the curriculum something that the school has done really well for the last five to seven

years is to really unpack the teaching as inquiry model. That then spins off into our other stuff. (Totara, deputy principal)

While inquiry is an important concept to them, educators in some of the schools had differing views on inquiry. Participants in two of the schools were quite consistent with colleagues in their views on inquiry, but in the other two schools, educators held differing or contentious views. One head of department commented that some of her colleagues thought inquiry could be informal processes such as teachers reflecting on their practice while driving home or short conversations with colleagues in the staffroom, but she disagreed and believed inquiry needed to be more systematic than this. Another teacher in the same school explained that not only did staff have differing views about inquiry but mixed messages coming from leadership caused teachers to feel tension and emotional discord:

> There was a lot of discontent in the staff. The term 'teaching as inquiry' is chucked around a lot. Some of the staff felt that the messages that were coming through were at odds with each other. One thing was said, and then another, and it was just a mess. (Kauri, head of department)

This teacher has illustrated the tension that occurs when the term inquiry is so pervasive that it can mean all things to all people. Since the literature on inquiry conveys contrasting influences and contradictory purposes, but these differences are rarely articulated in the literature, this can cause tension and confusion for educators enacting it.

Another, more-blunt example of differing views about inquiry comes from a principal who was fairly new to a school:

> There's a great history in this school of doing action research. I don't think it's really action research, but they call it action research. What hasn't been done is linking to the objectives of the school. They've just thought, 'oh we'd like inquiry about this'. Some of it's quite good, some of it is rubbish. (Matai, principal)

One of the issues beneath the brusque tone of this quote is the conversation in the literature on the competing demands of teacher choice and school coherence in determining the focus of teachers' inquiry. Perhaps this school has a history of encouraging teachers to choose their own focus for their inquiry, but the new principal supports more connections to the school's wider goals. Balancing those competing demands can cause friction.

Teachers feeling the tensions on the tightrope

Educators often expressed tensions surrounding inquiry as enacted in their schools. The inquiry processes involved competing demands and values. Feeling this tension could induce mixed emotions for these educators. This was particularly true for teachers in positions as heads of department (HODs) in their role overseeing other teachers and their inquiries. They were walking a tightrope between supporting the professional growth of teachers in their department on the one hand and intently focusing on the student performance expected by the school on the other. One HOD described balancing collegial relationships within the department and the expectations of school leadership to raise student achievement:

> As a HOD I can remember feeling a little bit *sort of torn* about it because in terms of the *relationships* that you have with your team and identifying these *deficiencies*, the areas where you are not getting *the results that you want* and then thinking about what does that mean for practice within the team. So that's been a learning curve. That's been really interesting around how you have honest conversations with people and be open to change and open to improvement. That was very *relational*. So that was an interesting part of self-evaluation which quite invested in the *responsibility* side of it as a very small [department] and that was very much trying to *raise achievement*. (Totara, head of department)

This head of department expresses feelings of "being torn" between several competing demands. The first is being relational with his team vs. focusing on teacher performance around student achievement. Ball (2003) argues that there is a tension between metric performances and authentic relationships. This HOD also expresses the tension between directing teachers' inquiry to a particular student achievement concern and leaving them free to choose their focus. He also feels the tension between fostering the sense of trust necessary to being open to change but being pressured to focus on identifying teacher deficiencies that are believed to lead to less desirable student achievement outcomes. The tensions and feelings of being torn illustrate Ball's (2003) argument that for teachers in a neoliberal educational culture, "the struggles are often internalized" (p. 216).

A HOD at a different school felt conflicted as she was describing how teachers in her department collected students' assessment results and feedback from students, which the HOD could monitor online:

> It is kind of like, *it sounds terrible in my head now*, but keeping an eye on them

without letting them know, *this sounds terrible*. Yeah keeping an eye on them so I know where my department is at. So if I am asked a question by [the principal], 'what's going on?', I can [respond]. (Rata, head of department)

The HOD has tacitly accepted what is asked of her, but when she explains it out loud, it sounds terrible to her. Once she describes the process, she realises it sounds like she is engaging in surveillance. The pressure to engage in these processes is not unique to this context. Ozga (2009) comments that data-rich systems can produce a "drive to constant surveillance" (p. 156). This teacher has not acknowledged that she is engaging in surveillance until she is saying it out loud; realising what she's doing appears to be at odds with her values. The danger that inquiry can slip into performativity and become a way to hold educators accountable for their performance was alluded to in the third tension found in the literature around inquiry. As Robertson (2008) cautions, "neoliberalism has transformed, in both predictable and unpredictable ways, *how* we think and *what we do* as teachers and learners, and it is therefore important we make these things evident to ourselves" (p. 12, emphasis in original).

Members of senior leadership teams and HODs explained that they tried to help teachers improve the quality of their inquiries. Their processes needed to balance between scaffolding teachers' learning and holding teachers accountable. A deputy principal described that the school's processes were walking a tightrope.

> We have tried to be really careful around the language of that and really don't use that word accountable. So we are really thinking about this is like a responsibility of an effective teacher. This is what a professional inquiring teacher would look like and this is what we want to see everyone doing and, you know, it is a really good thing. But it could slip into the principal wants to know that you've done your inquiry, you've done each part of that step by these dates and you've got a result. (Totara, deputy principal)

The principal explains that they want it to be a "good thing" for a "professional inquiring teacher" but he can recognise that it is very close to mandating and enforcing completing "an inquiry" which sounds like a distinct project and not like a genuine, ongoing part of teaching practice. He acknowledges that it is a fine line and even with the best intentions, the possibility is there to slip into accountability.

The educators' experience of tensions involved a conflict of emotions at times. A HOD described a situation in which the department's results were low on an end-of-year national assessment and he was called into the principal's office to explain the situation. Over a couple of years, the department inquired into the issues with the particular unit of study, experimented with different strategies, and eventually found one that raised achievement. He described the emotional aspects of this process:

> So that was a very interesting process of me experimenting with things that weren't working and feeling *very fragile*. We were all feeling *pretty anxious* about it because at the end of the day the school is holding, well it is seeing you as responsible I guess for what is happening and it is *not a very nice feeling* when you are behind the national average. And so it was *good*. It was quite *painful* but it was good to have to inquire into that and eventually, you know, get there. (Totara, head of department)

This teacher felt a range of emotions in this process of inquiring into a high-stakes assessment. Ball (2003) explains that these processes "have an emotional status dimension, as well as the appearance of rationality and objectivity. Thus, responses to the flow of performance information can engender individual feelings of pride, guilt, shame and envy" (p. 221).

Slipping into accountability

Educators' comments indicated that they occasionally slipped from a place where they struggled with the tensions of inquiry and fell into areas of accountability. This particularly happened in two areas that were raised as potential pitfalls in the literature: an overemphasis on narrow student outcomes and mandating reflection.

Overemphasis on narrow student outcomes

Taken as a whole, the advice in the literature is to walk a fine line to focus on student outcomes, but not too narrowly. In this study, teachers articulated that there can be a strong pressure to have a narrow focus on student achievement outcomes as the goal of their inquiries. Many of the teachers commented about being encouraged to be reflective practitioners but their comments reveal they are mainly being reflective about student performance on summative assessments. In a context of accountability, their inquiries appeared to often overemphasise performance on student summative assessments rather than

focus more broadly on student learning or other student outcomes. In one of the few comments to overtly criticise the pressure to perform, a teacher claimed:

> A lot of self-reflection and evaluation as a teacher inherently comes down to the [summative student] results. So there's a lot of it which is actually directed by the credit hunting. You know, why are these kids not getting credits, how do we get them more credits, you know, are they doing too much assessment and those kinds of things. So when we are looking at evaluating our roles and what we are doing invariably it comes to [student] assessments because that seems to be basically what we're judged on. (Matai, teacher)

Teachers' comments show they feel a great deal of pressure to raise student results. This creates tensions in how they interact with the practice of inquiry. While inquiry is supposed to be about becoming a reflective practitioner, it can turn into being a reflective practitioner for the purpose of getting students over the line. This echoes the warnings in the literature on the danger of an overemphasis on narrow student outcomes (e.g., Charteris & Smith, 2017). Zeichner (1996) cautions us that it is important to consider just what it is we want teachers to reflect about.

Mandating reflection

In the conversations in the literature around the competing demands of teacher choice and school coherence, many authors are critical of the practice of *requiring* teachers to engage in practitioner inquiry and holding them accountable for completing "an inquiry". Some of the schools in this study had succumbed to that practice. A deputy principal described how the school's processes had evolved to ensure certain inquiry practices.

> It was happening informally and anecdotally across the curriculum but we wanted to formalize that… For me it *ensured that all teachers were being more reflective* of their own practice. Because for some time we—we're still trying to head towards more inquiry within the teaching practice, so teaching as inquiry. (Rata, deputy principal)

There is something inherently contradictory about mandating and ensuring reflection, as opposed to encouraging and facilitating it. Benade (2012) notes the tensions associated with requiring practitioner inquiry and argues that compelling teachers to engage in inquiry using teacher appraisal is a form of surveillance. He further argues that mandating inquiry ensures

that practitioner research becomes "domesticated" as a way to raise student achievement, showing that this is intertwined with the previously discussed practice of an overemphasis on student outcomes.

While the practice of requiring inquiry is sometimes criticised in the literature, this is what New Zealand policy documents, including the new Teaching Standards, expect. For example, an ERO (2011) good practice document explains that, in schools where inquiry processes were effective, "inquiry was generally an expected and monitored component of school practice" (p. 27). Schools are expected to monitor that teachers are engaging in inquiry. However, in practice there is a fine line between monitoring and surveilling inquiry, and a real potential to slip into accountability.

Acknowledge the tensions but also the promise of inquiry

After giving plenty of attention to the challenges of enacting inquiry, I choose to end on a slightly more positive note. In the following quote, the educator is not looking at inquiry unrealistically; he acknowledges the tensions, but also holds that the promise of inquiry is possible. A HOD acknowledged that some teachers may be doing inquiry half-heartedly, but that most can see the intent behind the school's efforts and the promise of inquiry processes for improvement:

> I am sure there is some degree of people showing seemingly that they're doing [inquiry], but actually they haven't necessarily completely bought into it….[But] I think I'm being honest, the culture of the school is so strong that I think people could see the benefits. They can see that actually this is about their practice and improving the teaching and learning. (Totara, head of department)

It takes a great deal of work on the part of schools and the system as a whole for inquiry processes to be about authentic improvement and not merely accountability. But this quote implies that a key to the promise of inquiry may be whether the intent behind it is genuine.

Conclusion

This paper has examined the current conversations in the literature and explored some of the current debates over the meaning of inquiry. The existence of different versions of inquiry, with different traditions, is rarely acknowledged. Practicing inquiry as advised by the literature requires walking a tightrope on a narrow range of practice, because even desirable

behaviours can slip into dangerous territory. The literature has identified potential ways that inquiry can fall into accountability, including: a narrow focus on student outcomes, becoming overly data-driven, processes that are merely compliance exercises, school leaders choosing the focus of inquiry for teachers, or a contrived, paint-by-numbers formulaic inquiry process. Perhaps the most insidious danger is the potential for inquiry to become "accountability in the guise of agency" (Smardon & Charteris, 2017, p. 180). By knowing what the dangers are and what to look for, educators might avoid falling into them.

Some of the pitfalls described in the literature were present in the current study and some were not. For example, a process that authors warned about in the literature (e.g., Bendikson, 2018; Cardno et al., 2017; Hill, 2016; Smardon & Charteris, 2012) that was less prevalent in these schools was that inquiry becomes a tick-the-box compliance exercise. These teachers did not tend to view inquiry as a distinct 'project' they needed to complete for appraisal but rather did view it as a way to check their effectiveness and improve their practice. But to what end? Ideally teachers would engage in inquiry processes because they focus on the teacher and the learner, but the focus has often shifted to a narrow definition of student success based on summative assessments and pressure to make the teacher and school look good.

Of all the dangers that inquiry can succumb to that are enumerated in the literature, the discourse on accountability seems to have become the dominant one. Unresolved tensions, contested views of inquiry and "the apparent similarity of language used to indicate two markedly different philosophical bases" (Couch, 2018, p. 130) has pushed the educators in this study to walk a fine line and occasionally slip from an agentic to a performative use of inquiry processes. As enacted, this is often not an accountability that holds teachers accountable for embodying an inquiry stance, which may have been its original intent (Sinnema & Aitken, 2011). Rather, it is often inquiry in service of raising student summative achievement.

Educators in this study expressed the tensions they feel when accountability has snuck in, and the emotional aspects that can arise in the process. As demonstrated in the literature, even behaviours suggested as good practice have the potential to stumble into accountability. These teachers describe ways inquiry could slip or has already slipped, in many cases without them fully realising it. They struggle with such issues as: feeling torn between competing demands; realising some of their practices sound terrible when uttered aloud;

wrestling with tensions between the concept of the professional teacher and accountability; and an overemphasis on summative student achievement outcomes. These educators feel tension because of accountability pressures to reach certain student achievement outcomes, but they have not been explicitly pushing back on the pressure they feel. Despite their acquiescence, their words illustrate their mixed feelings. Teachers comply with the performative pressure but a deeper look at their comments shows that they often feel uneasy about it. Møller (2009) argues that educators are increasingly "caught in the cross-fire" of educational reforms focused on accountability (p. 42). He goes on to explain that it can be stressful and emotional when one's "personal accountability", that is, the values that are sacred to a person, are in conflict with other kinds of accountability, such as managerial accountability. Ball (2003) argues that performance surveillance deeply affects the attitudes, practices, and identities of educators. He goes on to suggest that performativity works from the outside in, through regulations, controls and pressures, but also from the inside out. It can provoke feelings of "guilt, uncertainty, instability and the emergence of a new subjectivity…A kind of values schizophrenia is experienced by individual teachers where commitment, judgement and authenticity within practice are sacrificed for impression and performance" (Ball, 2003, p. 221). The voices of the teachers in this study suggest that a "values schizophrenia" exists in the way that inquiry is currently enacted. Perhaps this is due to "structural and individual schizophrenia of values and purposes" (Ball, 2003, p. 223) of inquiry and the unacknowledged tensions inherent in different visions of inquiry.

Chapter Seven

Introduction to Article Two

The topic for Article Two was, like Article One, a subject that I had not identified a priori, but that I identified during data analysis. While Article One on inquiry focuses on the effects of data use processes on the individual teacher, Article Two widens the lens to focus on changes to teacher practice and organisational changes that came about after schools' implementation of a data use tool. In Chapter Five, I explained that during the realist interviews, I presented my participants with a list of intermediate outcomes of school self-evaluation processes. The topic for Article Two came from the analysis of the responses to that list. Participants identified the importance of one of the mechanisms included on the list, "increased tracking and monitoring of students". While not every participant chose this item on the list, participants at all schools explained that they used a "traffic light" tracking and monitoring tool, explained its importance, and how its data use processes worked. In analysing participants' descriptions of those processes at their school, I identified the positive outcomes of those processes, but also what I would classify as unintended consequences, such as an increased focus on summative student outcomes, particularly among certain groups of students. In this way, I explore a nuanced understanding of the tensions and mixed results of these data use processes.

One of the themes running through this thesis, the two competing discourses of accountability, is prominent in this article. This article highlights that the participating schools are leaders in data use and have data tools that allow them to track and monitor individual students' progress toward NCEA and identify which students might be at risk of not achieving. On the other hand, more so than the other articles, this one highlights external accountability. Article Two picks up on a theme that was mentioned in Article One, which is the way that schools focused on and responded to student achievement data in a context of neoliberal external accountability. I explore how neoliberal theories of competition interact with theories of school self-evaluation. In the previous article, I explain that the educators in this study need to walk a fine line in their focus on improving valued educational outcomes for students. In a context of accountability, teachers' inquiries appear to often overemphasise student outcomes on summative assessments rather than focus more broadly on student learning or other student outcomes. Teachers' comments suggest they operate in an environment that over-values the metrics of achievement and this context influences the

schools' processes. The danger is that over-relying on student outcomes risks "gaming the system", for example, using narrow assessment measures to ensure student assessment results appear to be favourable. Article Two explores this topic in depth. More so than the other articles, this article also highlights the way that school competition and the flexibility of the NCEA assessment system intersect with the accountability agendas.

I decided to publish this article in a New Zealand education journal. This article discusses practices known as "triage" that have been documented in overseas schools, but to my knowledge, not in New Zealand literature. I wanted to draw attention to these practices for a New Zealand audience and highlight that external accountability pressures can lead to unintended consequences of data use.

Article Two

Data Use in New Zealand Secondary Schools: Tracking, Traffic Lights, and Triage

Introduction

Data use has become an important part of educational reform in various countries around the world. Advocates argue that using data is central to the school improvement process (Daly, 2012; Lai & Schildkamp, 2013). Data use means different things in different contexts, but, broadly speaking, it refers to teachers systematically collecting and analysing data to guide and support educational decisions. Data are defined broadly as any information that is systematically collected and organised from relevant stakeholders (e.g., parents, teachers, and students) to represent some aspect of schools (Lai & Schildkamp, 2013). This could include student assessment and achievement data, attendance data, student demographic information, classroom observations, student work samples, or survey results from students, teachers, and parents. It could be argued that contemporary education cannot be fully understood without paying proper attention to the accumulation and flow of data in schools (Selwyn, 2014). Scholars call for educational research to develop nuanced approaches to understanding what it means to live and work within what Selwyn (2014) refers to as "data deluged conditions" in educational contexts.

Research on data use has become its own area of scholarship in the past 15 years or so. In the beginning, much of the literature was prescriptive rather than descriptive, advocating for data use or providing how-to guides rather than analysing what actually happens when educators use data in their practice. More recently, an increase in research on data use processes has attempted to understand how people engage with data in their everyday work and how that relates to instructional change and organisational learning (Coburn & Turner, 2011, 2012).

Coburn and Turner's (2011) framework provides conceptual guidance for the scholarship around data use interventions and advocates research focusing on: (a) the process of data use; (b) the context in which it occurs; (c) characteristics of data use interventions; and/or (d) potential outcomes. This framework also recognises that data use is an act of interpretation and sensemaking that involves noticing data in the first place, making meaning of it, and constructing implications for action (Bertrand & Marsh, 2015; Coburn & Turner, 2011; Datnow, Park, & Kennedy-Lewis, 2012; Datnow et al., 2017). The framework also emphasises that the process of data use is shaped in significant ways by the organisational

and political contexts in which it occurs (Coburn & Turner, 2011, 2012; Farrell & Marsh, 2016; Mausethagen et al., 2018). Finally, Coburn and Turner's (2011) framework highlights the importance of examining the outcomes of data use interventions and the pathways between intervention activities and outcomes. While student learnings are the most frequently studied outcome of data use interventions, the framework also suggests a broader range of outcomes that may be important when studying the phenomenon of data use. The least-studied aspects of data use interventions are organisational change and change in practice. These two outcomes are the subject of this paper.

Practices of data use are more established in the New Zealand context than in some other countries because of the well-developed culture of data-informed teacher inquiry and a system of school self-evaluation (Nusche et al., 2012; Sinnema & Aitken, 2011). In addition, data use has been shown to be effective in New Zealand in large-scale government-funded schooling improvement interventions that emphasised collaborative data use; outcomes included improved achievement for Māori, Pacific, and low-income students (e.g., Lai & McNaughton, 2016). Each New Zealand school is expected to perform ongoing evaluation for improvement that generally consists of noticing data, investigating further, collaboratively making sense of data, taking action, and then monitoring and evaluating the impact of the action taken (ERO, 2015). While there is an expectation of school self-evaluation, the literature shows that self-review is highly variable across New Zealand schools and only a minority of schools have strong processes in place (Mutch, 2012; Nusche et al., 2012; Timperley, 2013).

This research is part of a larger study on schools that are effective at self-evaluation and data use that explored how school self-evaluation leads to school improvement. Despite the New Zealand education system's reliance on school self-evaluation, little empirical research has shown how these processes work. Most research on data use in New Zealand schools has occurred within the context of the previously mentioned large-scale interventions; in contrast, this study examined everyday teacher and school leader practice. This study tried to open up the "black box" of data use in schools to explore the mechanisms of the change process while attending to the context in which school data use occurs in New Zealand. Using a case study of selected New Zealand high schools, this study explored the perceptions of teachers and school leaders whose schools had recently implemented a student assessment tracking and monitoring technology tool. These "traffic light" systems organised and displayed assessment results, allowing schools to track students' progress and

identify which students might be at risk of not achieving. This study explored the ways this data tool informed the strategies used for school improvement and how use of this assessment data contributed to organisational change and changes in practice. This paper will demonstrate that the data became a powerful force and led to structural changes within the schools. This study found many similarities but also nuanced differences among the schools' data use practices. After a brief overview of the international literature on data use, this introduction describes the relevant education context in New Zealand, with an emphasis on data use.

Selected Data Use Literature

It has been argued that the push for data use comes from two competing agendas—one for external accountability and one for internal accountability (Lai & Schildkamp, 2016). In an external accountability framework, teachers and schools are held externally accountable for student educational outcomes, primarily through summative assessment data. A culture of external accountability tends to be reactive and is motivated by the need to raise assessment scores as an end in itself. On the other hand, a culture of internal accountability emphasises teacher inquiry, school self-evaluation, and formative assessment. These different accountability agendas shape how assessment data are used. School leaders and teachers often have to navigate both agendas simultaneously, which can create tension for educators (Lai & Schildkamp, 2016). New Zealand's education system simultaneously embraces these two conflicting cultures of data use, which will be explored more fully later in this section.

The literature extols many benefits of data use in schools. Research suggests that data use is central to the school improvement process (Coburn & Turner, 2011; Daly, 2012; Lai & Schildkamp, 2013). A synthesis of the literature found that schools that gather data about their practices and their students and then use these data to inquire into the effectiveness of their teaching and school practices made significant improvements in student achievement (Timperley et al., 2007). Some policy makers even argue that the only way to increase student achievement is for schools to make decisions based on data (Bertrand & Marsh, 2015). Some scholars also highlight the equity goal of data use (Bertrand & Marsh, 2015; Datnow et al., 2017), contending that expanded access to data allows schools to operate more equitably (Selwyn, 2014); the goal of data use is success for all students (Datnow & Park, 2018); and data use can help accomplish the equity goal of supporting all students to reach their full potential (Datnow et al., 2017; ERO, 2015).

While the literature describes many benefits of data use in education, it also touches on some pitfalls, mostly centred on pressures of data use for external accountability (Booher-Jennings, 2005; Coburn & Turner, 2012; Farrell & Marsh, 2016; Mandinach & Schildkamp, 2020; Mausethagen et al., 2018). Much of the external pressure described in the data use literature occurs in contexts with government-driven rewards or sanctions in response to attaining targets on high-stakes testing such as No Child Left Behind policy in the US or reward funding based on NAPLAN performance in Australia. In other contexts, the external pressure comes from a type of accountability called "consumer accountability" with the purpose to strengthen consumer (parent) responsiveness (Ranson, 2003). Some authors, particularly those with a critical theory orientation, argue that in a culture of neoliberalism, data are used to monitor and judge schools, teachers, and students in order to reinforce market principles of accountability, consumer choice, market competition, and efficiency (Ball, 2003; Court, 2004; Selwyn, 2014). The literature on data use notes that the misuse and abuse of data are more likely in high-stakes accountability systems that put pressure on teachers to use data in certain ways, and demonstrates the negative outcomes of accountability policies on teachers' pedagogical practices and on students (Booher-Jennings, 2005; Farrell & Marsh, 2016; Lingard & Sellar, 2013; Mausethagen et al., 2018). One outcome of an increased emphasis on assessment data is a narrowing of the curriculum, "as the 'tail wags the dog', encouraging schools and teachers to teach to the test/target/inspection/performance management goal" (Thrupp, 2005, p. 42).

Another practice of data misuse or abuse occurs when schools focus their efforts on a particular small group of students, such as those on the cusp of proficiency, thus limiting educational opportunities for others (Booher-Jennings, 2005). Booher-Jennings describes a situation where teachers divided students into three groups—safe cases, students on the threshold, and hopeless cases. Under the auspices of data-driven decision making, these teachers then diverted resources to those students most likely to increase the school's scores and away from those viewed as hopeless cases rather than meeting the needs of individual students. Gillborn and Youdell (2000) call it "educational triage" when schools classify students based on whether or not the students have the potential to improve overall school scores and enhance the school's reputation (p. 14). *Triage* refers to the disproportionate investment of instructional resources precisely on students close to proficiency cut points potentially at the expense of other students and their educational needs. "Gillborn and Youdell's application of the idea of triage, a practice usually operating in dire circumstances such as the battlefield or the emergency room, to education perceptively captures the

dynamics … [in which] some students must be sacrificed in the name of improving the school's aggregate scores" (Booher-Jennings, 2005, p. 241). Some authors argue that, in a neoliberal or market-driven context, it is inevitable that organisations will spend money in areas where measurable successes can be demonstrated (Ball, 2003).

New Zealand and Data Use

Conflicting cultures of data use

The New Zealand education system has several features that make it a constructive context in which to study data use. While New Zealand faces many of the same educational issues experienced by other countries, including equity concerns for indigenous and other priority students, its education system also has some defining characteristics. One of the distinguishing features and strengths of the New Zealand education system is its emphasis on teaching and school leadership as a process of inquiry, as evidenced by *The New Zealand Curriculum* (MOE, 2007). Another distinguishing characteristic is its decentralised decision making and school self-governance in which responsibility for academic rigour and student achievement lies with individual self-managing schools. Another feature is that New Zealand's system of evaluation sits in contrast to school inspections in some other countries since it is centred on school self-evaluation. These elements require that New Zealand schools develop self-review systems and regularly use data to inform their teaching and improvement practices.

Other aspects of the New Zealand education system sit in contrast to these progressive practices, particularly the changing ethos of educational accountability. Court (2004) argues that New Zealand schools have a market version of accountability, with schools being like businesses competing for parent clients in an education market focused on school choice. In a system of competition created by policies of school choice, schools' reputations become all-important, as for example, when popular magazines such as *Metro* and *North and South* publish high schools' qualification results. A recent comprehensive report by a government taskforce contends that New Zealand's system of school competition was intended to improve access to quality schools through parent choice, but in practice there is no evidence that this competition or choice has improved overall school performance or student outcomes (Tomorrow's Schools Independent Taskforce, 2019). In 2005, Thrupp argued that a new emphasis on student achievement outcomes in target setting signalled a growth in managerial policy, but that it was still "soft touch" because schools set their own targets. However, during the time of data collection for the current study, the National Government (not ERO

or the MOE) had set a Better Public Service target that 85% of students would achieve Level 2 of the NCEA. This produced a great deal of pressure on schools and educators. In a recent opinion piece, a teacher wrote that come "hell or high water we needed to get to that level" (Education Central, 2018). Whether through government targets or popular media, the NCEA Level 2 achievement numbers are used as one of the main indicators of school quality, and represent external accountability of schools. This is one way in which the accountability climate in New Zealand education has shifted to include a range of managerial-oriented government policies (Thrupp, 2017).

New Zealand's education system simultaneously embraces these two conflicting cultures of data use, an educational system culture of internal accountability within a context of data use for external accountability (Lai & McNaughton, 2016; Lai & Schildkamp, 2016). It has a culture of teacher inquiry, practitioner agency, and school self-evaluation, but simultaneously had the pressure of consumer accountability, school choice, and a government-set 85% target for student achievement. In the data use literature, this tension between internal and external accountability is often not addressed, yet it is important for understanding what happens when educators engage with data in their everyday work (Mausethagen et al., 2018).

Tracking student achievement data

The literature shows that New Zealand school self-evaluation is highly variable across schools and only a minority of schools have strong processes in place (Mutch, 2012; Nusche et al., 2012; Timperley, 2013). One challenge is that, until very recently, most schools in New Zealand lacked sufficient long-term data to track student achievement and lacked access to appropriate computerised data management systems (Lai & Hsiao, 2014). Data management systems allow schools to organise information about student achievement, identify groups and individuals who need particular attention, and display information in a variety of formats. Many schools still struggle with effectively utilising these systems. The lack of effective tracking and monitoring may be particularly problematic in New Zealand because of the complexity of the NCEA system in which students incrementally accrue credits throughout the year, and students and parents may not be aware that a student is not on track to achieve a particular level or to gain the particular combination of credits required for University Entrance.

The Starpath Project, a university–government partnership initiative, found that this is also an equity issue since a key barrier to tertiary study for underrepresented groups (Māori, Pacific, and low-income students) was the lack of long-term data on student participation and achievement that would allow for tracking of students and effective goal-setting (Kiro et al., 2016; Starpath, n.d.). As a result, programmes like the Starpath Project emphasised real-time tracking and monitoring of students' NCEA results and early warning systems to identify students at risk of not achieving. However, an evaluation of the project indicated that even with support from Starpath, more than 40% of project schools struggled to maintain a database tool to collect longitudinal data on student achievement (Kiro et al., 2016). Furthermore, they found that the education system as a whole "lacks the capacity for adequate longitudinal tracking of student achievement at present. There is an urgent need to improve the Student Management Systems used by schools to ensure that an adequate and strategic range of data is consistently and accurately recorded over time" (Kiro et al., 2016, p. 5).

The importance of active tracking was reinforced by Daniell's (2018) analysis of secondary schools with NCEA Level 3 and University Entrance attainment rates significantly above the mean (high performing) and schools with attainment rates significantly below the mean (low performing). Schools reflected a range of socioeconomic levels, school size, student demographics, and geographic location. The majority of high-performing schools had systems for regular, active tracking of students' progress that enabled schools to identify students at risk of not achieving. This tracking process typically involved multiple stakeholders, including students, parents, and senior school leaders. In contrast, low-performing schools were less likely to identify at-risk students as they used "passive" tracking processes, such as providing students with their credit-tracking report but "where responsibility for any subsequent action was devolved to the students" (Daniell, 2018, p. 220). The majority of high-performing schools also put in place interventions to enable students to attain qualifications, such as the provision of short courses or credit top-up opportunities through external holiday programmes or in-school programmes. No low-performing schools had similar interventions.

All of the schools in the current study had effective tracking and monitoring databases in place. These "traffic light" systems allowed schools to track students' progress on New Zealand's complicated NCEA assessment system and identify which students might be at risk of not achieving. Schools used these databases to track students' real-time results on

internal assessments and then code them either red, amber, or green (RAG ratings). This study explored the effect of these traffic light tracking systems on the processes of selected schools.

Theoretical Framework and Methods

This study is informed by realist research and evaluation principles and methods (Emmel et al., 2018; Pawson & Tilley, 1997). Realists emphasise that outcomes of initiatives are shaped by the particular context in which they occur and by human agency, through the reasoning of the people involved. This way of thinking is oriented towards understanding how, why, for whom, and under what conditions interventions produce specific results. This was a suitable approach for this study because of the goal to "open the black box", or understand processes of *how* school self-evaluation leads to school improvement. This study attempted to explore the mechanisms or pathways between data use interventions and outcomes in these schools (Coburn & Turner, 2012). Realist research and evaluation usually involves the development and testing of theories about the various causal processes through which an initiative activates responses from participants (Manzano, 2016; Pawson & Tilley, 1997).

The first phase of this study, that Manzano (2016) calls "theory gleaning", involved reading New Zealand and international school self-evaluation literature and policy documents and then conducting a series of three focus groups including an open-ended qualitative activity with heads of department in one school. The results of these were used to inform questions for 1-hour-long individual "realist interviews" (Manzano, 2016; Pawson, 2013) with 13 educators at four high schools. The findings from the focus groups and literature were used to develop a list of 11 initial potential explanations or theories (Pawson & Tilley, 1997) about how school self-evaluation processes work in each school (for example: changed teacher pedagogy; greater tracking and monitoring of student achievement; changes to class offerings to respond to student needs). During the individual interviews, each participant was asked to choose three processes from the list that were a priority in their school and explain how each worked in their specific school, allowing a more nuanced, focused explanation of how processes operate in various settings (Manzano, 2016).

Participants

Because school self-evaluation is highly variable across New Zealand schools and only a minority of schools have strong processes in place, this research aimed to examine secondary

schools that are effective at self-evaluation. A purposive sample (Patton, 2001) of five high schools was chosen based on their previous identification as successful at self-evaluation, similar socioeconomic level, and student achievement levels that were above comparable schools. To identify schools, ERO reports of all 37 mid-decile schools in two geographic areas of the North Island were read along with other school documents, school websites, and independent publications on school achievement levels. Within these selection criteria, diversity was sought among the schools in terms of size, location, single-sex/co-educational, state/integrated, and length of tenure of the principal. Within each school, three people were usually interviewed: a principal or deputy principal, and two staff who were not part of the school's senior leadership team. Many were heads of department, and all were knowledgeable about their school's self-evaluation processes.

The four schools were each given a pseudonym based on a New Zealand native tree, and are described briefly below:

- Tōtara School is a very large, co-educational state school with a very diverse student body and an established principal. It has high University Entrance results for both boys and girls for its decile level.

- Kauri School is a large, co-educational, integrated (Catholic) school, with a less-diverse student population that is approximately two-thirds European. It has high NCEA Level 2 achievement rates, a fairly new principal, and a new deputy principal.

- Mataī School is a very large, co-educational state school with a diverse study body, a sizeable Māori population, and a fairly new principal. Māori and Pacific student achievement are above the national average.

- Rātā School is a medium-sized, single-sex state school with a large Māori population and an established principal. Māori student achievement at Rātā is significantly higher for NCEA at all levels compared to Māori students nationally.

Data Analysis and Focus

Interviews were recorded, transcribed, and imported into the software package NVivo 11. Before coding began I listened to interview recordings several times, and read and re-read

interviews as a whole to try to see the big picture and grasp basic themes in the data. Theoretical thematic analysis was used to analyse the data (Patton, 2001). Because this study focused on the pathways from school self-evaluation to school improvement, analysis focused on participants' discussions of the processes they selected from the list of 11 intermediate outcomes described above. Participants highlighted how practices like tracking and monitoring worked in their school. The identification of the importance of the traffic light tracking and monitoring tool led to the focus of this paper. In analysing the way participants talk about how the processes of tracking and monitoring work in their setting, the importance of the traffic light tool became apparent. Therefore, all segments of text related to their tracking and monitoring data were then grouped together. When I noticed the pattern of the prevalence of these practices, I searched for additional examples among the interview data by re-reading the transcripts and coding using NVivo. I asked questions of the interview data such as: What data are teachers attending to? What tools are they using? What changes have occurred around data use in recent years? What variations existed among different schools? and What interventions did they put in place in response to their data? This allowed an analysis of how schools use data and the impact the traffic light tool has had on schools.

Results and Discussion

The analysis of the data revealed the significance of the traffic light tool and the changes in practice that resulted from its use. The tracking and monitoring tool became important to these schools and led to structural changes within the schools. Before exploring these changes, the traffic light tool will be explained.

Importance of the tracking tool and changes in practice

The schools in this study had all implemented a traffic light software tool for tracking and monitoring student achievement of NCEA credits throughout the year. The software tools usually kept track of each student's credits accumulated from ongoing internal assessments, credits still needed to pass the level, and the ratio of credits passed to credits attempted. It classified students as either red, amber, or green according to how likely they were to achieve the NCEA level. These ratings were used by deans, tutor group teachers, academic counsellors, and/or the senior leadership team to identify students who might be struggling with particular subjects or assessments.

One school explained that they had originally implemented the tool after they analysed the previous year-end results and realised how many students had been close to the necessary 80-credit threshold but had not actually achieved it. Because of NCEA's complexity and its flexibility in building credits, students sometimes choose not to complete certain internal assessments, thinking that they will have enough credits to meet the requirement, only to discover at the end of the year that they do not (Hipkins et al., 2016). School leaders are also often caught unaware. A deputy principal explained why they had initially created and implemented a traffic light tool after learning that a number of students in the previous year had been close to achieving NCEA:

> We were surprised to discover how close some students had got to getting over the line. Why didn't we know that these kids were this close and why didn't we do anything about that? (Totara, deputy principal)

This is probably the main problem the traffic light tool was intended to solve. The objective is that the tracking tool produces simple to understand ongoing data that are easily assessable to both teachers and students.

Use of the tracking tool had led to changes in school practice and operation. For example, all schools shared more data about students' progress with the students and their whānau. It led to the creation of online portals showing real-time results of NCEA internal assessments. This was done to help students better understand their progress so they could make their own plans for improvement. Thus, students became participants in the data use process as well. These schools did not publicly display students' colour-coded assessment results. In previous research in other contexts, Neuman (2016) found that exhibiting data in that way can further marginalise students when they are reminded that they are not achieving.

Another important structural change as a result of the use of the data tool was an altered role for the tutor teacher, and in some schools, the change was fairly significant. In one school, the tutor teachers' responsibilities had changed to involve monitoring the number of credits students were achieving, and following up with the student, whānau, and other teachers. In fact, this new role was part of their appraisal. A head of department at the school explained the new role:

> The role of a house group tutor involves checking credits making sure these kids are on track. It has become quite a big job and we actually get appraised as a house group

tutor and the things that they look at are attendance, they look at how well you have communicated with whānau or caregivers, how well you have communicated with the dean [about students'] academics and where they might have fallen down. And if you've gone to a teacher that the kid might have to say look, such and such has 72 credits and they need 8 more, is there any chance they could get them in your class? (Rata, head of department)

The traffic light tool influenced who in the school interacted with whom around the data and in what ways. It appears that the change in the role of the tutor teacher de-privatised teachers' work and made student achievement of NCEA a more collective responsibility. Datnow and Park (2018) argue that shared responsibility is an important element of data use for equity.

Several teachers commented on how important this traffic light data tool had become in their school. When asked to identify three activities in their school that were priorities as part of the process of using assessment data for school improvement, a teacher indicated the significance of the traffic light tool:

> The tracking and monitoring of student achievement has been huge in this school. That is probably the first, second and third [priority] really. (Matai, teacher)

Other teachers lamented the overall way NCEA results had become so crucial:

> It's like NCEA has almost replaced everything else that teachers do … It all comes back to where are the credits? ... So it is almost like the success rate of students becomes the only tool by which you measure the effectiveness of a teacher. (Matai, teacher)

> It had very much been a school culture of academic results at the cost of everything else. It was almost as though that part of the school didn't even exist, it was just 'Oh look at our good NCEA results'. (Kauri, teacher)

These comments demonstrate the pressure of external accountability on the way schools focused on and responded to student achievement data. They suggest the system as a whole was over-valuing the metrics of achievement and this context influenced the schools' processes.

Focusing attention and school responses

In all schools in this study, the existence of the traffic light monitoring system focused attention on students on the cusp of achieving an NCEA qualification (e.g., NCEA Level 1 or 2). This section will explore this focus and the schools' responses. The data in the tracking system called attention to students on the cusp and encouraged educators to, at the very least, notice this group. This tool put disproportionate attention on this particular group of students. A teacher described the fundamental purpose of the tool this way:

> It's a traffic light system that allows you to see red, amber and green as to how many credits students have achieved and who is going to be in the zone where you are going to have to be concerned about them. (Rata, teacher)

One of the teachers wondered if merely identifying these students constituted school self-evaluation when she pondered her own genuine question:

> Is it self-review when you identify your cusp kids? (Kauri, teacher)

Because data use is an act of interpretation, the first step is when individuals *notice* certain data or patterns in the data, so this noticing process is an important part of research on data use (Coburn & Turner, 2011). The data tools schools utilise can shape these interpretative processes by influencing what teachers and school leaders notice and talk about in data use routines (Coburn & Turner, 2011). Tools can shape the attention, thinking, and behaviour of members of the organisation and they can convey to teachers the types of data that they should notice, and how they should discuss and interpret these data (Dyson, 2018). Some scholars worry that a traffic light design is problematic since it oversimplifies the results, focuses attention on the cusp students, and does not provide guidance on appropriate next steps (Mandinach et al., 2018).

In response to this focus on cusp students, all schools then used these data to target student-level interventions at specific groups of students on the threshold of achieving. This tool had shifted educators' ability to respond in real-time, whereas participants explained that, previously, analysis of their achievement data would occur retrospectively after the school year was over when they would analyse cohorts of students rather than individual students. A deputy principal explained the new proactive use of interventions in a timely manner:

> We're putting in interventions here. What it's been like in the past is here's the

coroner's report and we just go, 'Oh that's terrible'. We've now gone from being

spectators to being players. We're putting the cones out. (Totara, deputy principal)

Most of the interventions can be considered short-term solutions. These included
remediation classes, conducted at lunchtime or after school, and homework and study groups
facilitated by a resource person. One school referred to these opportunities as "sprinter
schools". All schools also offered opportunities for reassessment in a previously
unsuccessful standard or assessment in an additional standard in order to gain enough credits
for a qualification at the end of the year.

Different participants responded differently to these targeted interventions. One teacher had
identified that these decisions were not merely data driven but had political implications. He
objected to the opportunities not being presented equally when students who were identified
as being on the cusp of meeting the standard were offered additional assistance in small-
group settings:

> They are in smaller groups and there are several people monitoring them during that
> period, working with them and making sure they've got the resources and are
> working without distraction and that kind of thing. So it is more of a targeted
> response, which becomes political. We have senior classes that are 28, 29, 30
> [students], which is not ideal and so some of these kids are coming out of those
> classes and in period 6 they are in a group of six or seven or eight [students] which
> makes it much easier for the teachers to work with them. Obviously, it would be
> easier if they were all in classes of 20 to start off with. (Mataī, teacher)

This teacher objected to the structural reorganisation of the school and appeared to view it as
an example of educational triage (Gillborn & Youdell, 2000), when decisions are made to
disproportionately focus instructional resources on students close to the threshold of
achievement.

Focused on cusp students

While all schools focused on cusp students, the data use practices at one school, Mataī, were
more focused than at the other schools. Participants in the school held differing opinions of
this focus. In the following quotes, two participants discussed the traffic light tool helping to
identify students who are close to achieving NCEA. On the one hand, the principal illustrates
his reasoning that not all students are going to achieve the standard so he should focus on

students who are close. On the other hand, a teacher laments this focus on the cusp students and explains that this causes frustration for teachers in the school.

> In terms of tracking, our Year 11, 12, and 13 students are RAG rated, red, amber, green … And then you've got to cut your cloth.[3] There are some kids in the red that I can't help, so I've got to spend all my time on the amber kids to push them over. If we work hard on our red kids, there will be more failure, because we can push a kid from an F to an E, but the amber kids we can push from a D to a B. We can push these kids forward and get them over. (Mataī, principal)

> So it is good to be able to identify, but I think most of what we are doing genuinely seems to be identifying the ones who are close. The ones who are within 10 credits of passing. There doesn't seem to be much that we do for ones who are 40 credits away and are unable to get there … and I guess that is where a lot of frustrations come in. (Mataī, teacher)

While the teacher questions the focus on the cusp students, the principal implies that the only worthwhile improvement is one that converts a cusp student to a student who achieves a qualification. The principal's approach is a more stark example of educational triage (Gillborn & Youdell, 2000), when a school's instructional resources disproportionally focus on students on the cusp of achievement at the expense of other students. The principal appears to feel external accountability pressure to improve NCEA achievement results which has shaped his view of how he should respond to the data. The principal's data-informed decisions may be influenced by the pressure of wider political drivers, such as the repercussions of assessment scores printed in popular media. The principal may feel more pressure to conform to external agendas of accountability because he is in a position of ultimate responsibility(Hammersley-Fletcher, 2013). The literature shows that school leaders play an important role in shaping data use in their schools and can influence whether it will result in the thoughtful use of data, or accountability-driven data use oriented towards short-term fixes (Coburn & Turner, 2011; Datnow et al., 2017; Datnow & Park, 2018).

The principal's response to cusp students could be considered rational under accountability regimes, but not equitable. Targeting cusp students is rational, from a standpoint that this is the most efficient way to maximise outcomes for the least cost if organisations are looking

for areas with high potential for impact and strong value for money. "When proficiency rates are the target, 'triage is a sensible response, as it optimizes resources and effort' while focusing squarely on the desired outcome" (Horn, 2018, p. 401). Ball (2003) reiterates that if a school feels pressure to increase the number of students who gain a qualification, "they would be unlikely to 'invest' in work with children with special needs where the margins for improved performance are limited. In the hard logic of a performance culture, an organization will only spend money where measurable returns are likely to be achieved" (p. 223). When schools focus on cusp students, this can limit educational opportunities for other students and therefore compromise the equity goal of data use; "equity cannot be achieved unless data use efforts focus on all students" (Datnow & Park, 2018, p. 144).

Data as a starting point

Tōtara and Rātā Schools took a different approach to responding to the traffic light data. While all schools in the study had short-term solutions, personnel at these schools also used the data as a stimulus for further data gathering with students. School personnel would use the data from the traffic light tool to identify potentially struggling students and then offer academic counselling to identify individual students' needs and provide advice and supports to help resolve problems. A teacher described an example of this process where a student's lack of achievement in a class was determined to be due to certain family obligations that were causing him to be late for school and other issues. The school offered him special allowances such as permitting him to arrive late for class but giving him additional makeup work online. The teacher explained:

> That is not going to work every time, but the only way you find that out is by tracking the students and caring for them as an individual and in an ideal world that would happen. Obviously some students still fall through the cracks, but with technology and with the traffic light system we are implementing a really good programme where we do actually cater for the students. (Rātā, teacher)

In this example, the teacher did not respond to the student's poor achievement data by suggesting that the student engage with a quick-fix short-term solution, but instead gathered more information from the student to determine the problem, and then changed her own practice earlier in the year to accommodate the student's personal home life circumstances. Examples like this where the data were the beginning of a conversation were perhaps the ideal use of the student achievement data. The data became the catalyst for investigating

academic issues and seeking solutions that may involve pastoral care or other support issues. Data could be a starting point for deeper solutions and not merely short-term solutions. This example also highlights the need for diverse sources of data that can even include information about a student's transportation and home situation (Mandinach & Schildkamp, 2020).

Conclusion

A traffic-light data tool became a powerful force in the schools in this study. Data use practices have shaped these teachers' work and caused structural changes in these schools, such as altering teachers' roles in their tutor groups. The data informed which students were the focus of attention and received additional services. In all schools, there was a disproportionate focus on cusp students and all had targeted interventions aimed at those students. Sometimes the data were a starting point, a catalyst for identifying students and then having further conversations with them to determine barriers to success and possible solutions. This individualised approach could be productive, helpful, and possibly lead to real student learning and not merely short-term fixes. However, sometimes the data use tracking and monitoring tool led to data misuse, such as an emphasis on short-term solutions for cusp students, potentially at the expense of other students and their educational needs. This study shows that educational triage happens in New Zealand. The practice of triage has become normalised and the traffic light tool allows it to be more targeted. The practice of triage was influenced by external accountability pressures that equate NCEA Level 2 achievement numbers with school quality and are communicated through government targets or popular media. This is important in a context of consumer accountability and school competition where schools' reputations become vital.

It should not be misinterpreted that this paper is arguing against data use or against the specific traffic light tracking and monitoring tool in this study. Of course, schools need an accurate way to track student progress. No one would argue that students and teachers should not be aware of how likely it is that a student will achieve a qualification. No one wants students to slip through the cracks, believing that they will achieve NCEA only to be surprised at the end of the year. However, this study does suggest that data management tools are necessary but not sufficient for effective and equitable data use.

This study shows that data use can be a powerful force, with the potential for good. But it also raises some concerns about unintended consequences of data use. It is important that

educational research makes data use practices visible (Coburn & Turner, 2011; Selwyn, 2014) to draw attention to potential unintended consequences. The hope of this article is to highlight the unintended consequences of data use practices to learn from them and put in place processes that better support student learning. If educators are aware of the possible effects of their data use, they will be better prepared to mitigate the less desirable practices and help prevent unintended inequitable outcomes. This study contributes to the literature by highlighting the less-studied outcomes of changes in practice and organisational changes that occurred as a result of assessment data use in selected New Zealand high schools.

Some literature and policy about educators' data use implies that simply having accurate data available will lead to improved practice (Datnow et al., 2012; Mausethagen et al., 2018). While it is important to have access to accurate data, the presence of data on its own does not ensure improved practice (Mandinach & Schildkamp, 2020). In much of the educational literature, "data are treated as an unquestionably positive force" (Booher-Jennings, 2005, p. 261). This study has shown that data use can focus attention on individual students so their needs can be met, but also that data can be used to disproportionately target some students at the expense of others, and potentially undermine the equity goals of data use. These results demonstrate the need to call attention to these issues in the research literature, teacher preparation programmes, and teacher professional development.

Chapter Eight

Introduction to Article Three

Article Three highlights internal accountability and focuses on internally-actioned organisational ECB. The subject of ECB was the only topic that I had identified a priori, prior to conducting my study. I had a previous interest in ECB and sought to understand how New Zealand secondary schools build evaluative capacity. Internationally, ECB is an increasingly important topic within schooling systems (Organisation for Economic Co-operation and Development, 2013) and a topic of increasing interest in the field of evaluation, as expectations for internal evaluation is growing within all types of organisations including schools.

While there is growing demand for teachers to be able to use data, international research has shown that schools often find self-evaluation processes to be difficult and that school personnel do not have sufficient capacity to carry out effective self-evaluation. Creating sufficient capacity to carry out effective self-evaluation can be facilitated by ECB, the intentional work to increase the ability of an organisation to conduct and use evaluation (Stockdill, Baizerman, & Compton, 2002). ECB is a complex phenomenon involving individual learning, organisational change, and a variety of outcomes (Preskill & Boyle, 2008), which makes it a topic well-suited for my realist analysis of these change processes.

While the general topic of ECB was identified ahead of time, the theoretical lens of the sociocultural learning theory that I use to examine schools' ECB processes was developed inductively. As other articles do, I examine the ECB literature in such a way as to investigate the underlying theories and assumptions embedded in this field and in doing so, I identify that ECB is normally based on a more behaviourist learning theory. This article explores how different learning theories can be used as a framework for ECB, and how the field of evaluation could benefit from theories about learning from the field of education.

Article Three also picks up on several themes that were explored in the previous articles. Article One explores how schools encourage teachers to engage in inquiry. In that article, members of senior leadership teams and heads of department explain that they try to help teachers improve the quality of their inquiries. Their processes need to balance between scaffolding teachers' learning and giving teachers freedom on one hand, and encouraging school coherence and holding teachers accountable on the other. Article Three delves into

some of the specifics of how the participating schools built capacity internally. This article also reinforces themes from Article Two, such as the importance of the principal and school leaders in setting the tone of school evaluation, as well as the importance of the tools and routines schools employ.

I chose for this article to be published in an evaluation journal because I viewed one of the contributions was bringing theory from another field, education, into evaluation capacity building.

Article Three

Drawing on Theoretical Knowledge to Build Evaluation Capacity

Introduction

Many organisations now carry out self-evaluation, which requires the capacity to do and use evaluation. Organisations lacking sufficient capacity often engage in ECB to increase their ability to conduct and use evaluation (Stockdill et al., 2002). Nearly all ECB literature describes how professional external evaluators build capacity in specific organisations or systems. In contrast, this article explores how people who are not professional evaluators are building evaluation capacity in their own organisations. It did so by studying selected secondary schools in New Zealand, which are expected to carry out ongoing self-review processes. Previous literature shows that New Zealand schools' evaluation capacity is highly variable with few schools having strong processes in place, possibly because of insufficient systemic ECB (Timperley, 2013). This study aimed to contribute to the conversation around internally-actioned organisational ECB by identifying secondary schools that are effective at self-evaluation and exploring how they build capacity. After a brief look at the literature on evaluation capacity, ECB, and the context of school self-evaluation in New Zealand, this introduction turns to an explanation of sociocultural learning theory. This learning theory was a lens for understanding how New Zealand secondary school teachers build evaluative capacity.

School Self-Evaluation in New Zealand

Since New Zealand schools are self-managing, the system depends on school self-evaluation. Self-evaluation is evaluation that is initiated and conducted by the school itself to assess its functioning and support its decision making, learning, and school improvement (Schildkamp & Vissher, 2010). The scope of this activity can vary from individual teacher reflective practice through to school-wide analysis of data to inform planning and resource allocation (Mutch, 2012). Each New Zealand school is expected to carry out an ongoing process of evaluation for improvement but can select the approach as well as the tools used (ERO, 2015). An Organisation for Economic Co-operation and Development (OECD) report has noted that, in New Zealand, "school self-review is at the heart of quality assurance and improvement processes. The basic premise is that schools are best placed to analyse their own contexts" (Nusche et al., 2012, p. 89).

Evaluation capacity

In New Zealand schools, as in any type of organisation, conducting self-evaluation requires evaluation capacity. Evaluation capacity is the competencies and structures necessary to carry out evaluation and the ability to use evaluation findings in decision-making processes (Bourgeois, Whynot, & Thériault, 2015). Evaluation capacity encompasses the evaluation knowledge, skills, and attitudes of individual people. These include the ability to develop evaluative questions, use logic models, conduct surveys and interviews, analyse data, as well as having a favourable attitude towards evaluation (Huffman et al., 2008). Evaluation capacity also emphasises the organisational-level processes necessary to create and sustain quality evaluation. These can include strategic planning, continuous improvement strategies, professional development policies, and positive attitudes towards evaluation (Stockdill et al., 2002).

Evaluation capacity in the context of school self-evaluation is an issue in many countries around the world. Most economically advanced countries now prioritise school self-evaluation, but school personnel often lack sufficient evaluation capacity to carry it out effectively (MacBeath, 2006; McNamara & O'Hara, 2008b; Vanhoof & Van Petegem, 2013). Self-evaluation processes "demand knowledge and skills from schools which cannot be simply pulled out of a hat … It is no easy matter for schools to initiate and implement a systematic and cyclical process of self-evaluation" (Vanhoof & Van Petegem, 2013, p. 275). Even if they can obtain self-evaluation data, many schools then find it difficult to use this information to make school improvements (Schildkamp & Vissher, 2010). Several reasons account for this difficulty, including time constraints and teachers' lack of knowledge in adapting their instructional strategies (Datnow et al., 2012).

Evaluation capacity building

Support for organisations to develop evaluation capacity can occur through ECB, the intentional work to increase the ability of an organisation to conduct and use evaluation (Stockdill et al., 2002). "The ultimate goal of ECB is sustainable evaluation practice—where members continuously ask questions that matter, collect, analyse, and interpret data, and use evaluation findings for decision-making and action" (Preskill & Boyle, 2008, p. 444). Building evaluation capacity is not merely skill-building, it also involves building evaluation practice and incorporating evaluation into the life of an organisation. ECB tries to create a system where the use of evaluation is standard practice, or the ordinary "way we do things around here" (Stockdill et al., 2002, p. 9).

The ECB discussed in literature is nearly always an intervention by a professional evaluator to enhance the capacity of an organisation to do and use evaluation. ECB generally involves an evaluator providing training, technical assistance, written materials, consultation, coaching/mentoring, or immersion approaches to one or more staff within an organisation or system (Huffman et al., 2008; Preskill & Boyle, 2008).

New Zealand school evaluation capacity and ECB

As in many other countries, the evaluation capacity of New Zealand schools is highly variable and only a minority of schools have strong evaluation processes in place (Mutch, 2012; Nusche et al., 2012; Timperley, 2013). An MOE initiative found that evaluative capacity was crucial to, but lacking in, school self-review processes (Timperley, 2013). Similarly, in a school improvement initiative, the majority of schools, even with support, did not have robust enough data to build a self-review process (Lai, 2013). A recent ERO national report indicated that secondary schools are not universally using effective self-evaluation processes (ERO, 2014). Only one-quarter of schools in the sample were judged to be effectively inquiring into their annual NCEA achievement data and then developing activities, innovations, or approaches to improve achievement. Since schools are self-managing, the "success of the system is very dependent on the capacity of individuals in a variety of settings with differing skills, knowledge, and ability levels" (Ryan & Timmer, 2013, p. 201).

Several scholars, such as Timperley (2013), cite the insufficient investment in systemic ECB in the New Zealand education sector. Lai (2013) agreed that the responsibility for school self-evaluation is given to "individual schools with little guidance from central government on how to do so" (p. 68). An OECD review of the New Zealand schooling evaluation system recommended improving support structures to promote school self-evaluation (Nusche et al., 2012).

Drawing on Learning Theory

Because evaluation capacity is variable in New Zealand schools and has lacked major, systemic ECB, this study aimed to identify secondary schools that are effective at self-evaluation and explore how they build capacity for these processes. As will be discussed in the results section, the findings showed that these secondary schools appear to be drawing on concepts of sociocultural learning theory to build evaluation capacity. The study did not set out to use this approach, but analysis of the data suggested that this theory reflected what

was happening in schools and could be a helpful framework for understanding ECB. To begin to understand how the schools in this study drew from sociocultural learning theory and what this theory can bring to ECB practice, a brief overview of this theoretical tradition is helpful. It is proposed that ECB pay greater attention to learning theories and, more specifically, sociocultural theory. Currently, the literature on ECB is largely atheoretical (Huffman et al., 2008). The related literature on interventions to increase teachers' use of data also lacks theory (Marsh & Farrell, 2014).

Before expanding this argument, it is necessary to explain what is meant by "theory" here, since the field of evaluation understands the concept of theory in "fragmented" and "often confusing" ways (Leeuw & Donaldson, 2015, p. 478). Evaluators draw on "evaluation theories" (i.e., theories of practice for conducting evaluations); for example, empowerment evaluation, or participatory evaluation. Another widespread use of "theory" in evaluation is a theory of change approach describing processes of change in an initiative by outlining linkages between the intervention and expected outcomes. These programme theories might be called *small-t theories* (Leeuw & Donaldson, 2015). These are contrasted with *capital T* theories, sometimes called *explanatory, substantive,* or *scientific theories.* Leeuw and Donaldson (2015, p. 472) argue that evaluation should draw on more capital T theories that are "tested and robust explanatory theories from the (social, behavioural and policy) sciences". Schwandt (2014) echoes this, suggesting that drawing on theoretical knowledge can help with dilemmas of practice. Similarly, this article argues that the field of evaluation could benefit from attending to and drawing on explanatory theories.

One category of theories for the field of evaluation and particularly ECB to potentially draw on are learning theories. These include behaviourism, cognitivism, metacognitivism, social constructivism, and sociocultural theories. For example, in the behaviourist model, learning is an individual activity, learners are fairly passive participants in the learning process, and knowledge is generally seen as transferable from context to context (Svinicki, 1999). The desired behaviours are divided into small, carefully sequenced steps, with each step taught to mastery before going on to higher level steps. Although they do not use the label of "behaviourist", Huffman and colleagues (2008) contend that ECB is often conceived of as a linear, step-by-step process, which is similar to a behaviourist approach.

Sociocultural theory

The sociocultural approach to learning sits in contrast with this approach. Originally based on the writings of Vygotsky (1978), a variety of sociocultural theories of learning have been proposed (e.g., Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991). Learning is not conceived as an isolated process of an individual's acquisition of information, but as a process of meaning-making and knowledge construction that occurs through social interactions and activity. Sociocultural learning theory argues that learning is inherently a social phenomenon that occurs through dialogue and reflection with colleagues (Lave, 1996; Vygotsky, 1978; Wenger, 1998). Through dialogue, practitioners can challenge each other's beliefs and interpretations, which "can lead to new shared understandings and deeper engagement in particular activities than would otherwise be possible by individuals operating alone" (Honig & Ikemoto, 2008). The sociocultural lens also emphasises the importance of the particular context where learning occurs and the learner's connection with, and involvement in, this environment.

Within this social learning, various "scaffolds" or supports help learners to deepen their engagement in particular learning processes (Vygotsky, 1978). These scaffolds support the learner's understanding and independent performance. One such support is assistance from an experienced teacher or "more knowledgeable other" (Vygotsky, 1978). A practice the teacher might use is *modelling* particular activities by demonstrating the practice and explaining their underlying thought process to bring "thinking to the surface" and make it "visible" (Brown et al., 1989, p. 33). Honig and Ikemoto (2008) claim that it is "particularly powerful" to engage "others in dialogue about the purposes and nature of the practices—so others know not just what participation in these practices entails but why they should participate in particular ways" (p. 333).

Another scaffold is the use of *tools* to help deepen individuals' engagement in particular practices. According to sociocultural learning theory, tools are the manifestation of new ideas (Marsh & Farrell, 2014; Wenger, 1998). Tools may be physical, such as a shovel, but tools can also be psychological; much as physical tools can support physical activities, psychological tools can support human learning (Vygotsky, 1978). Grossman, Smagorinsky, and Valencia (1999) delineate two kinds of tools: conceptual and practical. Conceptual tools are principles, theories, frameworks, and ideas designed to frame how people think about certain issues; people use conceptual tools as heuristics to guide their decision making and practice (Grossman et al., 1999). Practical tools are more concrete practices, resources, and

strategies. Tools help communicate messages about what individuals in a workplace should and should not do; however, the meaning of these tools is not set in stone, but is negotiated within a community (Brown et al., 1989).

Sociocultural theories also emphasise the concept of learning as participation in practice, not separate from practice. Schwandt (2005) points out that many capacity-building workshops, seminars, and courses conceive of *learning* as separate from everyday practices of *doing*: "we readily assume that 'learning' is some activity that takes place on a special occasion when a practitioner is not busy 'doing'" (p. 328). However, in sociocultural learning theory, learning and application are not separate processes.

In education, sociocultural learning theories are used to make sense of student learning and teaching, and several scholars have used these theories to understand school improvement efforts (Gallucci, Van Lare, Yoon, & Boatright, 2010; Honig & Ikemoto, 2008; Ikemoto & Honig, 2010; Marsh & Farrell, 2014). This study draws on sociocultural learning theory as a lens for understanding how New Zealand secondary school teachers build evaluative capacity.

Methods

Participants and settings

This study was part of larger research on school self-evaluation in New Zealand secondary schools. This study tried to identify selected secondary schools that are effective at self-evaluation and explore how they build capacity within their organisation. Four secondary schools were purposefully selected based on three criteria: they had previously been identified by ERO as successful at self-evaluation; they were mid-decile schools (deciles 4–8); and they had student achievement levels that were above average for their decile level. To identify schools, the ERO reports of all 37 mid-decile schools in two geographic areas of the North Island were read along with other school documents, school websites, and independent publications on school achievement levels. Schools with similar decile ranges were chosen because they have similar levels of funding available to them for self-review and capacity building. Within these selection criteria, diversity was sought among the schools in terms of size, location, single-sex/co-ed, state/integrated, and length of tenure of the principal.

The four schools were each given a pseudonym based on a New Zealand native tree, and are described briefly below:

- Totara School is a very large, co-ed state school with a very diverse student body and an established principal. It has high University Entrance results for both boys and girls for its decile level.

- Kauri School is a large, co-ed, integrated (Catholic) school, with a less-diverse student population that is approximately two-thirds European. It has high NCEA Level 2 achievement rates, a fairly new principal, and new deputy principal.

- Matai School is a very large, co-ed state school with a diverse study body, a sizeable Māori population, and a fairly new principal. Māori and Pacific student achievement is above the national average.

- Rata School is a medium-sized, single-sex state school with a large Māori population and an established principal. Māori student achievement at Rata is significantly higher for NCEA at all levels compared to Māori students nationally.

Data collection

The primary data collection method for this study was in-depth semi-structured interviews with 13 people. Within each school, three people were usually interviewed: a principal or deputy principal, and two staff who were not part of the school's senior leadership team. At the first school, an additional head of faculty was interviewed. Many of those interviewed were heads of faculty or department, and all were knowledgeable about their school's self-evaluation processes. Interviews lasted approximately 1 hour, ranging from 58 minutes to 79 minutes. (See below, Data analysis, for information about interview questions.) Documents from each school were also analysed to gain a more complete understanding of the schools' processes. These included ERO reports, background information on the schools, and tools used by the schools to build capacity, such as templates and handbooks.

Data analysis

Data analysis began with transcribing the audio recordings of the interviews and reading them to become familiar with the data. For this study, segments of interview data relating to ECB were pulled out from the wider interview data. These were primarily participants' responses to the interview questions: "How have teachers and school leaders in this school acquired the capacity to be able to carry out these (school self-evaluation) processes?"

"What resources have you drawn on to build capacity?" Any other instances in the interviews where participants discussed the concept of capacity building were also sought.

I began by writing detailed descriptions of each school's capacity-building efforts, with an emphasis on how they differed. After further data analysis and in-depth discussions with two colleagues, I noticed commonalities between the cases and therefore decided that a thematic analysis was more appropriate (Miles et al., 2014). I then re-read the data, assigned codes to segments of data, and grouped together similar data into themes. After several rounds of data analysis, I identified that educators were describing using processes consistent with sociocultural learning theory and that this framework could help to understand the way schools were building evaluative capacity. When I noticed the pattern of the prevalence of these practices, I developed codes for these and searched for additional examples among the interview data.

Results and Discussion

Lacking capacity

Participants suggested that some teachers in their schools lacked capacity in evaluation activities. They indicated that some evaluation tasks are new for teachers, who may find them difficult. Often there is an assumption that teachers are all capable of doing these tasks. A participant explained that this includes believing that teachers can effectively interpret their data:

> There's an assumption that everyone knows their data or has an ability to interpret it, because *I* can look at my data and say 'high', 'low', 'issue'. I can do that, so I think we make an assumption that all teachers can do that. (Totara, deputy principal)

Some admitted that, in the past, perhaps they had not focused on capacity building to the extent that was needed. The same deputy principal said, "You've got to teach people to evaluate effectively. I think maybe we've been guilty of just saying go on, review. And I think you have to teach people how to review." The remainder of this section explores *what* and *how* the schools taught teachers to evaluate effectively. Before discussing *how* schools attempted to build capacity, it is valuable to understand *what* skills and traits were the focus of ECB efforts.

Sense making

Schools in this study emphasised building capacity for sense making rather than narrow technical skills. These schools did not focus their capacity building on increasing teachers' technical data collection and analysis skills such as building and accessing student data systems, producing graphs, or making statistical calculations like determining effect size. Instead, the schools focused on learning how to engage in meaningful inquiry, particularly collaborative inquiry. The schools all had effective data management systems for tracking and monitoring student achievement but believed that capacity building was still important; they saw the monitoring systems as necessary but not sufficient for effective self-evaluation. They focused on building capacity for using the available data to improve school practices through asking questions, interpreting information, solving problems, and developing action plans based on the data. Participants explained that they believed the interpretation of data was more important than technical data skills:

> I don't think individual teachers necessarily need to have a lot more work on tracking the data … The important part is the interpretation that comes through in discussion within departments. That is where you can say we notice this year we had lower success in that [NCEA] standard than the other, why is that? (Matai, teacher)

> We want the domain of the teacher to be on reflection, not around the skills required for analysis of data. [We say to teachers] 'Okay, we're going to ask you to do unit review but we're not going to ask you to have to do all the calculations, we just want you to worry about the reflection.' (Rata, deputy principal)

This emphasis on reflection, interpretation, and improvement rather than on discrete technical skills is reflected in the concept of evaluative thinking. This kind of evaluative thinking requires sense making, setting priorities, asking questions that matter, translating findings into instructional practice, and ensuring equity. Evaluative capacity is not about the data in itself, "but about the quality of the knowledge that emerges from this process. As we see it, good knowledge is founded on asking good questions, having good data, and engaging in good thinking" (Earl & Seashore Louis, 2013, p. 199).

The schools' focus on interpretation, evaluative thinking, and improvement sits in contrast with much of the literature on capacity building for data use in schools that focuses on the more technical skills required to carry out evaluation. In a programme to support New Zealand schools, Gan, Irving, and McKinley (2014) have focused their capacity-building

efforts on enabling schools to develop and maintain databases and data management systems. A recent survey in the US found that, although most districts are providing schools with support on how to access student data from data systems, they are far less likely to provide educators with training on how to use the system to analyse student achievement or use the results to change their instructional practice (Means, Padilla, & Gallagher, 2010). Much of the professional development about data use in education focuses on teaching data literacy skills in isolation rather than on how to *use* data to potentially change practice (Wayman & Jimerson, 2014). The schools in this study appeared to understand the importance of building capacity in the often-neglected skills of using self-evaluation data to make school improvements.

How Capacity was Built

The previous section discussed *what* schools focused on in their ECB and we will now turn our attention to *how* they built capacity. Previous studies using sociocultural learning theory to understand school improvement efforts identified several practices that were important in the current study: collaborative work, modelling, and the use of tools (Gallucci et al., 2010; Honig & Ikemoto, 2008; Ikemoto & Honig, 2010; Marsh & Farrell, 2014).

Collaboration

For the schools in this study, ECB was collaborative and occurred through social interactions. Collaborative inquiry involves forming groups of educators who can work together as they use data to examine and improve their own practice or overall school functioning. Teachers in these schools engaged in many formal and informal social processes to build evaluative capacity. Even when teachers were conducting individual inquiries into their own practice, this was supported by collaborative groups. Teachers' individual inquiries were also linked to whole-school evaluation and development through capacity building. The group structures took different forms, but the most frequent were evaluative discussions in departments (called "co-construction meetings" in one school) or cross-curricular professional development in professional learning groups (PLGs). A head of faculty described that these cross-curricular groups were not ad hoc, but required planning:

> We also thought quite a lot about collaboration and how to put people together and who might be a critical friend to somebody … We all came together across faculties in a group of say 10 or 15 and we talked about a [capacity building] reading for instance, and we pooled our ideas and reflections. (Totara, head of faculty)

The purpose of the groups was to conduct self-evaluation collaboratively, support educators in conducting inquiry into practice, and make this inquiry productive (Wayman & Jimerson, 2014).

This emphasis on collaboration is different from the way teachers often work in isolation (Little, 1990), but is strongly supported by the sociocultural learning theory understanding of learning as an inherently social phenomenon that occurs through dialogue and reflection with colleagues (Lave, 1996; Vygotsky, 1978; Wenger, 1998). Schwandt (2017) explained this social learning process:

> In dialogues, people exchange experiences and perspectives, and this helps them to gain a better, fuller understanding of the situation … [D]ialogue across these perspectives … is understood as an ongoing, social learning process in which participants develop new and richer understandings of their practice. (p. 15)

The concept of educators learning through social interactions is reflected in a trend in the wider teacher professional development literature towards the social in adult learning (Timperley et al., 2007) since research suggests that educators learn collaboratively (Wayman & Jimerson, 2014). These observations are not limited to educators, however, and the importance of learning through social interactions may be valuable to the practice of ECB in any sector.

Modelling

A specific collaborative practice that was used by the schools to increase skills and improve attitudes was modelling. This was usually used when teachers or school leaders shared their experiences of conducting an inquiry into their own practice. In PLGs, teachers shared their experiences, and articulated their thinking and the decision making behind their actions. Then they reflected on what had occurred in the classroom as a result of experimenting with different practice. At Totara School, modelling had become a formalised, ongoing process in which each school term, eight or nine teachers were chosen to publicly share their inquiries:

> We are picking high quality inquiries to present and try to model that and say to people look, here is what we think is good practice … The teacher will talk about how it went or what has occurred and what was learned. And that has proved to be pretty powerful. (Totara, head of faculty)

Teachers modelling effective inquiries and reflecting on what they had learnt as a result brought their "thinking to the surface" and made it "visible" (Brown et al., 1989) so others could learn from it. A deputy principal explained that this process was effective even if teachers did not share the same content area (e.g., maths, history, physical education). Teachers could learn from others' experiences and apply it to their own content area. By modelling the underlying skills of self-evaluation, leaders attempted to increase understanding and capabilities to facilitate teachers doing this work independently in the future.

Although scholars emphasise the cognitive and metacognitive skills learnt from modelling (Lave, 1988; Marsh & Farrell, 2014), the teachers in this study seemed to also consider the affective aspects and to view modelling as a nonthreatening way to increase positive attitudes towards self-evaluation. A participant described how she modelled her own inquiry for a resistant colleague:

> Previously he had been very resistant. And I'm just quietly showing him my teaching as inquiry, being very careful. I'm just showing him what I'm doing and the video I've taken of students for my inquiry. And I say 'this is going to be my scanning, this is going to be my bringing in data' … We're starting to see some progress. It helps with buy-in. And ever so slightly it's shifting. (Kauri, head of faculty)

In this quote, the teacher makes explicit her own processes and does so *quietly* and *carefully*, implying that it is important to be gentle and nonthreatening in order to overcome her colleague's resistance. Modelling also assisted with alleviating evaluation anxiety through school leaders showing vulnerability. Participants explained that listening to how colleagues described their self-reflection was "empowering" to others and made them more open to discussing their own practice. At Totara School, a participant suggested that leaders showing their own vulnerability when modelling was valuable for building capacity among the staff:

> The deputy principals have been good at standing up and showing lots of failures around inquiry and [a deputy principal] last year got up and *really* exposed things that hadn't worked well in her class. The inquiry had helped, but it was mixed. That was good to see examples of self-reflection and review that hadn't necessarily worked. So you are putting yourself in quite vulnerable situations. (Totara, head of faculty)

Taut (2007) argued that leaders' modelling of learning from self-evaluation set the tone in an organisation and highlighted that learning from evaluation is valued.

Educators at the schools explained that having systems and processes in place and specific time set aside were crucial in embedding self-evaluation in the school. A Kauri teacher explained: "So I think the fact that the school very systematically has invested in the model and in coaching and in giving people time to do it has probably helped to shift everybody."

Professional evaluators in any field whose work includes ECB can learn from the practice of using modelling in ECB work. To create an organisational climate favourable to evaluation, ECB professionals and organisational leaders can model what effective evaluation use looks like (Preskill & Torres, 1999). This allows others to more fully understand the power of evaluation use and helps to build a shared commitment to evaluation. ECB professionals can also, for example, make explicit their thought processes while analysing data as an effective strategy for building evaluation skills.

Tools

The practices above were assisted by the use of various tools. As mentioned previously, according to sociocultural learning theory, tools can be conceptual or practical. An example of a conceptual tool from a school in this study was a framework for thinking about self-evaluation called the Spiral of Inquiry (Timperley et al., 2014). A deputy principal explained that this conceptual tool had guided their thinking around self-evaluation: "I think it's become a bit of a model for us, a framework, not only for teachers unpacking what's going on in their classroom but for us unpacking what we are doing as a school."

The practical tools used by the schools in this study were developed internally and varied depending on the particular emphasis of the school. For example, Matai School attempted to raise the profile of using data by using the tool of visual displays around the school containing data on student achievement and attendance. The principal explained the symbolic importance: "[I]t's a signal that this is important." Totara School had developed an inquiry handbook that was given to teachers at the beginning of the year and contained readings, the programme of work, the inquiry model, and a profile of effective teaching developed by the school. The school's effective teaching profile was actually a conceptual tool, conveying statements about the practices that they believed promoted effective instruction. Rata School had a digital template to make ongoing self-evaluation of each NCEA unit easier for teachers. The template contained boxes asking teachers what went

well, what did not go well, how they knew this, and what needed to change. Teachers were asked to answer these questions using three pieces of data: student feedback; student achievement; and their own reflections. The information on this template was used to guide collaborative discussions in their departments. A participant explained that the tools were a way to build teachers' capacity and help them be more evaluative:

> We ask everyone to complete [the document] but by asking them to complete it, we are providing scaffolding at the same time. We're not asking them to produce graphs because it's done automatically. So trying to take away some of the barriers to doing a process like this by providing the template and the scaffold that goes with it … It's helping the teachers be more reflective, it's helping them in their learning areas develop better processes for reviewing their units of work and the like. (Rata, deputy principal)

Tools reflect the principles and ideas of school leaders and shape the attention, thinking, and behaviour of members of the organisation. As a participant in research by Honig and Ikemoto (2008) stated, "The tool has to carry the theory as well as the action … We had to build the tools that produced the action rather than tell people to have the action" (p. 349). In this study, Rata School's template tool carries several theories: analysis and interpretation is more important than calculating statistics and producing graphs; teachers' self-reflection is important, as is student voice; data should be interpreted collectively. In sociocultural theory, tools reflect the theory but also "tools shape our consciousness" (Lofthouse & Leat, 2013, p. 11). Kallemeyn (2014) suggested that processes and protocols for capacity building in self-evaluation are meaningful because they convey to teachers the types of data that they should notice, and how they should discuss and interpret these data, and then discuss their implications for practice and plan for improvement.

The staff in these schools emphasised that these processes and tools need to be organised and ongoing, rather than ad hoc, in order to be effective. A participant explained:

> I keep on coming back to that systematic thing, that does seem to be important … It clearly doesn't rely on me or anyone else, it just happens [in our school now] … Schools are complicated organisations, you know, so the idea that this will just happen is just too much of a romantic kind of hope. (Totara, head of faculty)

The importance of ongoing use of tools and processes for ECB has been echoed in the literature. Preskill and Torres (1999) argue that building sustainable evaluation practice in organisations requires implementing systems, processes, policies, and plans to help embed evaluation into the way the organisation works. King (2016) argues that building these structures and processes helps to minimise organisations losing capacity because of staff turnover. In a school improvement effort, Ikemoto and Honig (2010) found that the tools developed and used during a capacity-building initiative enabled teachers to engage with research-based ideas in ways that shaped their thinking and actions. This was particularly true when the tools: were modelled; included knowledge from both research and practice; were linked to practitioners' local situations; and were adapted over time.

Conclusion

This study argues that the use of learning theory and specifically, sociocultural theory, contributes to the scholarly conversation about ECB. The way educators in this study conducted ECB provides a model for how organisations might increase their capacity to conduct and use evaluation in their everyday activities.

One way that sociocultural theory contributes to the conversation about ECB is by drawing attention to the learning process. Through a sociocultural lens, learning is not conceived of as merely the acquisition of a set of skills and a body of knowledge, but instead as a process in which individuals engage with other people and various tools to make sense of information and construct new knowledge (Vygotsky, 1978). It also emphasises the particular context where learning occurs and the fact that the learner is connected with, and involved in, this learning context. In this way, sociocultural theory is helpful for considering the relationship between individual practitioner learning and organisational support for ECB. Given the limitations of examining only four schools, the current study is best understood as exploratory. However, it does raise the possibility that both ECB facilitators and organisational leaders might consider drawing on learning theories when designing their efforts to encourage self-evaluation. Even a limited understanding of these theories could impact our ECB efforts because this enables us to design activities that incorporate our understanding of how people learn.

When designing activities and processes, both ECB facilitators and organisational leaders might want to consider if their efforts reflect the practices that sociocultural learning theory suggests are most effective for learning. For example, an organisation could support sense

making in a variety of ways, such as allowing practitioners to have time and space for collaboration. They might also want to consider the tools that support the learning process and how these tools prompt users to engage in new ways of thinking and reflection. ECB facilitators might also assist with tool use by modelling how to use a tool or engage in new work practices. Modelling can be particularly valuable when modellers use metacognitive strategies that make their thinking visible so that practitioners understand not just *what* to do but *why*. Modelling can help practitioners determine how they could apply new practices and tools in their day-to-day work.

This article also contributes to the largely atheoretical body of research on ECB. Even if they do not employ strategies from sociocultural theory, evaluators may want to pay greater attention to the concepts implied in the way they conduct ECB. It can be valuable to unpack the assumptions underlying their capacity-building activities and to make them more explicit. Much as evaluators often ask programme stakeholders to unearth the tacit theories underlying their programme activities, it might be beneficial to pay greater attention to the theories underlying the way we do ECB.

Chapter Nine

Introduction to Article Four

Article Four returns to the tensions of Article One, but moves from the micro to the macro. In this article I ask what the processes, interactions, and tensions experienced by individual teachers signify and how theory from the emerging field of Evaluative Thinking could contribute.

This article is a more conceptual piece than the other three. It was developed when I reflected on the tensions felt by teachers that were uncovered in Article One and the interventions that were put in place described in Article Two. I was noticing trends in the data from my participants and wanted to explore these further. In thinking about the tensions teachers were feeling around data use reflected in the previous articles, I realised these often revolved around values discussions, emotional reactions, and the realisation that these decisions were more complex than it may seem initially. I realised these lived experiences were in contrast to the way that DBDM is often discussed in the literature. In this article, I explore the tensions in my data and problematise the DBDM literature to examine what theory is embedded in it. I draw on the theory behind Evaluative Thinking and explore how the DBDM field could learn from the new discussions happening in the evaluation field about evaluative thinking and complexity, that is, how the conversations in ET could contribute to the understanding of DBDM in a way that isn't based on a purely rational model of decision making.

Article Four draws extensively on the concept of Evaluative Thinking, a concept that was very briefly mentioned in Article Three. Article Four also briefly discusses ECB, the main topic of Article Three. Those connections between the topics covered in these articles might lead the reader to think that this sequence was carefully planned before I wrote these, but I didn't come to draw on the theory of Evaluative Thinking (ET) easily. Perhaps one reason for this is that ET is a new area of scholarship that is still being defined and refined with new articles in this field unpacking the construct of ET. The concept of Evaluative Thinking has been introduced in educational evaluation (Clinton & Dawson, 2018; Earl & Timperley, 2015; McFadden & Williams, 2020) but most of this literature has not drawn on the new work being done in the field of evaluation to explicitly unpack the construct of ET.

I first wrote this as a 2,000-word abstract for a full paper presentation at the American Education Research Association. The paper was accepted, but I was not able to travel to present at the conference due to health issues. When I expanded the abstract to a full journal article trying develop these reflections on my data, I looked for the right theory to make sense of it. I wrote multiple versions of it, drawing on different theories before hitting on the developments in the field of evaluative thinking. I provide more details of this process of searching for the correct theoretical construct in the Discussion in Chapter 10.

I want this article to be published in an overseas journal because in countries such as the US and England conceptions of data use in education have been heavily influenced by the rationality of the early scholarship and policy in DBDM. I want it to be published in an education journal because I view a primary contribution of this paper is the way that theory from the field of evaluation can contribute to the field of education.

Article Four

Beyond the Rational Model of Decision Making: A Conversation with Evaluative Thinking

Introduction

As educators in many different countries are increasingly urged to use data to make decisions, the field of DBDM has become its own area of scholarship. The early discourse in the field has strongly influenced the dominant way of thinking in DBDM, highlighting a linear, rational model of decision making (Datnow & Park, 2018; Vanlommel, Van Gasse, Vanhoof, & Petegem, 2018), with policymakers and researchers expecting educational decision-making to become more rational (Lai & Schildkamp, 2013; Mandinach & Schildkamp, 2020). In the early days, it heavily relied on standardised quantitative data used in a summative manner (C. Brown, Schildkamp, & Hubers, 2017), and emphasised the importance of the actual data, with policymakers and the literature implying that simply having accurate data available would lead to improved practice (Datnow et al., 2012; Mausethagen et al., 2018; Spillane, 2012).

Initially, much of the literature on DBDM focused on advocating for data use or providing how-to guides rather than analysing what happens when educators use data in their practice. More recently there has been an increase in research on data use processes that has attempted to understand how people actually engage with data in their everyday work and how that relates to instructional change and organisational learning (Coburn & Turner, 2011). More recent literature has a greater emphasis on data analysis and interpretation, particularly collaborative sensemaking. Researchers have begun to acknowledge that the process of data use is more complex, contextual, and less rational than the linear model might suggest (eg., Bertrand & Marsh, 2015; Datnow & Park, 2018; Vanlommel, Van Gasse, Vanhoof, & Van Petegem, 2017). While this understanding is developing, the field has recently still been described as neo-behaviourist, cognitivist (Mandinach & Schildkamp, 2020), linear, and rational (Datnow & Park, 2018). The literature has increasingly attended to teachers' interpretation and thinking processes, however, the current frameworks guiding data use do not fully capture decision making in complex contexts such as education (Vanlommel, 2018). Despite widespread attempts to encourage data use, the field is a relatively new area of scholarship that remains undertheorized. This article attempts to contribute to the

conversation and show that the dominant discourse of rationality has often not captured the reality that teachers experience with data use.

In a recently-completed study on the data use processes as part of school self-evaluation in New Zealand high schools, I attempted to gain a greater understanding of teachers' perceptions of the processes of using data to make decisions about interventions for students who were at risk of not succeeding. As I reflected on my data from the larger study, I noticed a common thread was that teachers and school leaders often feel tensions when enacting data use processes (Dyson, 2020b). As a new field in education, our conceptual understanding of DBDM continues to evolve and has begun to include a recognition of teachers' interpretive processes, however, I found that the tensions experienced by teachers in my study were not expressed in the conceptual tools of DBDM. I looked at related fields of scholarship and realised there are conversations occurring around the closely-related concept of evaluative thinking (ET) in the field of evaluation that could help expand our understanding of DBDM. The evaluation field is grappling with some issues that might be relevant to the conversations in school data use so we can understand DBDM in a way that isn't based on a purely rational model of decision making. The ET body of literature is distinct from, and rarely overlaps with, the data use in education literature. This article is intended as an invitation to expand the conversation between these fields about these processes.

In this article I explore how we can draw on some of the debates and contestations from evaluation to inform our thinking of DBDM. This paper will outline some of the key underpinnings of the DBDM field, then briefly trace this field's evolution before an overview of the concept of evaluative thinking. Then I will explore tensions that teachers feel around data use by drawing on illustrative examples of my data from the larger study and an examination of the literature, discussing the intersection of DBDM and ET in four distinct but related ways. I will explore two areas where DBDM can possibly learn from ET: the treatment of complexity and values; one area where ET is undertheorized: collaborative social practice; and one area that is undertheorized in both fields: the role of emotion in these processes. By doing so, this paper hopes to continue the advancement of our conceptual understanding of DBDM and consider how the conversations around ET can inform DBDM.

The Evolution of DBDM

Educators in many different countries are increasingly urged to use data to make decisions. The theory behind this recommendation is that collecting and analysing data will improve

the quality of decisions about teaching and learning practice (Earl & Seashore Louis, 2013; Marsh, 2012; Timperley et al., 2007; Vanlommel, Van Gasse, Vanhoof, & Van Petegem, 2017). Some policymakers even argue that the *only* way to increase student achievement is for school personnel to make decisions based on data (Bertrand & Marsh, 2015; Schildkamp & Kuiper, 2010). The processes in which teachers and principals systematically collect and analyse data in support of a range of decisions to help improve the success of students and schools are typically referred to as data-driven decision making (DDDM), or more recently, data-based decision making (DBDM) or "data use". Data can be defined broadly as any information that is systematically collected and organised from relevant stakeholders (e.g., parents, teachers, and students) to represent some aspect of schools (M. Lai & Schildkamp, 2013). This could include student assessment and achievement data, attendance data, student demographic information, classroom observations, student work samples, or survey results from students, teachers and parents.

The dominant conceptual understandings of DBDM have been strongly influenced by the policy and early discourse in the field. In the beginning of the DBDM movement, it predominantly focused on summative, standardised assessment data to determine if predetermined goals and benchmarks had been achieved, and, if necessary, to adapt classroom instruction in response (Brown et al., 2017). In the beginning, much of the literature was prescriptive rather than descriptive, advocating for data use or providing how-to guides rather than analysing what actually happens when educators use data in their practice. Policymakers and researchers anticipated that educational decision making would become more rational (Schildkamp et al., 2013), starting from the hypothesis that rational processes lead to better judgement than intuitive processes (Vanlommel, 2018). Early literature had an "irrational exuberance" for making decisions based on data (Hess, 2009) and some literature and policy about educators' data use implied that simply having accurate data available will lead to improved practice (Datnow et al., 2012; Mausethagen et al., 2018; Spillane, 2012). When the field became its own subject of inquiry, research focused on the characteristics of data use interventions and the factors that enable and hinder data use with the aim of enhancing DBDM in education. The data use literature "often aimed to identify 'best practices', whilst under-conceptualising and under-theorising 'data' and 'data use'" (Mausethagen et al., 2018, p. 39). Early work often had a narrow definition of data focused on quantitative indicators of students' achievement, based on the assumption that the quality of educational decisions would increase if they were based on objective measures (Vanlommel, 2018). Over years, the field evolved from a narrow focus on assessment data

and accountability to a broader focus on multiple data sources and a focus on instructional and school development (Brown et al., 2017)**.**

Perhaps because of these underpinnings, scholars contend that DBDM predominantly starts from a rational model of decision-making that implies that teachers use rational, straightforward processes (Datnow & Park, 2018; Datnow et al., 2012; Vanlommel, Van Gasse, Vanhoof, Van Petegem et al., 2017). The data use model suggests a "rational-instrumental [process], i.e. the expectation that teachers will use the provided data and development will therefore take place" (Mausethagen et al., 2018, p. 39). The model of DBDM attempts to make education practice more rational, more certain, more predictable, and less ambiguous (Dahler-Larsen, 2012; Schwandt, 2008; Vanlommel, Van Gasse et al., 2017). In contrast to rational DBDM processes, the literature implies that teachers would normally make important decisions based on hunches (Datnow et al., 2017) and that their decision-making processes are thought to be "plagued by emotions, inconsistency, subjectivity, personal preferences, bias and reliance on intuition, habit, myth and so forth" (Schwandt, 2005, p. 317).

The rational model is based on several assumptions about the process of using data to put in place interventions for students at risk of not achieving. A deliberately over-simplified summary of this theory is that data will identify which students need additional assistance, data will be used to determine the best intervention for said student, and data will unambiguously determine if the intervention worked; the model is seductive in its simplicity (Dahler-Larsen, 2012). This model is also based on the assumption that teachers can easily access all data that they need, that they have knowledge of all alternatives and that the consequences of a decision are known and consistent (Vanlommel, 2018). The rational model approach assumes that the scale of the intervention is proportional to the scale of the effect (Williams & Hummelbrunner, 2011). Attempts to influence human behaviour, in the form of programmes, interventions, or policies, will search for silver bullet solutions, implying that there is one "right" answer, one right intervention (Dahler-Larsen, 2012). Therefore, human behaviour can be shaped by applying the right formulas or recipes, which gives the appearance of certainty (Horn, 2018).

More recently there has been an increase in *research on data use* processes that has attempted to understand how teachers and school leaders actually engage with data in their practice and how that relates to instructional change and organisational learning (Coburn &

Turner, 2011). An extensive body of research has studied the rational processes of data-based decision making described above and recently the field has been evolving (Vanlommel, 2018). An emerging literature has begun to acknowledge that in actual practice, data use is less rational and straightforward than the way it is usually conceived in the literature and in policy, primarily by recognising that data use is an act of interpretation and sensemaking (Bertrand & Marsh, 2015; Datnow et al., 2012). Sensemaking theory asserts that the meaning of information or events is not given; individuals and groups must actively construct understandings and interpretations, and create meanings for their experiences (Coburn & Turner, 2011; Datnow et al., 2012; Spillane & Miele, 2007; Weick, Sutcliffe, & Obstfeld, 2005).

While DBDM has emphasised the primacy of the actual data, there is an increasing body of work on the importance of teachers' *interpretation* of data (eg., Bertrand & Marsh, 2015; Lai & Schildkamp, 2013; Vanlommel, 2018). There is nascent recognition that decisions can never be completely driven by data; that the process is more complex than previously framed; and that teachers filter data through their own experiences and shaped by their school context (Bertrand & Marsh, 2015; Coburn & Turner, 2011; Datnow et al., 2012, 2017). Research that addresses interpretation processes have centred on teachers' cognitive strategies and practices in using evidence to inform their teaching (Bertrand & Marsh, 2015; Vanlommel & Schildkamp, 2019). Data use research has focused on technical aspects of teachers' 'data literacy' (Wayman & Jimerson, 2014), and touched on teachers' knowledge of how to interpret and respond to data (Marsh, 2012), teachers' attitudes and motivation to use data (Vanhoof, Vanlommel, Thijs, & Vanderlocht, 2014), and teachers' approaches to teaching (Vanlommel et al., 2018). Research has also begun to touch on teacher beliefs *about* data use, but teachers' belief systems have not frequently been addressed in data use research or educational reforms (Datnow & Hubbard, 2016). The recognition that interpretation and sensemaking are involved in DBDM is a step in the right direction toward taking into account how people actually make decisions, because in the literature, "data use is often considered to be a straightforward process, without sufficient attention to the complexity of the sensemaking processes and teacher beliefs that influence decisions" (Vanlommel & Schildkamp, 2019, p. 23). This field is just beginning to develop a theory of teacher decision-making processes that are based on teacher judgement and not based solely on rationality (Vanlommel, 2018).

This recognition that data use involves interpretation and sensemaking processes is one of the biggest developments in the evolution of the field of data use. There is increasing recognition that data do not speak for themselves, but are mediated by teachers' worldviews. While the field of DBDM has evolved, the state of the field is still described as neo-behaviourist, linear, rational, cognitivist (Mandinach & Schildkamp, 2020; Vanlommel, 2018). For example, Mandinach and Schildkamp (2020) argue that the field needs to move away from its neo-behaviourist and cognitivist perspective on data use, but then do not explain this concept. What would it mean to expand the conception of data use processes beyond merely the cognitive? With an emerging recognition of data interpretation and thinking processes, the data use literature focuses on cognitive activities and building these capacities. Yet the cognitive is only one domain that should be understood in the evolving understanding of the practice of DBDM. Before considering these questions more fully, we turn to a brief overview of the concept of evaluative thinking and then consider how the conversations around ET can inform these questions.

Evaluation and Evaluative Thinking

Evaluation is a contested field with a history of vigorous debates about many issues, including what methods to use, how to judge the quality of programmes, and even the very definition of evaluation (Patton, 2018b). But at its core, the field of evaluation involves conducting systematic, data-based inquiries about how, and how well, interventions work. Interventions can be any effort, programme, project, initiative, policy, organisation or activity aimed at bringing about change.

Like the field of DBDM, for much of the history of evaluation, rationalism has been the dominant approach. Schwandt (2019, p. 323) argues that evaluation "is largely characterized by a need for order—to define, appraise, and rationalize the world so it will be controllable, predictable, and understandable." Rationalism is modern society's concept that human beings can control the natural and social world. Some authors argue that evaluation tries to make organisations (and the social world more broadly) more predictable, controllable, efficient, reliable, objective, and less complex (e.g., Dahler-Larsen, 2012; Schwandt, 2019). Discourse in evaluation, and discourse about social programmes more generally, has been dominated for decades by underlying positivist epistemologies and mechanistic ontologies (Archibald, Sharrock, et al., 2018). Social initiatives rest on a certainty that:

> encouraged us to believe that: all systems can be treated like independent mechanical

systems; the future can be predicted accurately from analysing the past; relevant data both exists and can be measured and analysed; problems can be reduced to largely independent parts (a consequence of assuming linear, proportional interactions); the parts of a system can be idealized and regarded as identical (i.e., assume all people in a class will react in the same way to the same stimuli); those qualities of the constituent parts that are more subjective and less tangible can be discounted. (Archibald, Sharrock, et al., 2018, p. 77)

In their histories, both DBDM and evaluation have a similar dominant ethos of rationality, control, predictability, objectivity, accountability, and an emphasis on methods and procedures. However, Dahler-Larsen (2012, p. 16) argues that the field of evaluation has recently had to contend with the *inadequacies* of rationalism to explain interventions: "many good initiatives fail, there is limited knowledge about why this happens, the goals shift, there is disagreement about political goals, a simple distinction between means and ends cannot be maintained etc.". This has recently led to changes in the field of evaluation characterised by new ways of thinking that oppose this rationalising tendency and criticise traditional practices and approaches to evaluation (Schwandt, 2019).

One of the recent emphases in evaluation practice has been greater prominence of the concept of *evaluative thinking* (ET). A small corner of the evaluation field has long de-emphasised evaluation as a technical activity and highlighted the importance of thinking processes, most notably the authors Patton, House, Schwandt and Dahler-Larsen. For example, Patton (2012) argued, "evaluation as a field has become methodologically manic-obsessive. Too many of us, and those who commission us, think it's all about methods. It's not. It's all about reasoning" (p. 105).

Recently, evaluative thinking is increasingly recognised as a key construct in the field of evaluation (Buckley, Archibald, Hargraves, & Trochim, 2015; Patton, 2018a). In the last couple of years the field of evaluation has begun to unpack the construct of ET. Schwandt (2018) suggests that all definitions of ET highlight a commitment to analytical reflection on one's assumptions and claims, combined with continuous learning and an openness to changing one's views in light of evidence and reasoned arguments. One of the most widely-cited definitions of ET comes from Buckley and colleagues (2015):

Evaluative thinking is critical thinking applied in the context of evaluation, motivated by an attitude of inquisitiveness and a belief in the value of evidence, that involves

identifying assumptions, posing thoughtful questions, pursuing deeper understanding through reflection and perspective taking, and making informed decisions in preparation for action. (p. 378)

In this definition--the dominant way of thinking about ET--this process is portrayed as an individual activity requiring cognitive capacities for analytical and critical thinking along with certain intellectual dispositions (Schwandt, 2018). Several authors suggest that this definition, while not wrong, is too narrow and they have expanded on this underpinning definition (Patton, 2018a; Schwandt, 2018). For example, Patton (2018) suggests that rigorous evaluative thinking combines critical thinking, creative thinking, inferential thinking, and practical thinking. Schwandt (2018) made a new contribution to the understanding of ET by emphasising the collaborative, communal aspects of sense making practices, which will be discussed further later.

Recent conversations around the construct of ET have also focused on the role of values. A recent conceptual framework derived from a literature review of ET includes the cognitive aspects of ET, but also identifies the importance of *values* as a key aspect of ET (Vo, Schreiber, & Martin, 2018). Patton (2018) suggests that ET involves being aware of and articulating values and ethical considerations. Wehipeihana and McKegg (2018) argue that the ET literature needs to pay more attention to the role of values in the process of making judgements. "The values embedded in everyday processes of evaluative thinking are not often unpacked and examined in detail" (Wehipeihana & McKegg, 2018, p. 95). They go on to argue for the need to link values and culture and to acknowledge that cultures beyond the dominant Western one have different values and different ways of valuing. "Much of our understanding about evaluative thinking and reasoning at this point of time appears relatively 'culturally free,' that is, we have barely considered other ways of valuing, of deliberating, and reaching agreement on what is valued" (Wehipeihana & McKegg, 2018, p. 104).

Another recent conversation around the construct of ET is the concept of complexity. Archibald and colleagues (2018) link the concepts of ET and complexity and argue for the importance of building capacity in ET to deal with complexity. The wider field of evaluation is increasingly developing an understanding of complexity science (eg., Gates, 2016; Patton, 2010; Williams & Hummelbrunner, 2011). As part of unpacking their framework for understanding ET, Vo and colleagues (2018) highlight the complexity of contexts and

organisations, emphasising that ET underscores the need to "navigate uncertainty, ambiguity, and complexity" (p. 44).

This article argues that these recent conversations in the evaluation field might be relevant to the conversations in school data use. DBDM and evaluation have a similar history of being dominated by the rational approach, however, evaluation has recently been dealing with the limitations of this approach. The remainder of the article will explore how some of the debates and contestations from evaluation and ET can inform the thinking around DBDM. One of the interesting areas of exploration is the way that scholars working on unpacking the construct of ET have been expanding the definition beyond individual cognitive activity. In the following sections I will explore the tensions teachers feel around data use by drawing on illustrative examples of my data from the larger study and an examination of the literature. By doing so, this paper hopes to push forward our thinking about DBDM. I first briefly describe the origins of this data.

Origins of the Illustrative Data

The illustrative data is drawn from a larger study that attempted to better understand data use process by considering the perspectives of teachers whose schools had recently implemented new data tools to track and monitor students and had implemented interventions in response to this data. This study was informed by realist research and evaluation principles and methods (Emmel et al., 2018; Pawson & Tilley, 1997). This way of thinking is oriented towards understanding how, why, for whom, and under what conditions initiatives produce results, making it a well-suited approach for "opening the black box" of data use. Coburn and Turner (2011) emphasise the importance of examining the pathways or mechanisms between intervention activities and outcomes to understand more about data use and the nature of change that occurs within schools. Describing these mechanisms or underlying processes entails making explicit the individual or social psychological responses to data use processes. Because of a focus on participants' unobservable thought processes, individual interviews were emphasised.

Data collection began at one high school where six heads of department were each interviewed three times in open-ended focus groups. The results of these were used to inform questions for one-hour individual "realist interviews" (Manzano, 2016; Pawson, 2013) with 13 educators at four high schools. Within each school, three people were usually interviewed: a principal or assistant principal, and two staff who were not part of the school's

senior leadership team. Many were heads of department, and all were selected because they were knowledgeable about their school's self-evaluation processes. Documents from each school were also analysed to gain a more complete understanding of the schools' processes. The four schools were each given a pseudonym based on a New Zealand native tree.

The context of this study was high schools in New Zealand. Practices of data use are more established in the New Zealand context than in some other countries because of the well-developed culture of data-informed teacher inquiry and a system of school self-evaluation (Nusche et al., 2012; Sinnema & Aitken, 2011). Each New Zealand school is expected to perform ongoing evaluation for improvement that generally consists of noticing data, investigating further, collaboratively making sense of data, taking action, and then monitoring and evaluating the impact of the action taken (ERO, 2015).

Teachers' Tensions in Data Use and Contributions from Evaluative Thinking

In this section I will explore the ways that some of the conversations occurring around evaluative thinking can inform the conversations in DBDM by drawing on illustrative examples of my data from the larger study and an examination of the literature. I will explore two areas where DBDM can possibly learn from ET, one area where DBDM is more developed than ET, and one area that is undertheorized in both fields. The data focuses on the schools' use of student achievement data to create interventions for students who were at risk of not achieving.

No silver bullet

Participants in my study learned that the process of implementing and evaluating interventions in response to student data involves complexity. A deputy principal explained the reality that, even with abundant student data, these are complex issues with no simple solutions:

> What we have found now is we are getting very good at identifying these students. We can identify students who may possibly be at risk but what we are struggling to do is to make a genuine significant impact. So what is the silver bullet that actually gets this kid over the line? We've got spreadsheet after spreadsheet where we analyse. I can crunch all the data and give you the stats, but what's the actual soft story behind it… So we've got much better at doing that. *But the reality is there is no silver bullet. There are so many variables…We are not dealing with something that is*

static and reacts in the same way every time. It has to be a really personalised,
individualised approach…. Things are improving, but it's not this nice linear
progression of year on year it goes up. (Totara School, assistant principal)

(Emphasis is mine)

This quote encapsulates the reality of using data in education. The deputy principal conveys the following concepts: Individual people don't all react the same way to interventions. The overall impact of interventions is mixed. There is not consistent, linear progression. There are no silver bullets, or solutions that will work for everyone. Education programmes work under certain conditions, in certain contexts, for certain students, in certain respects, over certain lengths of time. Taken together, these concepts illustrate the complexity of implementing and evaluating interventions in response to student achievement data.

This is in contrast to much of the discourse around data use in education that invites a stance of certainty (Horn, 2018). As discussed previously, the literature implies that DBDM attempts to over-simplify complex, ambiguous education practice (Selwyn, 2014). The rational stance values "strategies that reduce the interpretive character of [practitioners'] decision-making and enhance clarity, consistency, and order" (Schwandt, 2005, p. 317). This complexity is also the reason there is no silver bullet solution that the participant above refers to. The silver bullet way of thinking about interventions is also built on the assumption that interventions are causes that will necessarily generate effects, and that the scale of this effect will be proportional to the scale of the intervention. The existence of a silver bullet would also mean there is a solution that works for everyone. However, "practices are concerned with the particular (rather than the general) precisely because, as the example above illustrates, they are about taking the right action in consideration of *this* situation, *this* person, at *this* time and place, in *this* set of circumstances" (Schwandt, 2005, p. 323).

The field of evaluation is increasingly recognising the concept of complexity and how it can inform their work (Schwandt, 2019). One of the pioneers of the concepts behind evaluative thinking, Thomas Schwandt, noted that an evaluator's job often includes the need to explain "the complexity of social systems and the limitations on our ability to predict, plan, and control their behaviour" (Schwandt, 2008, p. 148). There is a growing awareness that "many aspects of economic and social development are complex, unpredictable, and ultimately uncontrollable" (Archibald et al., 2018, p. 74). A paradigm shift is taking place away from predominantly linear and simple models of change to ones that are more complex, reflective,

and responsive (Archibald, Sharrock, et al., 2018). Archibald and colleagues (2018) link the concepts of complexity and ET and argue for the importance of building capacity in ET to deal with complexity. In other work to define evaluative thinking, the recent framework by Vo and colleagues (2018) highlights the need to "navigate uncertainty, ambiguity, and complexity" (p. 44).

The conversation around DBDM could benefit from a move from rational models to greater recognition of complexity. Vanlommel and colleagues (2017) explain that rational strategies for decision making are suited to solving well-structured problems that allow for fixed procedures, but that most educational problems are complex, so the rational frameworks that are prevalent in educational research on decision-making may not suit these problems. "Data use in education should be viewed within a complex and integrated framework, rather than mainly linearly" (Mausethagen et al., 2018, p. 46). Shifting this way of thinking allows educators to rely on complexity-aware data use practices to collaborate, reflect, and make adaptations.

Value conflicts

The schools in my study used student achievement data to create interventions for students who were at risk of not achieving. The interventions included remediation classes, conducted at lunchtime or after school, and homework and study groups facilitated by a resource person. One of the interventions offered in all schools occurred at the end of the school year and gave students the chance to be reassessed in a previously unsuccessful standard or assessed in a new standard. Different educators responded differently to this end-of-year opportunity. Their views reflect different value positions. A sample of teachers' views are presented:

> The idea was once they've got level 1 it is easier for them to get into other classes… [I wonder] what was the level of learning that took place or was it really just here is something, write your name on it and copy from the board or something. They are actually not learning anything apart from I can get seven credits in one day from putting my name on a bit of paper and doing something I didn't know about. I have philosophical objections to that. (Matai School, teacher)

> Sometimes you get positive responses [from teachers] and sometimes you definitely get negative responses. 'He hasn't done anything all year, why should I put the effort in now blah, blah, blah.' I feel that is the wrong way to look at things, you know, if

the kid has a chance to succeed it doesn't really matter which part of the year they want to do it in and, you know, our role is essentially helping these kids out. (Rata School, teacher)

The vision for me was if we have 15 students through and we get 12 of them an NCEA Level 1 certificate and for one of those students, even if it's only one, who turns around and says, wow, if it wasn't for that school I wouldn't have achieved this and then re-engages more readily in learning for next year, then for me that's a success. The 12 boys is a success in itself but if that one boy re-engages with that programme because he feels more supported, that's a wild success. (Rata School, assistant principal)

These quotes reflect different value positions or competing conceptions of what is the best way to achieve the ultimate goal of student learning and student success. The first teacher quoted above is concerned about high-quality learning, being ethical, and avoiding what he called "cheap" credits. The other two argue for the primacy of the value of staying in school and continuing to learn. Their various positions are whether to prioritise keeping students in school or to uphold the quality of the learning. The different teachers hold different values and express different priorities.

Traditionally, the field of DBDM has not emphasised the role of values in decision-making. Higham and Booth (2018) argue that although values influence educators' decisions, the role of values in educational decision-making has been downplayed and under-theorised for decades. The argument of policy makers early on in the DBDM field that the mere existence of data would allow for sound decisions does not encompass the concept of value judgements. There has been little recognition that teachers in similar situations make decisions in different ways because they apply different values (Vanlommel, 2018), and little recognition of the need to articulate those values.

One of the reasons for the lack of emphasis on values is that the field of DBDM has viewed data and data-driven decisions as neutral. The literature and policy predominantly treat data and data-driven decisions as neutral, objective, impartial, non-political, and scientific (Booher-Jennings, 2005; Selwyn, 2014; Spillane, 2012). Many social scientists and policy makers have been trained that data is objective and that data, statistics, and evidence remove the politics from decisions. In reality, there is an ethical worldview embedded in each step of data use (what we notice, data collection, data analysis, interpretation). Data are not neutral

but are "loaded with values, interests and assumptions" (Selwyn, 2014, p. 14). Data use is often positioned as removing politics from school decision-making (eg., Schildkamp & Kuiper, 2010), but there is evidence that it is deeply intertwined with data use processes (Coburn & Turner, 2011).

The current conversations around evaluative thinking has acknowledged that an important part of making good decisions involves making value judgements (Karaali, 2011). In their everyday practices, when teachers make decisions about their response to data, they face the question of *what should we do?* (Schwandt, 2018). This question involves multiple people weighing options and consequences, using their values. Figuring out *what we should do* is about appraising the value of various courses of action. Some scholars unpacking the construct of ET have recently highlighted that values and value judgements are key elements of evaluative thinking (Schwandt, 2018; Vo et al., 2018), with Vo and colleagues (2018) arguing that evaluative thinking is the process in which a person uses data and evidence to make defensible, reasoned value judgements. When people make judgements, they use moral criteria to evaluate strategies, programmes, initiatives, interventions. They often use language reflecting value judgements such as that's good, he's doing a good job, the programme is working, they're really getting better, and words like effective, quality, good, bad, better, improving, etc. Knowledge about the effectiveness of particular interventions is not on its own a sufficient basis for decisions about educational action. There is always the question as to whether particular interventions are desirable; what counts as ''effective'' crucially depends on judgements about what is educationally desirable (Biesta, 2007).

The earlier quotes from teachers suggest the ambiguity of these moral decisions. Educators often negotiate between contradictory imperatives and conflicting ethical principles (Hammersley-Fletcher, 2013). Data use in education involves choices and interpretations. "Practitioners are *always* facing contingencies, multiple demands on time and resources, competing conceptions of what is right to do, and so on as they make decisions" (Schwandt, 2005, p. 328). This is why Schwandt has referred to teaching practice as "rough ground" that cannot be smoothed with the mere existence of data, even good data, "because different ideas of what constitutes a good practice and a good practitioner always compete for attention and because the moral, the political and the instrumental are always intertwined" (Schwandt & Dahler-Larsen, 2006, p. 504). Higham and Booth (2018) advocate making explicit the connections between values and actions, and argue that doing so helps schools deal with pressures between competing sets of values. Educators must make judgements

about the appropriateness of their actions in relation to a range of technical, political, moral, and ethical concerns (Schwandt, 2008).

To face these kinds of situations and act appropriately, Schwandt (2008) emphasises the importance of educators learning practical reasoning or the capacity for discretionary judgement. Sanderson (2006) argues for an expanded sense of rationality to include consideration of values and ethical-moral choices. "Only this kind of practical rationality can guide us toward appropriate action in complex and ambiguous social contexts" (Schwandt, 2008, p. 149).

Collaborative Social Practice

Here, we turn the tables and explore an area where DBDM has theorised more than ET. In both DBDM and ET, the literature focuses on individual cognitive activities and regards the individual as the locus of attention (Schwandt, 2018). However, the literature in DBDM has had more conversations and developed more theory than ET about the understanding that these processes are not merely an individual cognitive activity, but are a collaborative social practice. The dominant way of thinking about ET is that it is a solitary activity requiring cognitive capacities for analytical and critical thinking. This conception was expanded recently when Schwandt (2018) made a "major contribution to understanding evaluative thinking" by arguing for understanding it as a collaborative social process (Vo & Archibald, 2018, p. 141). Schwandt (2018) argued that thinking and reasoning are essentially interactive social phenomena.

The recency of this argument sits in contrast to the field of DBDM that has an established literature base theorising data use as collaborative practice (see Bertrand & Marsh, 2015a; Coburn & Turner, 2012; Lai & McNaughton, 2016; Mandinach & Schildkamp, 2020; Spillane & Miele, 2007; Vanlommel, Van Gasse, Vanhoof, Van Petegem, et al., 2017; Wayman & Jimerson, 2014). This literature often draws on sensemaking theory (Weick et al., 2005). Sensemaking is understood to be not solely an individual activity but a collective process based in social interaction. Sensemaking not only highlights individual teachers' data interpretations, but also privileges the role that social learning plays. Authors argue that for effective data use to occur, social interaction and collaboration are needed to co-construct meaning and decide what is worth doing (Spillane & Miele, 2007; Wayman & Jimerson, 2014). Using the framework of sensemaking the DBDM field has recognised that data use is influenced by teacher interpretation and social interactions.

While the field of DBDM has a longer history and wider base of literature about collaborative social practice, it could learn from ET's conception of this practice in its intersection with the previous concept of values. In ET collective sensemaking is linked to the acts of deliberation and negotiation of different values and multiple perspectives. Schwandt (2018) explains that the process of deliberating claims of facts and values can be thought of an individual- cognitive process, but it actually often occurs as an interactive process among groups of people with different viewpoints trying to establish common ground on which to evaluate an intervention. This has roots in Ernie House's writing on deliberative democratic evaluation (House & Howe, 1999) arguing that a central function of evaluation in a democracy is to give voice to various stakeholders and support dialogue and deliberation. It is echoed more recently in Vo and Archibald's (2018) reflection on the central role that evaluative thinking can play in strengthening democratic societies. This emphasis means that "there is a need to focus on and promote inclusion, dialogue, and deliberation" (Vo & Archibald, 2018, p. 141).

In the DBDM literature, collaborative social practice is conceived differently than this. It typically focuses on data analysis and interpretation to identify student learning problems, possible causes of these problems, and possible action plans (Schildkamp, 2019). This is in contrast to the process in ET that purposely surfaces and highlights different values. This is perhaps due to different processes in the two fields; a typical evaluation might entail a professional evaluator collecting and analysing data but then various stakeholders deliberating the implications, whereas DBDM may involve educators analysing the data themselves. The data from my larger study reflected this. Collaborative data use in those schools focused on learning from other teachers' individual inquiries, reflecting on units of learning to see what worked and what didn't, and using tracking and monitoring software to identify students at risk of not achieving. None of the teachers in my study mentioned a process of deliberating different value positions in order to make decisions.

Emotion

When participants in this study engaged in data use processes, they often felt tensions, which often manifested in the affective domain. In this study it became apparent that there was an emotional dimension to educators' data use, and the study attempted to unpack these tensions. Teachers discussed the emotional reactions they sometimes experienced. These

reactions particularly occurred when teachers were involved in evaluating interventions in which they had personally participated. A teacher described his emotional reaction to data:

> At the start of the year we were looking very much at the NCEA data and then from that data really thinking hard about where do we need to improve. There was clear pressure I think to do that. There was an expectation really that you are very much responsible for those results. We were looking at patterns of poor achievement, we were looking at gender, ethnicity, year levels, achievement standards and then trying to put in place an intervention really around how can we address this. But it kind of made you feel quite fragile. It is around emotion I suppose and the affective side of all this, how does all that make you feel, how open are you to really genuinely wanting to do that and then taking the brave step to investigate. (Totara School, teacher)

This quote displays the vulnerability and emotion teachers may feel at examining their practice and their contexts, and the commitment to this process required in self-evaluation.

Emotional reactions to data use processes are rarely included in the DBDM literature. The role of emotion in decision making is generally only mentioned to admonish educators not to make decisions based on emotion (e.g., Vanlommel & Schildkamp, 2019). The emotional effect the actual data use process may instigate is not generally considered. When teachers are urged to inquire into their own practice, and to self-evaluate, we don't talk about how they might *feel* about that process, especially if things aren't going well. How does it feel if something they try doesn't succeed? Data use in education focuses on rational and cognitive models; however, being subjected to measures and targets also has an emotional aspect to it. "Thus, responses to the flow of performance information can engender individual feelings of pride, guilt, shame and envy" (Ball, 2003, p. 221). In the field of evaluation the literature has acknowledged the possibility of evaluation anxiety (Donaldson, Gooler, & Scriven, 2002; Dyson, 2018), but this is not widely explored.

Only a small number of studies in the data use literature have acknowledged the uncomfortable feelings that may arise around data use. In a survey of over 1700 teachers, Dunn and colleagues (2012) found that teachers' beliefs about their own ability to effectively analyse and interpret student data, combined with anxiety about the process, limited their ability to use data effectively. Marsh's (2012) review of interventions designed to support data use found that if teachers were afraid that their personal identity would be exposed, they

disengaged from the process; even when assurances of anonymity were given, teachers expressed concern. They were often afraid that district leaders would use the data for evaluative purposes, and as a result, they did not trust or feel comfortable with data discussions. Bocala and Boudett (2015) note that it is important for teachers to feel trust and safety in their data use. "Psychological safety" is the belief that one will not be punished or humiliated for speaking up with ideas, questions, concerns, or mistakes.

Relational trust creates the necessary conditions whereby school professionals can take risks, trial new approaches and seek assistance when needed, as well as engage in open and honest conversations, informed by different sources of evidence (Bryk et al., 2015). Organisational cultures that build relational trust enable teachers to de-privatise their practice and challenge existing assumptions about teaching and achievement (Coburn & Turner, 2011).

Emotional reactions are another example of the way in which the conception of data use could expand beyond merely the cognitive domain in our evolving understanding of the practice of DBDM. What has rarely been acknowledged in the literature is that both ET and data use involve not only *thinking* processes but also have an emotional aspect. The field of cognitive psychology tells us that humans are simultaneously thinking and feeling human beings. We can't divide affect and rationality.

Implications

In this article I have attempted to further open the black box of data use processes by exploring the tensions that teachers feel when enacting data use. I have also considered the way in which the field of data use in education could benefit from exploring some of the conversations occurring around evaluative thinking. I have discussed the intersection of DBDM and ET in four distinct but related ways. If the field of DBDM learned from some of the conversations happening in the field of evaluative thinking, as proposed, what would it actually mean for schools to emphasise helping their teachers and leaders learn to reason and think evaluatively? What would it mean for educators to have greater evaluative thinking skills? It would entail the awareness and skills to interpret data, weigh evidence, consider the inevitable contradictions, articulate values, and deal with complexity (Patton, 2018a).

In ET, the discussions of concepts of complexity have highlighted a mindset shift from primarily linear and simple models of change to ones that are more complex. This shift acknowledges that many aspects of social systems are complex, and we cannot always

predict and control complex systems. The field of DBDM could also benefit from a greater focus on navigating this complexity and uncertainty. Shifting this way of thinking would encourage educators to rely on complexity-aware monitoring and evaluation practices to collaborate, reflect, and make adaptations. Educators can contribute to ongoing adaptations in their settings once they have gained greater awareness, knowledge, skills, and tools needed to engage in ET in their day-to- day work. Scholars and practitioners in DBDM could investigate some of the new ways of thinking and approaches to evaluation practice linked to planning and decision making that reflect these assumptions of complexity (Schwandt, 2019).

Also from the conversation in ET, I explored the role of values and value determination and discussed the intersection of collaborative social practice and values. DBDM has been based on a rational model of decision-making that de-emphasised the ethical-moral dimensions of problem-solving, but in fact, social interaction and co-constructing meaning are key to deciding what is worth doing in response to data. Incorporating concepts from ET highlights an acceptance of value pluralism, group deliberation of value positions, and purposely surfacing different values.

Emphasising these elements has implications for the tools schools employ for DBDM and the way in which they engage in capacity building or professional development for data use. For example, schools can develop and use tools for deliberation that purposely surface different values, such as asking each educator to role play or "wear different thinking hats" to represent different stakeholders or different interests. This could entail an exercise to consider how a particular practice/policy/intervention impacts various groups, such as teachers, parents, X group of students, Y group of students, and administrators. These tools can frame how people think about what aspects of data use are important and can guide their decision making and practice.

Incorporating ET could also influence professional development in data use. Initiatives designed to build capacity and encourage teachers to use student data may be overlooking crucial elements that could influence their effectiveness. Capacity building can emphasise sense making rather than narrow technical skills such as building and accessing student data systems, producing graphs, or making statistical calculations like determining effect size. Instead, schools can focus on learning how to engage in meaningful collaborative deliberation. This emphasis on reflection, interpretation, and improvement rather than on

discrete technical skills reflects the concept of evaluative thinking. Capacity building can focus on building awareness, knowledge, and skills in how to deliberate well on available courses of action, empathise, surface value positions, balance competing aims, and make judgements. Rather than merely cognitive capacities, the focus would expand to increasing cognitive, emotional, social, political, and moral capacities. Deliberating, negotiating, and surfacing different values require that teachers feel trust and psychological safety in their data use, so it would be important to use capacity-building approaches that take into account social learning and reducing anxiety. As this work is exploratory, these implications are just the beginning and would require further research and scholarship on the potential implications of borrowing concepts from evaluative thinking.

Chapter Ten

Discussion and Conclusion

The aim of thesis is to explore the actual practice of data use and self-evaluation activities, the perspectives of those embedded in these activities, and the ways these activities fit within wider contexts. It finds that the outcomes for teachers and schools in regards to data use and self-evaluation are not uniform, but are complex, multifaceted, mixed outcomes. This thesis contributes to the scholarly conversation through its argument that these multiple outcomes occur because they are shaped by the organisational contexts (systems, routines, tools and beliefs within the school) and the wider social and political context in which they work, one that is shaped to a large degree by different agendas of accountability. This discussion distils some of the themes of the varied outcomes.

In returning to the research questions outlined in chapter one, the various questions were addressed in these sections of the thesis:

- What underlying theories have informed teachers' data use, both in the literature and empirically? (Article 1, Article 3, Article 4)

- How do teachers experience the changing nature of their work as a result of the new expectations for data use in school self-evaluation? (all articles)

- How have teachers and school leaders acquired the capacity to carry out these new expectations and new ways of working? (Article 3)

- Has school self-evaluation led to organisational change, and if so, how? (Article 2 explicitly, Article 1 and 3 implicitly)

- How has the accountability context influenced the processes of data use and school self-evaluation? (Chapter 2, Article 1, Article 2, and Discussion)

- What aspects of the overall educational context have influenced these processes? (Chapter 3, Article 1, Article 2, and Discussion)

Across the various articles, one of the contributions of this thesis is that it highlights how data use and school self-evaluation can bring about organisational change, changes in practice, and changes to teachers' work. Therefore, the first section of this chapter reiterates

some of the data use practices these schools utilise and some of the organisational changes that have occurred as a result of their data use practices. It then turns to a key theme of this thesis: how practices are influenced by two competing accountability agendas, followed by a discussion of how teachers experienced these competing agendas. This thesis contributes to the scholarly conversation by advancing our understanding of the complex processes that occur when schools use data to improve student outcomes and by giving voice to the experiences and interpretations of teachers as it highlights how they interact with and respond to school self-evaluation processes.

This chapter then reviews the theoretical and methodological contributions of this thesis. As part of a reflexive stance consistent with a realist approach, I also comment on my own messy process of integrating the empirical data and a range of social and behavioural theories. Finally, I offer brief implications for further consideration for the system level, school level, and research. Throughout this discussion I reflect on the significance of the work in this thesis and how it helps us to advance our understandings of school self-evaluation and data use and offers tentative contributions for both the evaluation and educational communities.

Data Use Processes and Organisational Change

This section distils results across the four articles and outlines some of the data use practices these schools utilise and some of the organisational changes that have occurred as a result of their data use practices. As a reminder, the schools were selected for this study since they had previously been identified as effective at self-evaluation. They are therefore leaders in these processes as it has been documented that many schools in New Zealand find self-evaluation activities to be difficult. The schools in this study have developed some effective data use processes that enable improved outcomes for students, and generally speaking, teachers have a collective commitment to these processes. Some specific data use processes will be specified, before an analysis of the impact of these processes.

The schools in this study seek to use data to improve their teaching and improve outcomes for students. They all have implemented tracking tools that allow them to monitor individual students' progress toward NCEA. Tracking and monitoring is of critical importance in New Zealand given the complicated NCEA system in which students incrementally accrue credits throughout the year. This allows these schools to identify which students might be at risk of not succeeding, so they can be supported and be less likely to "slip through the cracks".

These schools also now share more data about students' progress with the students and their families, so students know what they need to do to succeed. In addition, these schools can conduct detailed analyses of specific groups of students, for example determining how Māori boys are performing on NCEA, so schools can determine how well they are serving priority learners. Another practice is that teachers and departments evaluate their units of study, using multiple sources of data, such as student assessment results, student survey data, and teacher perspectives. If certain units are found to be problematic, teachers are accustomed to digging deeper into the data and experimenting to find solutions, while using data to document their attempts to rectify the problems. Teachers are accustomed to sharing the results of these inquiries with small groups within the school and discussing what has worked well or not worked well.

As a result of the identification of students at risk of not achieving, all schools put interventions in place for these students. Their tracking and monitoring tools have shifted their ability to respond during the school year, whereas previously they only analysed their achievement data after the school year was over. While schools' ability to respond in real-time is certainly important, out of all the data use practices, some of these end-of-year interventions have the potential to be the most problematic. Many can be characterised as short-term fixes that ensure that students gain random NCEA credits so they can "get over the line" for achieving NCEA, but may not be related to meaningful learning or have a coherent connection to the rest of their learning pathways (Daniell, 2018; ERO, 2017). More implications will be discussed in the external accountability section below.

The data use practices have also resulted in organisational and behavioural changes with the schools themselves. The literature on data use in schools less frequently attends to these changes, despite the fact that they may be quite consequential (Coburn & Turner, 2011). Some of these organisational changes are less structural than symbolic, but are powerful reminders of the importance of the data, such as prominent visual displays around the school containing data on student achievement and attendance. Others are new routines and structures, such as set times for professional learning groups to share their inquiries or small group meetings to review the units of instruction. These are not ad hoc, but require planning, time, and space to collaborate around data. In these groups, teachers share their experiences, and articulate their thinking and the decision-making behind their actions. Then they reflect on what has occurred in the classroom as a result of experimenting with different practice. This is one example of a change in practice in response to data that has resulted in changes to

the ways that teachers interact with each other. Data use also sometimes results in structural reorganisation for students, such as pulling out students who have not achieved sufficient credits and placing them in smaller classes. Data use can lead to changes in teachers' roles within the school, such as tutor teachers' roles changing so they are responsible for checking on students' credits, making sure they are on track to succeed.

The articles that make up this thesis have zoomed in the lens to consider how school self-evaluation activities and data use practices occur and how these lead to organisational changes. The next section will zoom out to consider some features of the New Zealand context that contribute to this story. This thesis considers how the micro-level studies of teachers' experiences reflect macro-level trends.

Two Competing Accountability Agendas

A theme running throughout this thesis is that data use processes and school evaluation are shaped by the wider social and political context in which they occur, particularly the accountability context. The literature conveys that school self-evaluation and data use come from two competing agendas—one for external accountability and one for internal accountability—and that these different accountability agendas shape how assessment data are used (Cochran-Smith & Lytle, 2009; Gannon-Slater et al., 2017; Hargreaves & Braun, 2013; Hofer et al., 2020; Lai & Schildkamp, 2016; Schildkamp et al., 2013). Much of the literature suggests that, depending on the accountability agenda, data use can either lead to positive outcomes and greater quality *or* to negative outcomes and reduced equity (e.g., Datnow et al., 2017). However, in my research, it appears to lead to both. It was not an either/or situation, it was *and*. In New Zealand secondary schools internal and external accountability co-exist and this leads to complex, multifaceted, mixed outcomes.

This thesis provides nuance and shading to the normally-dichotomous view of accountability and shows the range of perspectives that accountability can take. In the data use literature, the tension between internal and external accountability is often not addressed, yet it is important for understanding what happens when educators engage with data in their everyday work (Mausethagen et al., 2018). The co-existence of the two accountability regimes can be viewed as a clash in the progressive ideals and the neoliberal imperatives that co-exist within the New Zealand educational context, as described in Chapter Three.

As a reminder, internal accountability in the New Zealand education system manifests in several ways. New Zealand, more so than many other countries, has embedded the concept of Teaching as Inquiry in its curriculum (MOE, 2007; Sinnema et al., 2017). The education system views teachers as adaptive experts who have the ability and autonomy to make decisions based on data to solve problems at the local level (MOE, 2007). Teaching as Inquiry was intended as an internal accountability that holds teachers accountable for embodying an inquiry stance (Sinnema & Aitken, 2011). Internal accountability was also a part of the reason for the *Tomorrow's Schools* education reform, as some of the initial goals were to meet locally-determined educational needs and priorities, move educational decision-making closer to the community, and improve educational outcomes for disadvantaged groups (Timperley, 2013; Wylie, 2012). Another aspect highlighting internal accountability is the evolution of ERO's school review system from an external, compliance, accountability-oriented focus to a system of *complementary evaluation* that is more focused on improvement and is collaborative, negotiated, and complementary of schools' internal evaluation that is now at the heart of the system of school evaluation (ERO, 2017; Lai & Kushner, 2013; Mutch, 2012). Despite all the ways internal accountability is embedded in the New Zealand system, it sometimes bumps up against external accountability pressures in the form of inquiry for compliance, market accountability, and a top-down imposition of an 85% target for NCEA Level 2. Some of the impacts of these external pressures are addressed in the next section.

These two accountability agendas are a theme running through the four articles. Article One explores the multiple agendas of accountability and the tensions teachers experience in enacting inquiry. This article dissects the various theories of inquiry that can be in tension and explores how teachers experience these tensions, especially when inquiry slips from internal to external accountability. Article Two highlights that the participating schools are leaders in internal accountability and have data tools allowing them to identify and support students who might be at risk of not achieving. On the other hand, more so than the other articles, Article Two also highlights external accountability and raises some concerns about the unintended consequences of data use. Article Three highlights internal accountability and demonstrates that a strong reliance on internal accountability at a system level requires that teachers have a high level of evaluation capacity, so evaluation capacity building is a crucial element. While the article does not directly address the tension, there is a friction in the system because schools are accountable for their self-evaluation but many schools lack sufficient evaluative capability to carry out these processes effectively. Article Four returns

to the tensions of Article One, but moves from the micro to the macro. This article asks what the processes, interactions, and tensions experienced by individual teachers signify and how theory from the emerging field of Evaluative Thinking can contribute. The following section will address some of the impacts of external accountability in this system.

External Accountability Pressures

The most disquieting finding of this thesis is the potential impact of external accountability pressures on data use processes and school self-evaluation. Some authors have warned us that external accountability pressures have the potential to jeopardise authentic practitioner inquiry and self-evaluation, with a danger that these processes become corrupted, managerialised, appropriated, contrived (see Peters, 2016; Smardon & Charteris, 2017). In Articles One and Two, teachers articulate that they feel a strong pressure to narrowly focus on summative student achievement outcomes. The overseas literature notes that the misuse and abuse of data are more likely in high-stakes accountability systems that put pressure on teachers to use data in certain ways, and demonstrates the negative outcomes of accountability policies on students (Booher-Jennings, 2005; Farrell & Marsh, 2016; Lingard & Sellar, 2013; Mausethagen et al., 2018).

External accountability pressures can lead to negative outcomes. Accountability pressure in the form of imposed arbitrary numerical targets (85% NCEA) and high-visibility publication of results in magazines can result in schools gaming the system, with negative consequences for the learning and achievement of the most disadvantaged students. This can result in unethical, inappropriate practices to enhance schools' NCEA results, or in particular, the specific results that will appear in the media (ERO, 2017). Practices such as narrowing of the curriculum and concentrating attention on "cusp" students near the threshold of required achievement at the expense of high-needs students whose current performance falls further below the threshold have been shown overseas. In New Zealand, these practices are less-widely acknowledged and examples of inequitable practices are much more recently documented in our literature, for example, in data showing that students from low socioeconomic areas often receive an easier curriculum with fewer opportunities to engage with challenging material (Wilson et al., 2016).

This thesis describes another inappropriate practice in which schools devote undue attention to students on the cusp of achieving, as shown in Article Two. Data use, within the context of external accountability pressures, can amplify efforts to focus on students on the

threshold, and result in disadvantaging some students. When schools focus on cusp students, this can limit educational opportunities for other students and therefore compromise the equity goal of data use, since "equity cannot be achieved unless data use efforts focus on all students" (Datnow & Park, 2018, p. 144). An ongoing challenge for the New Zealand system is to actively attend to the 15% of students who are not achieving Level 2 NCEA rather than disproportionately focusing on students who are close to the threshold.

Teachers Navigating These Accountability Agendas

Teachers are in a contested space and need to negotiate between different accountability discourses that co-exist in New Zealand. Teachers' work has changed due to the increase of school self-evaluation activities, but most research in this area has not explored teachers' perspectives and how they make sense of the competing accountability demands in their context (Mausethagen et al., 2018). The co-existence of two agendas means that the outcomes of school self-evaluation for teachers are not uniform; there are a variety of outcomes, some positive and some negative. When there is conflicting advice, competing demands, and contradictory purposes, the tensions created from those contrasting influences can make it difficult for educators to navigate these processes.

The data in this study show that teachers have good intentions and still aspire for internal accountability, genuine inquiry, and authentic school self-evaluation. That remains the ethos. But the system often incentivises data use in service of getting students over the line, and this results in teachers engaging in practices such as narrowly focusing on summative student assessments. The progressivist, participatory, complementary tradition of New Zealand's school self-evaluation can cloak a neoliberal external accountability agenda. In some cases, teachers may not realise that their practices are being corrupted, such as the participant in Article One who only notices that her practice of teacher surveillance "sounds terrible" when she is explaining it to me. She is tacitly accepting the practices expected of her, but when she explains it out loud, it sounds terrible to her. The external accountability pressures often "seep under the door, with change happening so gradually that practices become normalised" (Smardon & Charteris, 2017, p. 181).

One of the findings of this thesis is that for teachers in this system, there is an affective component to these school self-evaluation processes. The affective component can manifest as teachers feeling the tensions of navigating these competing underlying accountability agendas or feeling the pressure of being subjected to external accountability pressures. In

New Zealand, accountability has been devolved to the school and then to the individual teacher, and my data shows that teachers feel the pressure of that accountability. Educators in this study expressed the strains they feel when accountability has snuck in, and the emotional aspects that can arise in the process. Møller (2009) argues that educators are increasingly "caught in the cross-fire" of educational reforms focused on accountability (p. 42). He goes on to explain that it can be stressful and emotional when one's "personal accountability", that is, the values that are sacred to a person, are in conflict with other kinds of accountability, such as consumer accountability.

In addition to these content contributions to the knowledge base, this thesis also seeks to add to the literature through its methodological and theoretical contributions. The following will explain these contributions.

Methodological Contribution

This thesis makes a methodological contribution to the field of evaluation in its use of two innovative research methods, Rich Pictures and realist interviewing, and its unique combination of the two methods. Realist evaluation approaches are growing in popularity, particularly in the UK and in the disciplines of applied health research, and they often utilise qualitative interviews, but additional explicit examples of realist interviewing are needed (Manzano, 2016). My study is unique in using the novel approach of Rich Pictures as an initial participatory activity to develop preliminary realist programme theories. While Dalkin and colleagues (2018) argue that the epistemology of Soft Systems Methodology, from which the method of Rich Pictures derives, is compatible with realist approaches, they explain that these methodologies had not been combined before. Their approach to combining the methodologies did not involve having participants create Rich Pictures, as in my study. I used Rich Pictures in phase one focus groups as a process to develop initial theories and to develop more detailed questions for the subsequent round of interviews, based on the recommendation of Pawson (2013) to use of an open-ended qualitative activity for these purposes. This ensures that the initial theory development is participatory and that theories are grounded in actual school practice and not merely an idealised version of "best practice" from policy or the literature. Using Rich Pictures in this participatory way increases the transparency, robustness, and accuracy of the theory building process (Dalkin et al., 2018). The iterative development, testing, and refinement of these theories helps ensure they are grounded in the experience of participants as well as broader insights from

the wider literature, which enhances relevance, empirical weight, and portability to other settings (Punton, & Vogel, 2020).

Rich Pictures had several benefits as a research method. Using Rich Pictures gave participants an opportunity to think about their teaching practice in a different way, which enabled the surfacing of prior assumptions, implicit theories, and tacit knowledge (Dalkin et al., 2018; Williams & van't Hof, 2016). Unlike some other forms of data collection I have used previously, using Rich Pictures, I was surprised how quickly participants revealed meaningful, honest, and occasionally vulnerable insights. The technique of Rich Pictures allowed people to feel more at ease, more relaxed and they could talk about deeper issues more easily. For example, when describing their picture, a teacher could say "here is the community looking in at what we're doing" as opposed to directly saying, "we feel external accountability pressure that everyone is looking at our assessment results." Other scholars have found that Rich Pictures enable learning about abstract, emotional, or tacit knowledge like the social atmosphere among the different actors or stakeholders (Oakden, 2014). It was also an effective way to start the series of focus groups, as it acted like an 'ice-breaker' and enabled communication between participants, and encouraged greater participation to engage with their further discussions (Williams & Van't Hof, 2016).

Realist interviews also had several benefits. One implication was prioritising the study of underlying change processes and mechanisms. Reflecting on my experience of realist interviews, they were more focused, specific, and granular than other types of qualitative interviews. For example, asking for specific examples of how particular processes work in their school and asking about the specific structures, routines and tools that are part of that process encourages participants' stories about specific experiences and issues. Other researchers have also found that because data collection efforts are guided by initial theories, realist interviews are more focused (Astbury & Leeuw, 2010).

One technique that was effective in the interviews was developing a list of 10 intermediate outcomes of school self-evaluation, then presenting these initial theories in realist interviews and asking participants to choose three items and explain *how* the processes work in their school. Walking me through the procedures encouraged participants to provide specific examples and stories related to the process. It encouraged participants to give concrete answers about their processes and kept participants focused on these and on topic. Pawson and Tilley (1997) explain that in realist interviewing the researcher makes clear to the

interviewee their theory, or potential explanations for a specific set of actions. Then they allow the respondent to clarify these theories by asking them to articulate their perception of the processes. The researcher teaches the respondent about their theory and then learns what the respondent has to say about this theory. It's the "here's my theory, what's yours?" strategy of data collection. However, asking about a preconceived theory can risk theory confirmation bias, which is the respondent agreeing with the researcher's theory for the sake of politeness (Dalkin et al., 2018). My technique of requesting that participants actively choose 3 out of the 10 intermediate outcomes helped avoid the potential problem of 'leading the interview'. In the future, I may write a methodological journal article to contribute to the developing understanding of realist interviews.

Contributions to Theory

This thesis seeks to contribute to theory in two different disciplines, evaluation and education. It brings theory from the field of education into evaluation and vice versa. Article Three introduces a learning theory from education into the field of evaluation. Article Four argues that advances in unpacking the theory underlying Evaluative Thinking (ET) in the field of evaluation can contribute to conversations around DBDM in education. The bodies of literature in these two fields are often distinct from each other, and rarely overlap. This thesis is intended as an invitation to expand the conversation between these fields about the processes of organisational self-evaluation.

Authors argue that evaluation capacity building efforts in the fields of both evaluation and DBDM are largely atheoretical (Huffman et al., 2008; Marsh & Farrell, 2014), and it is true that they do not seem to explicitly draw on particular theories. However, now that I take a step back and consider the fields of evaluation and education, I realise that it is not necessarily that they are atheoretical as it is that proponents may not be explicit about that theory, which presumes that this is the natural or only way to conceive things. A deeper look reveals that they have certain underlying theoretical assumptions, histories, epistemologies, ontologies, and beliefs about change and causality. This thesis is pushing back on some of the established ways of thinking in both fields and attempting to extend evolving theoretical understandings. Both evaluation and DBDM have histories grounded in theories that: are rational, cognitivist, and behaviourist; are based on a conception of causality that is linear, proportional, and simplistic; and treats change processes as controllable, predictable, universal, and understandable (Archibald, Neubauer, & Brookfield, 2018; Patton, 2018a;

Schwandt, 2019; Vanlommel, 2018; Vanlommel & Schildkamp, 2019). In their histories, both DBDM and evaluation have a similar dominant ethos of rationality, control, predictability, objectivity, accountability, and an emphasis on methods and procedures. While these fields more recently are beginning to acknowledge the importance of concepts such as collaborative sensemaking, interpretation, context, and complexity, this thesis continues to extend those developments. For example, Article Three offers a model of ECB that moves away from a behaviourist model and suggests the alternative sociocultural theory. Mandinach and Schildkamp (2020) argue that the field of educational data use needs to move away from its neo-behaviourist perspectives, but do not offer models that illustrate this. Article Four contributes to a field that is just beginning to develop a theory of teacher decision-making processes that are based on teacher judgement and not based solely on rationality (Vanlommel, 2018). It does this by suggesting DBDM draw on the efforts to unpack the construct of evaluative thinking (ET), particularly the conversations around complexity, values, and the affective domain.

My Journey with Theory

While this thesis makes a contribution to the theory in two fields, the process of drawing on substantive theory was not always straightforward or simple for me. I suspect I'm not alone in that feeling as a doctoral researcher. In Chapter Four, I discuss my relationship with theory and my background as an applied researcher. Here, I reflect on some of my growth wrestling with theory throughout my doctoral journey.

The back-and-forth conversation between my empirical data and a range of social and behavioural theories was a big part of my journey. In all of the articles, my analysis led to a deeper investigation of the broader social science theoretical literature– steering me into the fields of political science, psychology, adult learning theory and theories about complexity. I wrote multiple versions of several of the articles, not merely tweaking the drafts, but nearly-completely abandoning the previous version. A good example was the process of writing Article Four, *Beyond the Rational Model of Decision Making*. I first wrote this as a 2,000-word abstract for a full paper presentation at the American Education Research Association. When I expanded the abstract to a full journal article trying to develop these reflections on my data, I looked for the right theory to make sense of it. I wrote multiple versions of the article, including an earlier draft bringing in theory from the fields of psychology and economics before re-writing it after hitting on the developments in the field of evaluative

thinking. My theoretical hunches unravelled many times and had to be stitched back together (Punton et al., 2016). The process was like trying on a theory to see if it fits, and unfortunately often getting pretty far down a track with writing it up, only to discover either by myself or more often, in conversation with my supervisors, that the theory wasn't really a good fit. In each article the analysis involved this type of back-and-forth conversation between the empirical data and theory to hopefully combine diverse ideas in new ways (Astbury & Leeuw, 2010).

Drawing on critical theory was also sometimes a problematic part of the journey, particularly when publishing my articles. I did not submit them to specific critical theory-oriented journals, but to more general education or evaluation journals, since I wanted to problematise some of the unquestioned assumptions in these fields to a non-critical theory audience. In the peer reviews for both Article One and Two, reviewers commented that the article had too much of a "political stance" and I twice softened this stance and removed some passages with a critical theory orientation. Now that I'm reflecting on this journey, I see that essentially the reviewers compelled me to reduce my critical stance and my critique of the neoliberal education system. On reflection, I may have been subjected to external pressures like my research participants, and I was perhaps more even aware than my participants were of what was happening, and I complied in order to have my articles published.

Implications

This section offers areas for further consideration. It offers an implication for the New Zealand education system, one for schools' evaluation capacity building efforts, and one for further research.

The right balance of tight and loose

It is an ongoing challenge for the New Zealand education system to resolve the tension between internal and external accountability agendas. Supporting internal accountability while addressing external accountability concerns so that improvement and accountability work together and not at cross-purposes is an ongoing challenge for the system that requires further conversations in policy and research. This question is not just relevant for New Zealand but is one that many countries wrestle with. "The search for an appropriate balance between internal and external evaluation is underway in multiple locations and in many

countries…Educational policy-makers in various Western countries are currently struggling with this problem" (Vanhoof & Van Petegem, 2007, p. 102).

Because the accountability pressures in New Zealand contrast with many of the strong top-down, government-driven, prescriptive accountability measures found overseas described at the beginning of this thesis, they are more nuanced and initially less apparent. The accountability context here is a mixture of consumer accountability and the particular tight—loose—tight arrangement at the time of the study. Schools are responsible and accountable for meeting achievement outcomes imposed by central government within a managed, or 'steered', school marketplace (Court & O'Neill, 2011; O'Neill & Snook, 2015). At the time of this study, there were "tight" accountability targets (85% NCEA Level 2) combined with loose guidance and a great deal of freedom on how schools choose to reach these targets in a context of a loose curriculum, assessment, and decentralised decision-making policy. With a rigid target, flexible curriculum and assessment system, and intense public scrutiny, it is not surprising that schools would engage in practices to make themselves look good. This arrangement contributes to the fact that the outcomes of school self-evaluation are not uniform. The flexibility of the self-evaluation, curriculum and assessment systems becomes a double-edged sword in the context of external accountability pressures. In this environment, teachers can use data to personalise their responses and help students make genuine achievement gains, but it can also enable focusing on students' short term success or focusing on cusp students. When the level of accountability is appropriate, the flexibility and autonomy that characterise other aspects of the system are an advantage. But when the balance of accountability is not right, autonomy has the potential to worsen outcomes for students, particularly those most at risk.

Efforts to improve would not merely require having *less* accountability in the system, but having accountability that ensures genuine learning for all students and not a system that is perhaps too flexible in parts and too rigid in others. A challenge for the New Zealand education system is to find the right balance of "loose" and "tight" in its school evaluation system. Here, I return to the metaphor used in Article One of walking on a tightrope. If the rope is too tight or too loose, it is extremely difficult to navigate, but between those two extremes there is a "sweet spot" with just the right balance. Finding this is the challenge.

Evaluation capacity building in schools

In order for the balance of internal and external accountability to function properly, it is necessary to build capacity for internal accountability. A strong reliance on internal accountability requires a high level of evaluation capacity. There is a tension in the system because schools are accountable for their self-evaluation but it has been documented that many schools lack sufficient evaluative capability to carry out these processes effectively place (ERO, 2017; Kiro et al., 2016; Lai, 2013; Mutch, 2012; Nusche et al., 2012; Timperley, 2013) . Therefore ECB within schools is an important activity going forward. This section will describe two recommendations in this area.

The first recommendation is that government agencies provide greater support for schools in their ECB efforts. This thesis shows that the participating schools conduct their own ECB, develop their own tools, and in some cases even create their own tracking and monitoring database tools. Not every school in the country can do this. Government agencies have perhaps been too 'loose' and not offered sufficient support for evaluation capacity building in schools. Fullan and colleagues (2015) advocate for a reciprocal accountability which holds the education system accountable for providing the conditions schools need for improvement, such as adequate resources and capacity building, at the same time as holding schools and teachers accountable for performance and improvement. I am not advocating that central government become too 'tight' and require schools to use prescribed tools and processes; after working fairly independently for decades, schools would rebel against this, it would risk becoming a compliance exercise, and some schools already have tools, structures, and routines that are working well for them. Schools need to be able to work closely alongside an organisation that will help them build capacity, monitor their progress, and help them improve (Tomorrow's Schools Independent Taskforce, 2019). In my research, it surprised me to learn that many secondary schools still struggle with effectively using a database tool to collect longitudinal data on student achievement (Daniell, 2018; Kiro et al., 2016). Every school should have this ability, especially given the complexity of the NCEA assessment system, and government agencies, such as the MOE or ERO could assist with this. However, this thesis shows that having and managing an effective data management tool is necessary but not sufficient for enabling school self-evaluation that ensures meaningful successful outcomes for all students.

It is also suggested that schools draw from the lessons learned in this thesis when designing their efforts to build capacity; these lessons are not limited to the article specifically focused

on ECB. This thesis shows that ECB is not only about the discrete technical skills required to analyse data. Some of the relevant insights of this thesis include that self-evaluation and inquiry can involve affective elements for teachers; data use is a process of sensemaking and interpretation; accountability pressures pose a danger of gaming the system and as a result short-changing some students; many of the necessary decisions are value-laden; many of the problems and solutions are complex; and if teachers are coerced to do inquiry, it can become a tick-the-box compliance exercise.

One way to address several of these concerns is to pay greater attention to the theories underlying the tools and structures schools use to build capacity. This thesis reinforces that teachers' work is influenced by the systems, routines, and tools of the schools in which they teach. In the early days of the DBDM field, it was assumed that data would lead to improvement in a straightforward, linear way, but there is now the understanding that the processes of interpretation are key. Between the data and the response there is a whole series of interpretations, underlying assumptions, inferences, and judgements (Coburn & Turner, 2011). As data use is a process of noticing and interpretation, tools and routines shape what data people even notice and who interacts with whom around the data and in what ways. In an educational system like New Zealand's where schools have a great deal of freedom and autonomy, social norms and processes within each school perhaps become more important. These processes can be in the form of rewards and sanctions that invite the possibility of compliance and gaming the system or they can be in the form of tools and norms that promote authentic inquiry. Going forward, practitioners should consider the underlying assumptions embodied in their processes and researchers may view this as an area ripe for further attention.

Further research

Throughout this thesis I have underscored that there is not just one story told in this research, but instead there are multiple, multifaceted stories. Using a realist understanding, this thesis acknowledges that research cannot presume to have a single "objective," "correct" understanding of the world. That was not my aim. In this spirit, I acknowledge that this research could have been conducted in different ways and that the exploratory findings in this thesis could be examined in further research.

The first consideration is that this research took place in a particular time and place in New Zealand secondary schools. Since the time of data collection, some important elements of the

education system and political context have shifted. The governing party has changed in New Zealand, and therefore some education priorities and policies have altered. One significant development relevant to this thesis is that the government has eliminated the "tight" target for 85% of students to achieve Level 2 NCEA. Another change has been a willingness to begin looking at the self-managing schools model, for example through a comprehensive process by an independent taskforce. As the story that is told in this thesis is not one that is static, it would be interesting to conduct similar research with teachers to explore the impact of the changes to the accountability and wider political context.

Another interesting concept to study would be collaboration across schools regarding data use. Since school competition is an important concept in the New Zealand context, it would be fruitful to explore examples of collaboration. This could occur through the Communities of Learning/Kāhui Ako, an initiative created with the goal of reducing competition and raising student achievement by creating groups of schools working together. It could be productive to investigate how those groups are sharing data and building evaluative capacity, and to determine if this has productive outcomes. The type of needed collaboration is shown by the simple example that when I worked as a consultant with a Community of Learning/Kāhui Ako, I was surprised at how little primary school teachers understood about the nature of student assessment data at the secondary school level and vice versa. Future research could determine whether and how the Kāhui Ako initiative does actually build evaluative capacity for teachers, whether these communities do engender greater cross-school data use, and whether greater data sharing leads to improved outcomes for students.

Another fruitful area for further research would be to conduct in-depth observational analysis of schools' data use activities. The research in this thesis has been exploratory and has focused on teachers' perspectives. This thesis highlights the importance of interpretation and sensemaking, and future observational research could capture more detail and nuance of teachers' actual interpretative work around data. This study also highlights the organisational context and the importance of organisational structures, norms, and tools. An observational study could focus on the actual use of those tools and resources.

These three implications—for the system, schools, and research—should be viewed as what they are, areas for further consideration, not viewed as recipes for success. This thesis was not school effectiveness research intended to provide recipes for the creation of more effective schools. More importantly, such recipes are not possible, a point that I have made

in this thesis. I have argued against the concept that human behaviour can be shaped by applying the right formulas or recipes. Evaluation, education, accountability, agency, change processes. These are not simple problems that can be addressed with recipes, these are complex issues. My purpose in this thesis has been to engage with this complexity. I hope in some small way I have contributed to the wider conversations.

Appendices

Appendix A: Participant Information Sheet

THE UNIVERSITY
OF AUCKLAND
FACULTY OF EDUCATION
Te Kura Akoranga o Tāmaki Makaurau
Incorporating the Auckland College of Education

Faculty of Education
Epsom Campus
The University of Auckland
Private Bag 92601, Symonds
Street
Auckland 1150, New Zealand
Telephone 64 9 623 8899

Participant Information Sheet for School Leaders/Teachers

Title of Research Project: Self Evaluation in New Zealand Secondary Schools
Researcher: Lisa Dyson

Supervisors: Dr Carol Mutch and Dr Ritesh Shah

My name is Lisa Dyson and I am a PhD student at the University of Auckland in the Faculty of Education. I am conducting research into the use of effective self-review processes in NZ secondary schools. The proposed research will describe current practices in school self evaluation, assess its benefits for schools and investigate how school personnel acquire the capacity to carry out these processes. I'd like to invite you to participate in this study.

Project Procedures
Your participation in this study would involve taking part in an individual interview of approximately 60 minutes. The interview will be carried out at a time and place of your choosing. With your consent, the interview will be audio-recorded and later transcribed by me. The recorded interviews will not be shared with third parties, including your principal. I will analyse the interview data and a summary of findings will be made available to your school. I may also request school documents that may be relevant to the study such as school policies or action plans that support the information discussed in interviews. Your principal has consented for me to access these documents.

Data Care and Use
The data and any signed consent forms will be stored in a locked cabinet or on a password-protected computer at the Faculty of Education for six years. At the end of this time all paper data will be shredded and electronic audio files deleted. Pseudonyms will be used for all participating schools and individual participants to minimize risk of identification. The data may be used as part of the researcher's thesis and may be used in conference presentations and academic journals.

Participant Rights

Participation in this study is voluntary. If you choose to take part, you can withdraw at any stage. You can decline to answer any questions that you feel uncomfortable answering and may ask for the recording device to be switched off at any time during the interview.

You are welcome to discuss this research with me or my supervisors, Dr Carol Mutch and Dr Ritesh Shah. Our contact details are listed below.

Lisa Dyson   lisa.dyson@auckland.ac.nz   027 533 0075

Carol Mutch   c.mutch@auckland.ac.nz   623 8899 ext 48142

Ritesh Shah   r.shah@auckland.ac.nz   623 8899 ext 46356

For any concerns regarding ethical issues, you can contact the Chair of the University of Auckland Human Participants Ethics Committee, The University of Auckland, Private Bag 92019, Auckland, 1142, or phone 373-7599 extn 87830.

Appendix B: Consent Form

THE UNIVERSITY
OF AUCKLAND
FACULTY OF EDUCATION
Te Kura Akoranga o Tāmaki Makaurau
Incorporating the Auckland College of Education

Faculty of Education
Epsom Campus
The University of Auckland
Private Bag 92601, Symonds
Street
Auckland 1150, New Zealand
Telephone 64 9 623 8899

Consent Form: School Leaders/Teachers

This form will be retained securely for a period of six years.

Title of Research Project: Self Evaluation in New Zealand Secondary Schools
Researcher: Lisa Dyson
Supervisors: Dr Carol Mutch and Dr Ritesh Shah

I have read the Participant Information Sheet, have understood the nature of the research and why I have been selected. I have had the opportunity to ask questions and have them answered to my satisfaction.

- I agree to take part in this research.
- I understand that participation is voluntary and that I can withdraw at any time.
- I understand that participation involves an interview of approximately 60 minutes that will be audio taped.
- I understand that the data gathered may be used as part of the researcher's thesis and may be used in conference presentations and academic journals.
- I understand that all names will have a pseudonym in order to ensure confidentiality and that identifying information about the school will not be published.
- I understand that I can request a summary of the research results at the end of the study.

Name: _____

Signature: _____   Date: _____

APPROVED FOR 3 YEARS BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE on 2 April 2015 for a period of three years. Reference number: 014082.

Appendix C: Focus Group Questions

Can you give me an example of where you think your school self-evaluation is working really well? Why did you choose that example? What do you think is making it work well?

Can you think of an example of a time that this type of process led to a teacher changing their practice in some way?

Can you think of an example of a time that this type of process led to a more structural change? (curricular change, change to pastoral care, communication with parents)

Can you think of a time when your self-evaluation processes didn't get the result you expected? What went wrong? (What support factors weren't in place? And what were the barriers?)

Can you identify any types of teachers who are more likely to change something because of self-evaluation? Can you think of who is less likely to change something?

Have you had any teachers who are resistant to your self-evaluation efforts? What have you done in response?

What helps your self-evaluation to work and what are some of the barriers?

Do you think your self-evaluation processes have helped to raise student achievement? What about raising achievement for students who are underachieving? What evidence do you have?

Do you think that school self-evaluation leads to school improvement in your school?

Appendix D: Interview Questions

What is your particular role in your school's self-evaluation?

Describe how your school carries out ongoing self-review. What are some of the:

- structures,
- organisational routines,
- processes,
- tools and resources, specific data sources that are part of that process?

Can you give me an example of where your self-evaluation is working really well? Why did you choose that example? What do you think is making it work well?

Show them the list of intermediate outcomes. Walk through top 3 of these. (*Explained in Chapter Five)

What aspects of your self-evaluation have had the most impact? That is, they've resulted in the most changes or improvement?

Can you think of a time when your self-evaluation process didn't get the result you expected? What went wrong? (What support factors weren't in place? And what were the barriers?)

What helps your self-evaluation to work and what are some of the barriers?

Can you think of any changes in your own professional practice because of the school's self-evaluation?

What evidence do you have that your school's self-evaluation processes have helped to raise student achievement? What about raising achievement for students who are underachieving?

Can you give me an example of when your self-evaluation made a difference for targeted groups of students?

Are there any tools or processes your school uses that help support teachers' inquiry processes? The overall school self-evaluation processes?

How have teachers and school leaders in this school acquired the capacity to be able to carry out these processes? What resources/external expertise have you drawn on to build capability?

References

Adams, P. J., & Buetow, S. (2014). The place of theory in assembling the central argument for a thesis or dissertation. *Theory & Psychology*, *24*(1), 93–110. https://doi.org/10.1177/0959354313517523

Archibald, T., Neubauer, L. C., & Brookfield, S. D. (2018). The critically reflective evaluator: Adult education's contributions to evaluation for social justice. *New Directions for Evaluation*, *158*, 109–123. https://doi.org/10.1002/ev.20319

Archibald, T., Sharrock, G., Buckley, J., & Young, S. (2018). Every practitioner a "knowledge worker": Promoting evaluative thinking to enhance learning and adaptive management in international development. *New Directions for Evaluation*, *158*, 73–91. https://doi.org/10.1002/ev

Astbury, B., & Leeuw, F. L. (2010). Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation*, *31*(3), 363–381. https://doi.org/10.1177/1098214010371972

Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy*, *18*(2), 215–228. https://doi.org/10.1080/0268093022000043065

Bazeley, P. (2009). Analysing qualitative data: More than 'identifying themes'. *Malaysian Journal of Qualitative Research*, *2*(2), 6–22.

Benade, L. (2012). Challenging the domestication of critical reflection and practitioner reflectivity. *Educational Philosophy and Theory*, *44*(4), 337–342. https://doi.org/10.1111/j.1469-5812.2011.00824.x

Benade, L. (2015). Teaching as inquiry: Well intentioned, but fundamentally flawed. *New Zealand Journal of Educational Studies*, *50*(1), 107–120. https://doi.org/10.1007/s40841-015-0005-0

Bendikson, L. (2018). *Inquiry – Much harm done in its name?* Retrieved from http://www.uacel.ac.nz/language/en-us/publications/articletype/articleview/articleid/395/inquiry--much-harm-done-in-its-name#.WtL2JS-B28o

Bennett, G., & Jessani, N. (Eds.). (2011). *The knowledge translation toolkit, bridging the know–do gap: A resource for researchers*. New Dehli, India: SAGE.

Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, *52*(5), 861–893. https://doi.org/10.3102/0002831215599251

Bhaskar, R. (1978). *A realist theory of science*. Sussex, UK: Harvester Press.

Biesta, G. J. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory*, *57*(1), 1–22. https://doi.org/10.1111/j.1741-5446.2006.00241.x

Bishop, R., Berryman, M., Wearmouth, J., & Peter, M. (2012). Developing an effective education reform model for indigenous and other minoritized students. *School Effectiveness and School Improvement*, *23*(1), 49–70.

Bocala, C., & Boudett, K. P. (2015). Teaching educators habits of mind for using data wisely. *Teachers College Record*, *117*(4), 1–20.

Booher-Jennings, J. (2005). *Below the bubble: "Educational triage" and the Texas accountability system*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Bourgeois, I., Whynot, J., & Thériault, É. (2015). Application of an organizational evaluation capacity self-assessment instrument to different organizations: Similarities and lessons learned. *Evaluation and Program Planning*, *50*, 47–55. https://doi.org/10.1016/j.evalprogplan.2015.01.004

Brady, A. M. (2016). The regime of self-evaluation: Self-conception for teachers and schools. *British Journal of Educational Studies*, *64*(4), 523–541. https://doi.org/10.1080/00071005.2016.1164829

Brown, C., Schildkamp, K., & Hubers, M. D. (2017). Combining the best of two worlds: A conceptual proposal for evidence-informed school improvement. *Educational Research*, *59*(2), 154–172. https://doi.org/10.1080/00131881.2017.1304327

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*, 32–42.

Bryk, A., Gomez, L., Grunow, A., & LeMahieu, P. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford, UK: Open University Press.

Buckley, J., Archibald, T., Hargraves, M., & Trochim, W. M. (2015). Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation*, *36*(3), 375–388. https://doi.org/10.1177/1098214015581706

Cardno, C., Bassett, M., & Wood, C. (2017). A tale of two secondary schools ' efforts to embed 'teaching as inquiry ' within an appraisal system. *Leading & Managing*, *23*(1), 12–24.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: Doing it better. A practical guide to predicting if a policy will work for you*. Oxford, UK: Oxford University Press.

Charteris, J., & Smith, J. (2017). Sacred and secret stories in professional knowledge landscapes: Learner agency in teacher professional learning. *Reflective Practice*, *18*(5), 600–612. https://doi.org/10.1080/14623943.2017.1304375

Checkland, P., & Poulter, J. (2006). *Learning for action: A short definitive account of soft systems methodology and its use for practitioners, teachers and students*. Chichester, UK: John Wiley & Sons.

Clinton, J., & Dawson, G. (2018). Enfranchising the profession through evaluation: A story from Australia. *Teachers and Teaching: Theory and Practice*, *24*(3), 312–327. https://doi.org/10.1080/13540602.2017.1421162

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement*, *9*(4), 173–206. https://doi.org/10.1080/15366367.2011.626729

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education*, *118*(2), 99–111. https://doi.org/10.1086/663272

Cochran-Smith, M., & Lytle, S. L. (2009). *Inquiry as stance: Practitioner research for the next generation*. New York, NY: Teachers College Press.

Collins, A., Brown, J. S., & Holum, A. (2003). Cognitive apprenticeship: Making thinking visible. In L. B. Hall & M. W. Resnick (Eds.), *The principles of learning: Study tools for educators* (pp. 1–18). Pittsburgh, PA: University of Pittsburgh.

Connell, R. (2009). Good teachers on dangerous ground: Towards a new view of teacher quality and professionalism. *Critical Studies in Education*, *50*, 213–229. https://doi.org/10.1080/17508480902998421

Connell, R. (2015). Markets all around: Defending education in a neoliberal time. In H. Proctor, P. Brownlee, & P. Freebody (Eds.), *Controversies in education* (pp. 181–197). https://doi.org/10.1007/978-3-319-08759-7

Couch, D. (2018). From progressivism to instrumentalism: Innovative learning environments according to New Zealand's Ministry of Education. In L. Benade & M. Jackson (Eds.), *Transforming education* (pp. 121–133). https://doi.org/10.1007/978-981-10-5678-9

Court, M. (2004). Talking back to new public management versions of accountability in education: A co-principalship's practices of mutual responsibility. *Educational Management Administration & Leadership*, *32*(2), 171–194. https://doi.org/10.1177/1741143204041883

Court, M., & O'Neill, J. (2011). 'Tomorrow's Schools' in New Zealand: From social democracy to market managerialism. *Journal of Educational Administration and History*, *43*(2), 119–140.

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

Dale, R., & Ozga, J. (1993). Two hemispheres—Both New Right? 1980s education reform in New Zealand and England and Wales. In B. Lingard, J. Knight, & P. Porter (Eds.), *Schooling reform in hard times* (pp. 63–87). London, UK: Falmer Press.

Dalkin, S., Lhussier, M., Williams, L., Burton, C. R., & Rycroft-Malone, J. (2018). Exploring the use of soft systems methodology with realist approaches: A novel way to map programme complexity and develop and refine programme theory. *Evaluation*, *24*(1), 84–97. https://doi.org/10.1177/1356389017749036

Daly, A. J. (2012). Data, dyads, and dynamics : Exploring data use and social networks in educational improvement. *Teachers College Record*, *114*, 1–38.

Danermark, B., Ekstrom, M., Jakobsen, L., & Karlsson, J. C. (2002). *Explaining society: Critical realism in the social sciences*. London, UK: Routledge.

Daniell, G. (2018). *Dead ends and doorways: Attainment and equity in upper secondary school qualifications pathways* (Doctoral thesis, University of Auckland, Auckland, New Zealand). Retrieved from http://hdl.handle.net/2292/45083

Datnow, A., Greene, J. C., & Gannon-Slater, N. (2017). Data use for equity: Implications for teaching, leadership, and policy. *Journal of Educational Administration*, *55*(4), 354–360. https://doi.org/10.1108/JEA-04-2017-0040

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*(1), 7–28. https://doi.org/10.1007/s10833-015-9264-2

Datnow, A., & Park, V. (2018). Opening or closing doors for students? Equity and data use in schools. *Journal of Educational Change*, *19*(2), 131–152. https://doi.org/10.1007/s10833-018-9323-6

Datnow, A., Park, V., & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. *Journal of Education for Students Placed at Risk*, *17*(4), 247–265. https://doi.org/10.1080/10824669.2012.718944

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record*, *106*(6), 1145–1176.

Donaldson, S.I., Gooler, L.E., & Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation, 23*(3), 261–273.

Downey, C., & Kelly, A. (2013). Professional attitudes to the use of data in England. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 69–90). Dordrecht, The Netherlands: Springer.

Dunn, K. E., Airola, D. T., Lo, W., & Garrison, M. (2012). What teachers think about what they can do with data: Development and validation of the data-driven decision making efficacy and anxiety inventory. *Contemporary Educational Psychology*, *38*, 87–98.

Dyson, L. (2018). Drawing on theoretical knowledge to build evaluation capacity. *Evaluation Matters—He Take Tō Te Aromatawai*, *4*. https://doi.org/10.18296/em.0032

Dyson, L. (2020a). Data use in New Zealand secondary schools: Tracking, traffic lights, and triage. *Assessment Matters, 14*, 89–111. https://doi.org/10.18296/am.0043

Dyson, L. (2020b). Walking on a tightrope: Agency and accountability in practitioner inquiry in New Zealand secondary schools. *Teaching and Teacher Education*, *93*, https://doi.org/10.1016/j.tate.2020.103075

Earl, L. (2013). Looking forward: Evaluation in New Zealand. In M. K. Lai & S. Kushner (Eds.), *A developmental and negotiated approach to school self-evaluation* (pp. 179–194). Bingley, UK: Emerald Group.

Earl, L., & Katz, S. (2006). *Leading schools in a data-rich world: Harnessing data for school improvement.* Thousand Oaks, CA: Corwin Press.

Earl, L., & Seashore Louis, K. (2013). Data use: Where to from here? In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data based decision making in education: Challenges and opportunities* (pp. 193–203). London, UK: Springer.

Earl, L., & Timperley, H. S. (2015). *Evaluative thinking for successful educational innovation* (Report no. 122). https://doi.org/10.1787/5jrxtk1jtdwf-en

Education Act No. 80. (1989).

Education Central. (2018, April 4). Off the record: The problems with NCEA. https://educationcentral.co.nz/off-the-record-the-problems-with-ncea-spoiler-alert-its-notthe-schools-fault/

Education Council of Aotearoa New Zealand. (2017). *Our code our standards: Code of professional responsibility and standards for the teaching profession*. Wellington, New Zealand: Author.

Education Review Office. (2010). *Promoting Success for Māori Students Schools' Progress*. Wellington, New Zealand: Author.

Education Review Office. (2011). *Directions for learning: The New Zealand Curriculum principles, and teaching as inquiry*. Wellington, New Zealand: Author.

Education Review Office. (2014). *Towards equitable outcomes in secondary schools: Good practice*. Wellington, New Zealand: Author.

Education Review Office. (2015). *Internal evaluation: Good practice*. Retrieved from http://www.ero.govt.nz/publications/internal-evaluation-good-practice/

Education Review Office. (2016). *School Evaluation Indicators.* Wellington, New Zealand: Author.

Education Review Office. (2017). *Briefing to the incoming Minister*. Wellington, New Zealand: Author.

Education Review Office, & Ministry of Education. (2015). *Effective school evaluation: How to do and use internal evaluation for improvement*. Wellington, New Zealand: Author.

Emmel, N. (2013). *Sampling and choosing cases in qualitative research: A realist approach*. Los Angeles, CA: SAGE.

Emmel, N., Greenhalgh, J., Manzano, A., Monaghan, M., & Dalkin, S. (2018). *Doing realist research: Partnership and participation*. Los Angeles, CA: SAGE.

Evans, L. (2011). The 'shape' of teacher professionalism in England: Professional standards, performance management, professional development and the changes proposed in the 2010 white paper. *British Educational Research Journal*, *37*(5), 851–870. https://doi.org/10.1080/01411926.2011.607231

Farrell, C. C., & Marsh, J. A. (2016). Contributing conditions: A qualitative comparative analysis of teachers' instructional responses to data. *Teaching and Teacher Education*, *60*, 398–412. https://doi.org/10.1016/j.tate.2016.07.010

Fetterman, D., Kaftarian, H. J., & Wandersman, A. (2015). *Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability* (2nd ed.). Thousand Oaks, CA: SAGE.

Fullan, M., Rincon-Gallardo, S., & Hargreaves, A. (2015). Professional capital as accountability. *Education Policy Analysis Archives*, 23(15), 1-22.

Gallucci, C., Van Lare, M. D., Yoon, I. H., & Boatright, B. (2010). Instructional coaching: Building theory about the role and organizational support for professional learning. *American Educational Research Journal*, *47*(4), 919–963. https://doi.org/10.3102/0002831210371497

Gan, M., Irving, S. E., & McKinley, E. (2014). Early warning systems in schools: Tracking and monitoring students ' progress using NCEA achievement data. *Set: Research Information for Teachers*, *2*, 54–60.

Gannon-Slater, N., La Londe, P. G., Crenshaw, H. L., Evans, M. E., Greene, J. C., Schwandt, T. A., … Greene, J. C. (2017). Advancing equity in accountability and organizational cultures of data use. *Journal of Educational Administration, 55*(4), 361–375. https://doi.org/10.1108/JEA-09-2016-0108

Gates, E. (2016). Making sense of the emerging conversation in evaluation about systems thinking and complexity science. *Evaluation and Program Planning*, *59*, 62–73. https://doi.org/10.1016/j.evalprogplan.2016.08.004

Gates, E., & Dyson, L. (2017). Implications of the changing conversation about causality for evaluators. *American Journal of Evaluation*, *38*(1), 29–46. https://doi.org/10.1177/1098214016644068

Gibson, R. (1986). *Critical theory and education*. London, UK: Hodder and Stoughton.

Gillborn, D., & Youdell, D. (2000). *Rationing education*. Buckingham, UK: Open University Press.

Grossman, P. L., Smagorinsky, P., & Valencia, S. (1999). Appropriating tools for teaching English: A theoretical framework for research on learning to teach. *American Journal of Education*, *108*(1), 1–29.

Hall, C., & Noyes, A. (2009). New regimes of truth: The impact of performative school self evaluation systems on teachers' professional identities. *Teaching and Teacher Education*, *25*(6), 850–856. https://doi.org/10.1016/j.tate.2009.01.008

Hammersley-Fletcher, L. (2013). Value(s)-driven decision-making. *Educational Management Administration & Leadership*, *43*(2), 198–213. https://doi.org/10.1177/1741143213494887

Hargreaves, A., & Braun, H. (2013). *Data – Driven improvement and accountability*. Retrieved from http://nepc.colorado.edu/publication/data-driven-improvement-accountability/

Harvey, D. (2005). *A brief history of neoliberalism*. Oxford, UK: Oxford University Press.

Hess, F. M. (2009). *The new stupid: Limitations of data-driven education reform*. Retrieved from http://www.aei.org/outlook/100004

Higham, R., & Booth, T. (2018). Reinterpreting the authority of heads: Making space for values-led school improvement with the index for inclusion. *Educational Management Administration and Leadership*, *46*(1), 140–157. https://doi.org/10.1177/1741143216659294

Hill, M. (2016). Inquiry for appraisal. *New Zealand Journal of Teachers' Work*, *13*(1), 6–7. https://doi.org/10.24135/teacherswork.v13i1.92

Hill, M., & Colby, G. (2016). Teacher inquiry through impact projects: One school's journey. *Set: Research Information for Teachers*, *1*, 30–38. https://doi.org/10.18296/set.0034

Hipkins, R., Johnston, M., & Sheehan, M. (2016). *NCEA in context*. Wellington, New Zealand: NZCER Press.

Hofer, S. I., Holzberger, D., & Reiss, K. (2020). Evaluating school inspection effectiveness: A systematic research synthesis on 30 years of international research. *Studies in Educational Evaluation*, *65*. https://doi.org/10.1016/j.stueduc.2020.100864

Honig, M. I., & Ikemoto, G. S. (2008). Adaptive assistance for learning improvement efforts: The case of the Institute for Learning. *Peabody Journal of Education*, *83*(3), 328–363. https://doi.org/10.1080/01619560802222327

Horn, I. S. (2018). Accountability as a design for teacher learning: Sensemaking about mathematics and equity in the NCLB era. *Urban Education*, *53*(3), 382–408. https://doi.org/10.1177/0042085916646625

House, E. (2003). Evaluation theory: Introduction. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation: Part one: Perspectives* (pp. 9–14). https://doi.org/10.1007/978-94-010-0309-4

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

Huffman, D., Thomas, K., & Lawrenz, F. (2008). A collaborative immersion approach to evaluation capacity building. *American Journal of Evaluation*, *29*, 358–368. https://doi.org/10.1177/1098214008318798

Ikemoto, G. S., & Honig, M. I. (2010). Tools to deepen practitioners' engagement with research: The case of the institute for learning. In M. K. Stein & C. E. Coburn (Eds.), *Research and practice in education: Building alliances, bridging the divide* (pp. 93–107). Lanham, MD: Rowman & Littlefield.

Kallemeyn, L. M. (2014). School-level organizational routines for learning: Supporting data use. *Journal of Educational Administration*, *52*(4), 529–548. https://doi.org/10.1108/JEA-02-2013-0025

Karaali, G. (2011). An evaluative calculus project: Applying Bloom's taxonomy to the calculus classroom. *Primus*, *21*(8), 719–731.

King, J. A. (2016, November 2). *Evaluation capacity building through the years* [Webinar]. American Evaluation Association.

Kiro, C., Hynds, A., Eaton, J., Irving, E., Wilson, A., Bendikson, L., … Rangi, M. (2016). *Starpath Phase 2. Final Evaluation Report*. Auckland, New Zealand: The University of Auckland.

Lai, M. (2013). A thousand flowers blooming: The implications of school self-review for policy developers. In M. Lai & S. Kushner (Eds.), *A developmental and negotiated approach to school self-evaluation* (pp. 57–72). https://doi.org/10.1108/S1474-7863(2013)0000014004

Lai, M., & Kushner, S. (Eds.). (2013). *A developmental and negotiated approach to school self-evaluation.* https://doi.org/10.1108/S1474-7863(2013)0000014022

Lai, M., & McNaughton, S. (2016). The impact of data use professional development on student achievement. *Teaching and Teacher Education*, *60*, 434–443.

Lai, M., & Schildkamp, K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 9–21). London, UK: Springer.

Lai, M., & Schildkamp, K. (2016). In-service teacher professional learning: Use of assessment in data-based decision-making. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 77–94). New York, NY: Routledge.

Lai, M. K., & Hsiao, S. (2014). Developing data collection and management systems for decision-making: What professional development is required? *Studies in Educational Evaluation*, *42*, 63–70. https://doi.org/10.1016/j.stueduc.2013.12.006

Lave, J. (1988). *Cognition in practice*. Cambridge, UK: Cambridge University Press.

Lave, J. (1996). The practice of learning. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 3–34). Cambridge, UK: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York, NY: Cambridge University Press.

Leeuw, F. L., & Donaldson, S. I. (2015). Theory in evaluation: Reducing confusion and encouraging debate. *Evaluation*, *21*(4), 467–480. https://doi.org/10.1177/1356389015607712

Lemire, S., Kwako, A., Nielsen, S. B., Christie, C. A., Donaldson, S. I., & Leeuw, F. L. (2020). What is this thing called a mechanism? Findings from a review of realist evaluations. *New Directions for Evaluation*, *167*, 73–86. https://doi.org/10.1002/ev.20428

Lingard, B. (2013). The impact of research on education policy in an era of evidence-based policy. *Critical Studies in Education*, *54*(2), 113–131. https://doi.org/10.1080/17508487.2013.781515

Lingard, B., & Sellar, S. (2013). "Catalyst data": Perverse systemic effects of audit and accountability in Australian schooling. *Journal of Education Policy*, *28*(5), 634–656. https://doi.org/10.1080/02680939.2012.758815

Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record*, *91*, 509–536.

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education*, *118*, 143–166.

Lofthouse, R., & Leat, D. (2013). An activity theory perspective on peer coaching. *The International Journal of Mentoring and Coaching in Education*, *2*(1), 8–20. https://doi.org/10.1108/20466851211231585

MacBeath, J. (2003). *The self-evaluation file. Good ideas and practical tools for teachers, pupils and school leaders.* Retrieved from https://dera.ioe.ac.uk/5951/4/self-evaluation-models-tools-and-examples-of-practice.pdf

MacBeath, J. (2006). *School inspection and self-evaluation: Working with the new relationship*. London, UK: Routledge.

Maguire, M. (2018). International responses: Bob Lingard, Meg Maguire and David Hursh. In M. Thrupp (Ed.). *The search for better educational standards* (pp. 213–229). https://doi.org/10.1007/978-3-319-61959-0_10

Mandinach, E. B., Penuel, W. R., Shepard, L. A., Hamilton, L. S., Miller, S. R., & Gummer, E. S. (2018). *Data-driven decision making: Does it lack a theory of learning to inform research and practice?* Facilitated discussion at the annual meeting of the American Educational Research Association, New York.

Mandinach, E. B., & Schildkamp, K. (2020). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation*. https://doi.org/10.1016/j.stueduc.2020.100842

Manzano, A. (2016). The craft of interviewing in realist evaluation. *Evaluation*, *22*(3), 342–360. https://doi.org/10.1177/1356389016638615

Marchal, B., van Belle, S., van Olmen, J., Hoerée, T., & Kegels, G. (2012). Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research. *Evaluation*, *18*(2), 192–212. https://doi.org/10.1177/1356389012442444

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*(11), 1–48.

Marsh, J. A., & Farrell, C. C. (2014). How leaders can support teachers with data-driven decision making. *Educational Management Administration & Leadership*, *43*(2), 1–21. https://doi.org/10.1177/1741143214537229

Mausethagen, S. (2013). Talking about the test. Boundary work in primary school teachers' interactions around national testing of student performance. *Teaching and Teacher Education*, *36*, 132–142. https://doi.org/10.1016/j.tate.2013.08.003

Mausethagen, S., Prøitz, T., & Skedsmo, G. (2018). Teachers' use of knowledge sources in 'result meetings': Thin data and thick data use. *Teachers and Teaching: Theory and Practice*, *24*(1), 37–49. https://doi.org/10.1080/13540602.2017.1379986

Maxwell, J. (2012). *A realist approach for qualitative research*. Thousand Oaks, CA: SAGE.

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed.). Los Angeles: SAGE.

Maxwell, J., & Mittapalli, K. (2010). Realism as a stance for mixed method research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 145–167). Thousand Oaks, CA: SAGE.

McEvoy, P., & Richards, D. (2006). A critical realist rationale for using a combination of quantitative and qualitative methods. *Journal of Research in Nursing*, *11*(1), 66–78. https://doi.org/10.1177/1744987106060192

McFadden, A., & Williams, K. E. (2020). Teachers as evaluators: Results from a systematic literature review. *Studies in Educational Evaluation*, *64*. https://doi.org/10.1016/j.stueduc.2019.100830

McNamara, G., & O'Hara, J. (2008a). The importance of the concept of self-evaluation in the changing landscape of education policy. *Studies in Educational Evaluation*, *34*(3), 173–179. https://doi.org/10.1016/j.stueduc.2008.08.001

McNamara, G., & O'Hara, J. (2008b). *Trusting schools and teachers: Developing educational professionalism through self-evaluation*. New York, NY: Peter Lang.

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level: From accountability to instructional improvement*. Washington, DC: US Department of Education Office of Planning, Evaluation and Policy Development

Metro's best schools in Auckland 2019: How to pick the right school. (2019, December 19). *Metro*. Retrieved from https://www.metromag.co.nz/society/society-schools/the-best-schools-in-auckland-2019-metro

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A method sourcebook* (3rd ed.). Thousand Oaks, CA: SAGE.

Ministry of Education. (2007). *The New Zealand Curriculum*. Wellington, New Zealand: Learning Media.

Møller, J. (2009). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Educational Change*, *10*(1), 37–46. https://doi.org/10.1007/s10833-008-9078-6

Mutch, C. (2012). Complementary evaluation: The development of a conceptual framework to integrate external and internal evaluation in the New Zealand school context. *Policy Futures in Education*, *10*(5), 569–586. https://doi.org/10.2304/pfie.2012.10.5.569

Mutch, C. (2013). Progressive education in New Zealand: A revered past, a contested present and an uncertain future. *International Journal of Progressive Education*, *9*(2), 98–116.

Mutch, C. (2017). Understanding progressive education and its influence on policy and pedagogy in New Zealand. In C. Mutch & J. Tatebe (Eds.), *Understanding enduring ideas in education: A response to those who 'just want to be a teacher'* (pp. 31–47). Wellington, New Zealand: NZCER Press.

Neuman, S. B. (2016). Code red: The danger of data-driven instruction. *Educational Leadership*, *74*(3), 24–29.

Nusche, D., Laveault, D., MacBeath, J., & Santiago, P. (2012). *OECD reviews of evaluation and assessment in education: New Zealand 2011*. http://doi.org/10.1787/9789264116917-en

Oakden, J. (2014). *If a picture paints a thousand words: The use of rich pictures in evaluation*. Retrieved from http://pragmatica.nz/wp-content/uploads/pragmatica-nz/sites/326/150716-Rich-pictures-in-evaluation-vxx.pdf

O'Brien, S., McNamara, G., O'Hara, J., & Brown, M. (2019). Irish teachers, starting on a journey of data use for school self-evaluation. *Studies in Educational Evaluation*, *60*, 1–13. https://doi.org/10.1016/j.stueduc.2018.11.001

O'Neill, J., & Snook, I. (2015). What will public education look like in the future and why? *New Zealand Journal of Educational Studies*, *50*(2), 195–209. https://doi.org/10.1007/s40841-015-0022-z

Organisation for Economic Co-operation and Development. (2013). *Synergies for better learning: An international perspective on evaluation and assessment.* Paris, France: Author.

Organisation for Economic Co-operation and Development. (2015). *OECD economic surveys: New Zealand 2015*. https://doi.org/10.1787/eco_surveys-nzl-2015-en

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results (Volume I): What students know and can do*. https://doi.org/10.1787/5f07c754-en

Ozga, J. (2009). Governing education through data in England: From regulation to self-evaluation. *Journal of Education Policy*, *24*(2), 149–162. https://doi.org/10.1080/02680930902733121

Ozga, J. (2017). Education policy should not be driven by performance data. *Nature Human Behaviour*, *1*. https://doi.org/10.1038/s41562-016-0014

Patton, M. (2001). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: SAGE.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: SAGE.

Patton, M. Q. (2010). *Developmental evaluation. Applying complexity concepts to enhance innovation and use*. New York, NY: Guilford Press.

Patton, M. Q. (2012). Contextual pragmatics of valuing. *New Directions for Evaluation*, *133,* 97–108. https://doi.org/10.1002/ev.20011

Patton, M. Q. (2015). *Qualitative research and methods: Integrating theory and practice.* (4th ed.). Thousand Oaks, CA: SAGE.

Patton, M. Q. (2018a). A historical perspective on the evolution of evaluative thinking. *New Directions for Evaluation*, *158*, 11–28. https://doi.org/10.1002/ev.20325

Patton, M. Q. (2018b). Evaluation science. *American Journal of Evaluation*, *39*(2), 183–200. https://doi.org/10.1177/1098214018763121

Pawson, R. (2013). *The science of evaluation: A realist manifesto*. Los Angeles, CA: SAGE.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London, UK: SAGE.

Peters, M. (2016). Practice makes mperfect. *New Zealand Journal of Teachers' Work*, *13*(1), 4–5.

Preskill, H., & Boyle, S. (2008). Insights into evaluation capacity building: Motivations, strategies, outcomes, and lessons learned. *Canadian Journal of Program Evaluation*, *23*(3), 147–174.

Preskill, H., & Torres, R. (1999). *Evaluative inquiry for learning in organizations*. Thousand Oaks, CA: SAGE.

Punton, M., & Vogel, I. (2020). Keeping it real: Using mechanisms to promote use in the realist evaluation of the building capacity to use research evidence program. *New Directions for Evaluation*, *167*, 87–100. https://doi.org/10.1002/ev.20427

Punton, M., Vogel, I., & Lloyd, R. (2016). *Reflections from a realist evaluation in progress: Scaling ladders and stitching theory*. Retrieved from https://www.ids.ac.uk/publications/reflections-from-a-realist-evaluation-in-progress-scaling-ladders-and-stitching-theory/

Ranson, S. (2003). Public accountability in the age of neo-liberal governance. *Journal of Education Policy*, *18*(5), 459–480. https://doi.org/10.1080/0268093032000124848

Reid, A. (2004). *Towards a culture of inquiry in DECS*. Adelaide, Australia: Department of Education and Children's Services.

Robertson, S. L. (2008). "Remaking the world": Neoliberalism and the transformation of education and teachers' labor. In M. Compton & L. Weiner (Eds.), *The global assault on teaching, teachers, and their unions stories for resistance* (pp. 11–27). New York, NY: Palgrave Macmillan.

Ryan, K. E., & Timmer, J. D. (2013). New Zealand style school review: A view from outside. In M. Lai & S. Kushner (Eds.), *A developmental and negotiated approach to school self-evaluation* (pp. 195–206). Bingley, UK: Emerald Group.

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland*. New York, NY: Teachers College Press.

Saldaña, J. (2016). *The coding manual for qualitative research* (3rd ed.). Thousand Oaks, CA: SAGE.

Sanderson, I. (2006). Complexity, "practical rationality" and evidence-based policy making. *Policy & Politics*, *34*, 115-132.

Schaffer, E., Reynolds, D., & Stringfield, S. (2012). Sustaining turnaround at the school and district levels: The high reliability schools project at Sandfields Secondary School. *Journal of Education for Students Placed at Risk*, *17*(1–2), 108–127.

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*(3), 482–496. https://doi.org/10.1016/j.tate.2009.06.007

Schildkamp, K., Lai, M., & Earl, L. (2013). *Data-based decision making in education: Challenges and opportunities*. Dordrecht, The Netherlands: Springer.

Schildkamp, K., & Vissher, W. (2010). The utilization of a school self-evaluation instrument. *Educational Studies*, *36*(4), 371–390. https://doi.org/10.1080/03055690903424741

Schmitt, J. (2020). The causal mechanism claim in evaluation: Does the prophecy fulfill? *New Directions for Evaluation*, *167*, 11–26. https://doi.org/10.1002/ev.20421

Schwandt, T. A. (2005). On modeling our understanding of the practice fields. *Pedagogy, Culture & Society*, *13*(3), 313–332. https://doi.org/10.1080/14681360500200231

Schwandt, T. A. (2008). Educating for intelligent belief in evaluation. *American Journal of Evaluation*, *29*(2), 139–150. https://doi.org/10.1177/1098214008316889

Schwandt, T. A. (2014). On the mutually informing relationship between practice and theory in evaluation. *American Journal of Evaluation*, *35*(2), 231–236. https://doi.org/10.1177/1098214013503703

Schwandt, T. A. (2017, April). *The centrality of engagement and ethics to the task of evaluating for equity*. Paper presented at the Educational Review Office Professional Forum, Auckland, New Zealand.

Schwandt, T. A. (2018). Evaluative thinking as a collaborative social practice: The case of boundary judgment making. *New Directions for Evaluation*, *158*, 125–137. https://doi.org/10.1002/ev

Schwandt, T. A. (2019). Post-normal evaluation? *Evaluation*, *25*(3), 317–329. https://doi.org/10.1177/1356389019855501

Schwandt, T. A., & Dahler-Larsen, P. (2006). When evaluation meets the 'rough ground' in communities. *Evaluation*, *12*(4), 496–505. https://doi.org/10.1177/1356389006073745

Selwyn, N. (2014). Data entry: Towards the critical study of digital data and education. *Learning, Media and Technology*, *40*(1), 64–82. https://doi.org/10.1080/17439884.2014.921628

Sinnema, C., & Aitken, G. (2011). Teaching as inquiry in the New Zealand Curriculum: Origins and implementation. In J. Parr, H. Hedges, & S. May (Eds.), *Changing trajectories of teaching and learning* (pp. 29–48). Wellington, New Zealand: NZCER Press.

Sinnema, C., Meyer, F., & Aitken, G. (2017). Capturing the complex, situated, and active nature of teaching through inquiry-oriented standards for teaching. *Journal of Teacher Education*, *68*(1), 9–27. https://doi.org/10.1177/0022487116668017

Smardon, D., & Charteris, J. (2012). Between a rock and a hard place: Teacher professional learning. *New Zealand Journal of Teachers' Work*, *9*(1), 27–35.

Smardon, D., & Charteris, J. (2017). Raising the bar for teacher professional learning and development? Or just cruel optimism? *New Zealand Journal of Educational Studies*, *52*(1), 177–183. https://doi.org/10.1007/s40841-017-0075-2

Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, *118*(2), 113–141. https://doi.org/10.1086/663283

Spillane, J. P., & Miele, D. B. (2007). Chapter 3: Evidence in practice: A framing of the terrain. *Yearbook of the National Society for the Study of Education*, *106*(1), 46–73. https://doi.org/10.1111/j.1744-7984.2007.00097.x

Starkey, L., & Eppel, E. (2019). Digital data in New Zealand schools: Policy reform and school leadership. *Educational Management Administration & Leadership*, *47*(4), 640–658. https://doi.org/10.1177/1741143217745881

Starpath. (n.d.). *The Starpath Project: Raising achievement for students from low socio-economic backgrounds*. https://cdn.auckland.ac.nz/starpath/starpath/casestudynumber1

Stockdill, S. H., Baizerman, M., & Compton, D. W. (2002). Toward a definition of the ECB process: A conversation with the ECB literature. *New Directions for Evaluation*, *93*, 7–25.

Svinicki, M. D. (1999). New directions in learning and motivation. *New Directions for Teaching and Learning*, *80*, 5–27.

Taut, S. (2007). Studying self-evaluation capacity building in a large international development organization. *American Journal of Evaluation*, *28*(1), 45–59. https://doi.org/10.1177/1098214006296430

Thrupp, M. (2005). "Official school improvement" in England and New Zealand: A cautionary comparison. *The New Zealand Annual Review of Education*, *14*, 39–58. https://doi.org/10.26686/nzaroe.v0i14.1487

Thrupp, M. (2017). Introduction—Nine years of National-led education policy. *Waikato Journal of Education*, *22*(1), 3–10.

Thrupp, M. (2018). *The search for better educational standards*. https://doi.org/10.1007/978-3-319-61959-0

Timperley, H. S. (2013). The New Zealand educational context: Evaluation and self-review in a self-managing system. In M. Lai & S. Kushner (Eds.), *A Developmental and negotiated approach to school self-evaluation* (Vol. 14, pp. 23–38). https://doi.org/10.1108/S1474-7863(2013)0000014002

Timperley, H. S., Kaser, L., & Halbert, J. (2014). *A framework for transforming learning in schools: Innovation and the spiral of inquiry*. Melbourne, Australia: Centre for Strategic Education.

Timperley, H. S., & Parr, J. (2008). *Making a difference to student achievement in literacy: Extending our understanding of the Literacy Professional Development Project*. Wellington, New Zealand: UniServices.

Timperley, H. S., Parr, J., & Bertanees, C. (2009). Promoting professional inquiry for improved outcomes for students in New Zealand. *Professional Development in Education*, *35*(2), 227–245.

Timperley, H. S., Wilson, A., Barrar, H., & Fung, I. (2007). *Best evidence synthesis iterations (BES) on professional learning and development*. Wellington, New Zealand: Ministry of Education.

Tomorrow's Schools Independent Taskforce. (2019). *Our schooling futures: Stronger together: Whiria ngā kura tūātinitini: Final report by the Tomorrow's Schools Independent Taskforce*. Wellington, New Zealand: Ministry of Education.

Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry*, *16*(10), 837–851. https://doi.org/10.1177/1077800410383121

Vanhoof, J., Vanlommel, K., Thijs, S., & Vanderlocht, H. (2014). Data use by Flemish school principals: Impact of attitude, self-efficacy and external expectations. *Educational Studies*, *40*(1), 48–62.

Vanhoof, J., & Van Petegem, P. (2013). Judging and explaining the quality of school self-evaluations: Indicators and findings on meta-evaluation from a Flemish perspective. In M. Lai & S. Kushner (Eds.), *A developmental and negotiated approach to school self-evaluation* (Vol. 14, pp. 275–291). Bingley, UK: Emerald Group.

Vanhoof, J., Van Petegem, P., & De Maeyer, S. (2009). Attitudes towards school self-evaluation. *Studies in Educational Evaluation*, *35*(1), 21–28. https://doi.org/10.1016/j.stueduc.2009.01.004

Vanlommel, K. (2018). *Opening the black box of teacher judgement: The interplay of rational and intuitive processes* (Doctoral thesis, Antwerp, Belgium, University of Antwerp). Retrieved from https://doc.anet.be/docman/docman.phtml?file=.irua.4dcaf5.150060.pdf

Vanlommel, K., & Schildkamp, K. (2019). How do teachers make sense of data in the context of high-stakes decision making? *American Educational Research Journal*, *56*(3), 792–821. https://doi.org/10.3102/0002831218803891

Vanlommel, K., Van Gasse, R., Vanhoof, J., & Petegem, P. Van. (2018). Teachers' high-stakes decision making. How teaching approaches affect rational and intuitive data collection. *Teaching and Teacher Education*, *71*, 108–119. https://doi.org/10.1016/j.tate.2017.12.011

Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research*, *83*, 75–83. https://doi.org/10.1016/j.ijer.2017.02.013

Vanlommel, K., Van Gasse, R., Vanhoof, J., Van Petegem, P., Garner, B., Thorne, J. K., … Selwyn, N. (2017). Irish teachers, starting on a journey of data use for school self-evaluation. *International Journal of Educational Research*, *5*(1), 75–83. https://doi.org/10.1177/1474474015575473

Vo, A. T., & Archibald, T. (2018). New directions for evaluative thinking. *New Directions for Evaluation*, *158*, 139–147. https://doi.org/10.1002/ev.20317

Vo, A. T., Schreiber, J. S., & Martin, A. (2018). Toward a conceptual understanding of evaluative thinking. *New Directions for Evaluation*, *158*, 29–47. https://doi.org/10.1002/ev.20324

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wayman, J. C., & Jimerson, J. B. (2014). Teacher needs for data-related professional learning. *Studies in Educational Evaluation*, *42*, 25–34.

Wayman, J. C., Midgley, S., & Stringfield, S. (2006). Leadership for data-based decision making: Collaborative educator teams. *Paper presented at the Annual Meeting of the American Educational Research Association*, San Francisco, CA.

Wehipeihana, N., & McKegg, K. (2018). Values and culture in evaluative thinking: Insights from Aotearoa New Zealand. *New Directions for Evaluation*, *158*, 93–107. https://doi.org/10.1002/ev.20320

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Westhorp, G. (2018). Understanding mechanisms in realist evaluation and research. In N. Emmel, J. Greenhalgh, A. Manzano, M. Monaghan, & S. Dalkin (Eds.), *Doing realist research* (pp. 41–58). Thousand Oaks, CA: SAGE.

Williams, B., & Hummelbrunner, R. (2011). *Systems concepts in action: A practitioner's toolkit*. Palo Alto, CA: Stanford University Press.

Williams, B., & van't Hof, S. (2016). *Wicked solutions: A systems approach to complex problems* (2nd ed.). Retrieved from https://www.lulu.com.

Wilson, A., Madjar, I., & McNaughton, S. (2016). Opportunity to learn about disciplinary literacy in senior secondary English classrooms in New Zealand. *The Curriculum Journal*, *27*(2), 204–228. https://doi.org/10.1080/09585176.2015.1134339

Wohlstetter, P., Datnow, A., & Park, V. (2008). Creating a system for data-driven decision-making: Applying the principal-agent framework. *School Effectiveness and School Improvement*, *19*(3), 239–259. https://doi.org/10.1080/09243450802246376

Wylie, C. (2012). *Vital connections: Why we need more than self-managing schools*. Wellington, New Zealand: New Zealand Council for Educational Research.

Wylie, C. (2020). New Zealand – Steering at a distance and self-managed schools. In H. Ärlestig & O. Johansson (Eds.), *Educational authorities and the schools* (pp. 351–369). https://doi.org/10.1007/978-3-030-38759-4_19

Zeichner, K. (1996). Teachers as reflective practitioners and the democratization of school reform. In K. Zeichner, S. Melnick, & M. L. Gomez (Eds.), *Currents of reform in pre-service teacher education* (pp. 199–214). New York, NY: Teachers College Press.