TEACHERS' BELIEFS ABOUT THE TEACHING, LEARNING AND ASSESSMENT OF MATHEMATICS AND THEIR RELATIONSHIPS WITH BOTH STUDENTS' ACHIEVEMENT, AND TEACHERS' SUMMATIVE JUDGMENTS

BY

SHELLEY VAILE

A thesis
submitted to the Victoria University of Wellington
in fulfillment of the requirements for the degree of
Master of Education

Victoria University of Wellington
(2018)

Acknowledgments

This thesis would not have been finished without the support of my supervisor, Dr Michael Johnston. I am grateful for your advice and guidance and unfailing kindness, especially when life got challenging. I have learnt so much from you over the last two years; many, many thanks.

To all of the teachers who have contributed to this study, thank you so much for participating. This work would not be possible without you.

To my colleagues at Karori West Normal School, thank you for your encouragement, interest and generosity, I learn so much from you all. It is humbling to work with such a dedicated group of education professionals. Special acknowledgement to Janice and Janice for the support and encouragement you have both given me throughout this process. To Jenny, Lynn and Brigit for the many professional discussions and your generosity in offering your time and support whenever I needed to share my thoughts. I love working with you all.

To my friends, Linda, you were there at the beginning. I appreciate your encouragement and your thoughtful, wise counsel. Angela, for those great inspiring discussions, they really helped clarify my thinking. All of my other friends who have helped in any way I look forward to having more time to spend with you soon.

To my lovely family, Mum, Lauren, Pete, Jules and Julie, whether it was a kind word, a gift of chocolate, a text, entering data, or proof reading, your unflagging support and encouragement has kept me going when things got tough. Finally, to Gary, Emily and Jordan, thank you for giving me the space to complete this work. I know it has been tough over the last few months, I dedicate this work to you, I love you very much.

Shelley

Table of Contents

List of Tables

List of Figures

Abstract

Teachers in New Zealand are required to make judgments of students' achievement in reading, writing and mathematics against the National Standards and to report these to the Ministry of Education annually (Ministry of Education, 2009b, 2013). The process for making these judgments is complex and there are many factors that contribute to the variability of teachers' judgments (Smaill, 2013; Ward & Thomas, 2013, 2015). It is likely that the beliefs teachers hold about effective pedagogy in mathematics, about the primary purpose of assessment and about the nature of students' ability in mathematics will contribute to the variability of teachers' judgments against the mathematics standards.

This research explored the beliefs teachers hold about four elements of effective pedagogy in mathematics; the extent to which teachers endorse a discrete approach, a connectionist approach, a procedural approach and a conceptual pedagogical approach. Teachers' beliefs about the primary purpose of assessment and the nature of students' ability in mathematics were also explored.

The purpose of this research was to elucidate relationships between teachers' beliefs and their application of assessment criteria when making summative judgments of students' achievement in mathematics. An additional purpose was to elucidate any relationships between these elements of teachers' beliefs and increased students' achievement and in mathematics.

The degree to which teachers are conservative in their application of assessment criteria when making "at the standard" judgments of students' achievement was found to be related to a coherent set of beliefs. Increased students' achievement was found to be only weakly related to teachers' beliefs about the nature of students' mathematical ability.

Introduction

Primary school teachers in New Zealand are required to make summative judgments of students' achievement against the National Standards in reading, writing and mathematics each year (Ministry of Education, 2009a, 2009b). These standards are descriptive statements of the strategies students use when making sense of texts, when producing written responses and when solving problems and modeling situations in mathematics. The standards are aligned with the New Zealand Curriculum and specify the expected level of students' achievement at each year level. Teachers' summative judgments, or 'overall teacher judgments,' as they are commonly dubbed, are based on their assessments of students throughout the year (Ministry of Education, 2009a, 2009b). Teachers are given the flexibility to decide which pieces of evidence they will use to base their judgments on. There is no national testing regime in New Zealand primary schools.

The National Standards in mathematics, or mathematics standards, cover the breadth of the mathematics curriculum, from the end of students' first year at school until the end of primary schooling in Year 8. According to the Ministry of Education, informal assessment of the solution methods that students use to solve problems as part of the normal classroom programme should form part of the basis of these judgments; however students' achievement in numeracy should be the most critical factor in determining teachers' summative judgments against the mathematics standards (Ministry of Education, 2009a).

Thomas, Tagg, and Ward (2005) explored the reliability of teachers' judgments of students' achievement in numeracy, their study formed part of the annual report and evaluation of the Numeracy Development Projects. The consistency of teachers' judgments of students' achievement against the New Zealand Number Framework were compared to those of experts. The majority of teachers made judgments in agreement with those of experts. Levels of agreement ranged from 72% to 78% over five tasks. Of the teachers whose judgments were not in agreement, two thirds were more conservative than the experts' judgments. The

degree to which teachers are conservative or liberal when making judgments of students' achievement is one important factor when discussing the variability in teachers' summative judgments. The consistency of teachers' judgments is the other important factor.

What teachers believe about the fundamental nature of students' learning processes, about effective pedagogy and about assessment in mathematics may all be related to teachers' application of assessment criteria in making summative judgments of students' achievements in mathematics. A very few research studies (Dweck, Chiu, & Hong, 1995; Wyatt-Smith & Klenowski, 2013) have explored relationships between peoples' beliefs and their judgments. Wyatt-Smith and Klenowski (2013) explored links between the epistemological views held by teachers and their application of assessment criteria. Their research suggested that teachers' epistemological views of the subject they were assessing were related to the process they used to make judgments; with the judgment process used often mirroring their epistemological view of the subject.

Teachers' beliefs about mathematics pedagogy have been the subject of much research (Bahr, Monroe, Balzotti, & Eggett, 2009; Buehl & Fives, 2009; Delandshere & Jones, 1999; Fives & Buehl, 2008; Gill & Hoffman, 2009; Nisbet & Warren, 2000; Peterson, Fennema, Carpenter, & Loef, 1989; Stipek, Givven, Salmon, & MacGyvers, 2001; Wood & Sellers, 1997). Several researchers (Askew, Rhodes, Brown, Wiliam, & Johnson, 1997; Baturo, 2004; Peterson et al., 1989) have explored the relationships between teachers' pedagogical beliefs and students' achievement in mathematics. These researchers have identified sets of pedagogical beliefs that they associated with increased achievement. These included beliefs about students ' ability to learn mathematics (Delandshere & Jones, 1999; Fives & Buehl, 2008; Gill & Hoffman, 2009; Stipek et al., 2001) and teachers' beliefs about assessment in mathematics (Brown & Harris, 2010; Brown, Harris, & Harnett, 2012; Brown, Lake, & Matters, 2011; Suurtamm, Koch, & Arden, 2010). Teachers' assessment beliefs have also been explored, including links between assessment practice (Brown & Harris, 2010) and pedagogical approach to mathematics (Nisbet & Warren, 2000; Peterson et al., 1989).

In this study teachers' judgments of students' achievement in mathematics, as well as their beliefs about effective pedagogy, the purpose of assessment and the nature of students' ability in mathematics were explored. The aim was to elucidate any relationships between teachers' beliefs, their application of assessment criteria when making summative judgments and their students' achievement in mathematics.

National Standards

Over the last twenty years there has been a trend towards standards-based assessment in countries like Australia, Canada, New Zealand, the UK and the USA. Standards are attractive to policy makers because they serve a dual purpose: they are seen as both a means of raising students' achievement, and as a means of ensuring accountability on the part of schools and teachers (Ministry of Education, 2011; Smaill, 2013). One aspect of the New Zealand response to this trend has been to revise the assessment and reporting requirements at the primary level of schooling. The Ministry of Education mandated the implementation of National Standards in reading, writing and mathematics in all English medium primary schools in 2010. These standards are described after the first, second and third year of schooling and subsequently at the end of each year until the end of primary schooling in Year 8. Concurrently, with the implementation of the National Standards, schools' reporting requirements to the Ministry of Education were revised to require primary schools to report student achievement against the Standards each year. More specifically, schools must report the numbers and proportions of their students at each year level who are above, at, below, and well below the National Standards in these three areas (Ministry of Education, 2013). The data are used to inform the allocation of resources by the Ministry of Education, with the aim of supporting schools to raise students' achievement. The aggregated data are also published on the Ministry of Education website (Ministry of Education, 2015b).

Most countries with national standards systems require students to sit state-wide or national standardised tests to gauge students' achievement. The approach taken in New Zealand has been quite different; assessments for

National Standards are based on classroom teachers' overall judgments of each student's achievement for their year level. These are based on a range of formal and informal assessments that have contributed to the teachers' cumulative knowledge of each student's achievement. From this broad range of evidence teachers form an overall judgment of students' achievement. This approach is likely to lead to increased validity when compared with some overseas models requiring students to sit standardised tests. Most jurisdictions in which National Standards have been implemented have relied on a single assessment, which is likely to give an inaccurate indication of individual students' true level of achievement. There are many factors that may impact on a students' performance in a one-off assessment, including anxiety, health and well being on the day of the assessment, and guessing. On the other hand, teachers build up detailed knowledge of students' achievement over an extended period of time, especially when they are using a wide variety of formal and informal assessments, which is likely to invest their judgments with greater validity than a single test.

There is however likely to be much variability between teachers' judgments, which makes the reliability of this approach questionable. Schools are expected to develop moderation processes with the aim of supporting teachers to increase the consistency of their National Standards judgments. Ward and Thomas (2015) reported on their analysis of data from a survey of principals about the moderation practices used in their schools. The majority of principals reported that their schools used systematic processes for the moderation of teachers' judgments in relation to National Standards in the three learning areas. Of concern, however was a finding that a substantial proportion of schools had developed their own resources for teachers to moderate against (Ward & Thomas, 2015). This would likely further reduce the reliability of teachers' judgments, because schools were not moderating against a common set of resources and there is no formal system in place for moderating teachers' judgments across schools. Furthermore Ward and Thomas (2015) found evidence that the reliability of National Standards judgments was low.

As the requirement for teachers to make judgments against National Standards is relatively recent, the process of making judgments is complex, and public confidence in the Standards depends on adequate reliability, this is an important area for research in New Zealand primary schools.

Teachers' summative judgments of students' achievement

It is important to understand the assessment context within which teachers' judgments are made. There are six principles upon which the assessment strategy in New Zealand has been developed: the student is at the centre, the curriculum underpins assessment, building assessment capability is crucial to achieving improvement, an assessment capable system is an accountable system, a range of evidence drawn from multiple sources potentially enables a more accurate response, and effective assessment is reliant on quality interactions and relationships (Ministry of Education, 2011).

According to the New Zealand Curriculum, assessment should be student centered; this requires teachers to choose appropriately the assessments they use to best suit students' individual needs. Teachers are encouraged to choose from the range of assessment tools available and to use informal assessment methods to meet the needs of their students and to capture assessment information from their daily interactions with students. This implies that the tools or strategies used to assess students' understanding of what they have been learning will vary across schools and may even vary within the same classroom. This presents a threat to the reliability of National Standards data as the tools used to measure students' achievement are likely to vary widely.

A broad range of assessment tools are in use in New Zealand schools, across all curriculum areas, but with a particular emphasis on reading, writing and mathematics. The Ministry of Education has focused on building the assessment capabilities of teachers by providing professional development and resources in this area over the last 20 years (Brown & Harris, 2010). Professional development combined with pre-service training could potentially increase the reliability of teachers' judgments; however, this is conditional on coverage of the

teaching population and consistency in the delivery of the professional development.

The Ministry of Education view that an assessment-capable school system increases accountability at all levels of this system (Ministry of Education, 2011, p. 17) is consistent with moves towards increased accountability across the education sector. Smaill (2013) identified increased accountability as one of the two purposes for introducing a National Standards system. Using teachers' own judgments as a basis for an accountability measure potentially presents a threat to reliability because teachers may feel pressure to base their judgments on evidence that shows students' achievement in the best possible light, rather than necessarily presenting the most accurate, on-balance reflection of students' achievement.

Teachers' differing interpretations of assessment criteria are likely to lead to assessing different aspects of students' achievement, which, again, presents a threat to the reliability of teachers' judgments. This was evident in the analysis by Ward and Thomas (2013), of teachers' ratings of the importance of various assessment tools when making summative judgments of students' achievement in mathematics. For example, the IKAN tool was rated as either moderately or highly important by 42% of the teachers surveyed. This tool specifically assesses students' number knowledge and, as such, is largely irrelevant when making judgments about students' ability to solve problems. One possible explanation for teachers rating this assessment tool so highly is that they might have misinterpreted the assessment criteria, focusing only on measuring what students know, rather than on how they apply that knowledge in solving problems.

Teachers are required to aggregate the information from a range of tools to form their overall judgments. They might differ in the weightings they assign to different sources of assessment information and may combine the different pieces of assessment information in different ways. It is likely that this approach presents an additional threat to the reliability of teachers' judgments; teachers

base their summative judgments on the ongoing assessments they have made of students' achievement throughout the year.

There is a growing body of research on teachers' judgments of students' achievement (Brookhart, 2013; Connolly, Klenowski, & Wyatt-Smith, 2012; Cooksey, Freebody, & Wyatt-Smith, 2007; Crisp, 2013; Klenowski, 2013; Sadler, 2009; Wyatt-Smith & Klenowski, 2013). In her review of research into teachers' judgments, Brookhart (2013) found that much of the research focused on describing variability in teachers' judgments. Included in her review were several recent research studies, which suggested that teachers may not typically distinguish between achievement and other non-achievement factors like effort when making their judgments.

Teachers application of assessment criteria when making summative judgments

Sadler (2009) described some of the difficulties when assessment criteria are presented as descriptive statements as is the case with the National Standards. Such statements are open to differences in interpretation, which are likely to increase the variability of teachers' judgments.

Selections of annotated exemplars of students' solution methods accompany the assessment criteria for the National Standards in mathematics at each year level. There are additional exemplars on the Ministry of Education sponsored websites (Ministry of Education, 2010). These provide teachers with examples of students' work and describe the elements inherent in each sample that are commensurate with particular assessment criteria. These are designed to support teachers to make consistent judgments. Ward and Thomas (2013) reported that 69% of teachers in their sample used these exemplars when making judgments against the National Standards in mathematics.

Teachers are instructed to make judgments based on students' solution methods employed "independently" and "most of the time" (Ministry of Education, 2009a, p. 12). This adds to the complexity of the judgment process. Teachers' must not

only determine if students' solution methods are consistent with those specified in the assessment criteria, they are required to make decisions about students' levels of independence and consistency when solving similar problems.

Under the National Standards system teachers are expected to aggregate multiple sources of assessment evidence to make summative judgments. Teachers are advised that students' achievement in number is the most critical component in meeting the standard (Ministry of Education, 2009a, p. 12). Teachers must factor this into their relative weighting when aggregating students' achievement against the assessment criteria. This specification adds to the complexity of the teachers' judgment process.

Several research studies (Connolly et al., 2012; Wyatt-Smith & Klenowski, 2013) have found differences in teachers' application of assessment criteria depending on the subject being assessed. Teachers of mathematics and the applied sciences took a more analytical approach than their colleagues assessing language, social science and practical subjects who were more likely to adopt a global or holistic approach to judging students' achievement. The teachers of mathematics and applied sciences made judgments against each of the assessment criteria then combined each aspect (usually additively) to form an overall judgment. Teachers of language, social sciences and practical subjects made a global judgment on the piece of work as a whole without strictly adhering to the stated criteria. This would seem to suggest that the judgments of the teachers of mathematics and applied sciences may be more reliable of the two groups of teachers.

Possible causes of variability

Recent research (Connolly et al., 2012; Crisp, 2013; Wyatt-Smith & Klenowski, 2013) has identified a variety of factors contributing to variability in teachers' judgments. One of these factors was the context in which the judgments are made. Different elements of context have been studied (Connolly et al., 2012; Crisp, 2013; Wyatt-Smith & Klenowski, 2013): the subject area, the school and the classroom have been shown to contribute to variability in teachers' judgments. Crisp (2013) reported findings that suggest teachers' judgment processes vary according to the subject being assessed. Crisp's study was

however focused on secondary teachers who are subject specialists, this may, or may not hold true for primary teachers who are much less likely to have specialist knowledge of the subject being assessed.

Previous experience in making judgments has been shown to reduce variability (Connolly et al., 2012; Crisp, 2013), with teachers' judgments becoming more consistent as they have the opportunity to practice. The effect of experience is enhanced when teachers have been provided with the opportunity to collaborate in making judgments, and have been supported to explain and justify their judgments through social moderation. However these reductions in variability are conditional on the quality of the feedback received (Smaill, 2013). The effectiveness of social moderation has been related to the expertise of the participants. Recent research (Crisp, 2013; Wyatt-Smith & Klenowski, 2013) has suggested that teachers' judgments become more consistent over time as they have opportunities to practice and refine their skills. Teachers who are expert have been found to make judgments that are more consistent than those of novices. The consistency of novice teachers' judgments has been significantly increased when provided the opportunity to collaborate with and receive feedback from experts.

Different teachers may interpret the information they get from an assessment tool differently, especially if they are using a tool that is new to them. This may well lead to teachers making different teaching decisions. In their study exploring the reliability of teachers' judgments of students' achievement in numeracy, Thomas et al. (2005) found that two thirds of the teachers whose judgments differed from those of experts rated students numeracy achievement below the experts' ratings. In many of these cases teachers explained that they had rated the students' lower as they wanted to consolidate students' understanding at the current level before moving them up to the next instructional group.

Smaill (2013) discussed some possible causes of variability in teachers' summative judgments of students' achievement in relation to the National

Standards. First, teachers are required to make judgments on a range of evidence from a variety of sources. Second, teachers require high levels of assessment literacy when analysing and comparing the information from different assessments, especially when the information is contradictory. Third, the solution methods employed by students to solve problems are the basis upon which judgments are made and assessment tools vary in their suitability for evaluating students' solution methods.

Beliefs

The identification and measurement of pedagogical beliefs held by teachers of mathematics has been the focus of numerous studies (Bolden & Newton, 2008; Buehl & Fives, 2009; Fives & Buehl, 2008; Nisbet & Warren, 2000). These studies have focused on describing the beliefs held by different groups of teachers, comparing and contrasting the beliefs

Ultimately of primary importance in the study of teachers' beliefs is their relationship with students' achievement. Sets or systems of beliefs held by teachers that have been associated with increased student achievement have been described in many studies (Askew et al., 1997; Bahr et al., 2009; Peterson et al., 1989; Wood & Sellers, 1997).

Much research into teachers' beliefs describes findings in terms of associated sets or systems of beliefs held by teachers (Delandshere & Jones, 1999; Gill & Hoffman, 2009; Peterson et al., 1989; Rubie-Davies, Flint, & McDonald, 2011; Song & Looi, 2012; Stipek et al., 2001). Pajares (1992) also discussed the importance of thinking about beliefs in terms of connections among beliefs rather than as independent subsystems

Teachers' beliefs about the primary purpose of assessment in mathematics

The New Zealand Curriculum advocates an assessment-for -earning approach (Brown & Harris, 2010; Ministry of Education, 2007, 2009a, 2011; Smaill, 2013). The focus of this approach is on teachers' ongoing use of formative assessment to identify students' learning needs, and to inform teachers' pedagogical decisions of

how best to address these needs. Brown and Harris (2010) noted that over the last 20 years the Ministry of Education has devoted significant resources to the development of assessment tools and professional development focused on teachers using assessment formatively to improve learning. These developments include the Assess to Learn (ATOL) professional development programme and the development of assTTle and e-asTTle, which are both computer–assisted assessment tools.

An important element of formative assessment is the feedback given to students by teachers about their learning and identifying their next steps. Hattie and Timperley (2007) discussed the importance of feedback for enhancing students' achievement in their review of research in this area. The differential effects of the quality and nature of feedback were explored, addressing issues like the timing of feedback and the level of the feedback, whether it is related to the task, the process, regulatory or personal issues. Brown et al. (2012) explored New Zealand teachers' conceptions of feedback and the practices that they associated with feedback. The findings suggest that both primary and secondary teachers strongly supported the notion that feedback is to improve learning.

Summative assessment is assessment carried out for reporting purposes. The practice of teachers assessing what students have learnt at the completion of a unit of work, in order to evaluate students' level of achievement is summative. The timing of the assessment is one of the key features of summative assessment. This is most often at the end of a unit of work or towards the end of the year. A common misconception is that the assessment tool used is what distinguishes summative assessment from formative assessment. However, the assessment tool used does not determine whether the assessment is summative or formative, but rather the purpose of the assessment. Tools designed to be used formatively can be used for reporting purposes and vice-versa. In their study of the assessment practices in Ontario, Suurtamm et al. (2010) reported that teachers predominantly used paper and pencil tests, quizzes and performance tasks for summative assessment purposes (determining a mark on a report card). The

same teachers reported using a wider variety of assessments including paper and pencil tests, quizzes and performance tasks for formative assessment purposes.

Teachers' beliefs about effective assessment have been the focus of recent research (Brown & Harris, 2010; Brown et al., 2012; Brown et al., 2011; Suurtamm & Koch, 2014; Suurtamm et al., 2010). Relationships have been found between these beliefs and teachers' assessment practices. Brown and Harris (2010) discussed assessment in terms of four major purposes: improved learning and teaching, school evaluation, student evaluation, and irrelevance. Relationships have been shown between these purposes of assessment and teachers' beliefs about the nature of teaching, learning and curriculum. Two of these purposes align with the purposes of formative and summative assessment, assessment to improve teaching and learning is consistent with the purpose of formative assessment and student evaluation is consistent with the purpose of summative assessment. The degree to which teachers endorse the primary purpose of assessment in mathematics as either formative or summative will therefore very likely relate to their assessment practices and their pedagogical beliefs.

Teachers' beliefs about effective pedagogy when teaching mathematics. A pedagogical approach that is connectionist or more discrete.

The term *connectionist* was first used by Askew et al. (1997) to describe the pedagogical approach taken by a group of teachers identified as being effective teachers of numeracy. These teachers explicitly linked ideas and concepts in their pedagogy. This approach is advocated in the New Zealand Curriculum (Ministry of Education, 2007, 2009a): teachers are advised that it is important to connect the ideas within and across the three strands of the mathematics curriculum (Ministry of Education, 2007, 2009a). Making connections within the strands refers to teachers making explicit the relationships between the students' prior learning in areas related to the new learning. An example of this would be linking multiplication with students' prior knowledge of addition as well as their prior knowledge of multiplication, from this basis students build new learning about multiplication. Making connections across the strands requires teachers to

explicitly link students' knowledge across different strands of the mathematics curriculum. Using the example of teaching multiplication, teachers explicitly link their teaching to the calculation of area and volume for different shaped spaces from the measurement and geometry strand of the curriculum. Anthony and Walshaw (2007) discussed the importance of teachers having knowledge and awareness of the conceptual connections inherent in the levels of the mathematics curriculum that they taught.

The term *discrete* pedagogical approach has been used throughout this research to describe the approach of presenting mathematical ideas separately. Askew et al. (1997) categorised aspects of both a transmission orientation and a discovery orientation to the teaching of numeracy as presenting mathematics in discrete packages with different concepts taught separately. Teachers taking a discrete pedagogical approach would for example teach multiplication separately from addition. Delandshere and Jones (1999) found that the teachers in their study took a discrete pedagogical approach to the teaching of mathematics. They described the teaching of one of the teachers in terms of "a set of disconnected procedures or specific tasks." (Delandshere & Jones, 1999, p. 231).

The degree to which teachers believe that an effective pedagogical approach explicitly connects concepts in mathematics for students or conversely presents the concepts as discrete components of mathematics will very likely be related to how they teach mathematics and may therefore also be related to their students' achievement.

A pedagogical approach that is conceptual or more procedural

A conceptual approach to the teaching of mathematics is centered on the development of students' understanding of mathematical ideas. This approach emphasises students developing their own solution methods to solve problems as it is argued (Beishuizen & Anghileri, 1998) that this allows students to solve more complex problems and to develop a deeper understanding of the principles underlying the mathematics they are using. Students' are encouraged to share their strategies and solution methods and to justify their choice of strategy based

on the effectiveness for the particular problem. This approach was central to the Numeracy Development Project (Numeracy Professional Development Projects, 2008), a countrywide professional development project based on the developmental framework of students' acquisition of number knowledge and use of increasingly complex strategies for solving number problems. This framework forms the basis of the number component within the New Zealand Curriculum. The framework had dual aims of increasing students' achievement by developing teachers' content knowledge and pedagogical approach to the teaching of number. Multiple teaching resources have been developed to support the use of a conceptual approach to the teaching of all aspects of mathematics.

A procedural approach focuses on learning conventional mathematical procedures to solve problems. This approach emphasises the importance of students developing mastery of set procedures and learning which procedures to apply to different types of problems. A procedural approach values students' adherence to the learnt procedures regardless of the efficiency of that solution method. Gill and Hoffman (2009) also found evidence of a procedural approach as the teachers in their study thought it was important to teach the rules first especially prior to attempting problem solving. Rittle-Johnson and Koedinger (2009) found that an iterative approach to teaching concepts and procedures lead to increased student achievement when solving novel tasks then an approach where the concepts were taught separately from the procedures.

Teachers' beliefs about the effectiveness of a pedagogical approach that is conceptual or procedural will very likely shape the approach they take when teaching mathematics. Peterson et al. (1989) found that teachers with conceptual pedagogical beliefs tended to their use of a more conceptual approach to teaching mathematics and this was related to increased students' achievement on problem solving tasks. This is an important area of research as teachers adopting a more conceptual approach has been linked with increased student' achievement when solving complex mathematical problems.

Teachers' beliefs about students' ability to learn mathematics

Implicit theories of ability are the theories that individuals hold about their own or others' ability. Implicit theories are described by two opposing views. The degree that people believe ability is fixed, stable and largely unchanging is referred to as an *entity* theory. Conversely the degree to which people believe that ability is malleable, changeable, able to be mediated, is referred to as an *incremental* theory. Much research has been conducted exploring the implicit theories people hold (Dweck et al., 1995; Garcia-Cepero & McCoach, 2009; Ilhan & Cetin, 2013; Jones, Bryant, Snyder, & Malone, 2012; Jones & Egley, 2007; Jonsson, Beach, Korp, & Erlandson, 2012) about the nature of attributes ascribed to themselves and others. These attributes include morality, intelligence and honesty. The implicit theories held by teachers have been found to be related to the teaching decisions they make.

Teachers' implicit theories of students' mathematical ability have been inked to their pedagogical beliefs in many studies (Delandshere & Jones, 1999; Fives & Buehl, 2008; Gill & Hoffman, 2009; Stipek et al., 2001). These studies consistently reported that entity beliefs are related to procedural and discrete pedagogical beliefs, which were found to be consistent with a more transmission orientation to the teaching of mathematics. Conversely teachers holding incremental views of students' ability in mathematics were found to be related to conceptual pedagogical beliefs and these in turn have been linked with teachers using a problem solving approach to their teaching of mathematics.

Dweck et al. (1995), in their review of research on implicit theories, found that the implicit theories held by people influence their inferences, judgments and reactions. When exploring teachers' judgments of students' achievement, it is likely that these judgments are related to the implicit theories they hold about students' ability. This an important dimension to explore when researching teachers' judgments and their relationship to teachers' beliefs.

What relationships are there between teachers' application of assessment criteria against the National Standards in mathematics, either conservative or liberal, and each of four dimensions of teachers' beliefs: a pedagogical belief that a

connectionist approach is effective for teaching mathematics, a pedagogical belief that a conceptual approach is effective for teaching mathematics, an incremental view of students' mathematical ability, and a belief that assessment is primarily for formative purposes?

A supplementary question was also explored.
What relationships are there between students' achievement in mathematics and the four elements of teachers' beliefs: a pedagogical belief that a connectionist approach is effective for teaching mathematics, a pedagogical belief that a conceptual approach is effective for teaching mathematics, an incremental view of students' mathematical ability, and a belief that assessment is primarily for formative purposes?
A correlative approach was taken to answer these two questions.

Literature Review

The research on teachers' beliefs can be categorized as descriptive (research focused on describing or measuring teacher's beliefs); explanatory (research focused on how beliefs are clustered together); correlative which elucidates the relationships between beliefs and other factors including student achievement; and change which identifies the process that has led to a change in beliefs. These categories will be used throughout this review.

The literature on teachers' summative judgments has been traditionally focused on describing the variability in teachers' judgments according to the review by Brookhart (2013). Two major themes ran through the research; teachers tended to mix student achievement with non-achievement factors like effort and work habits when making summative judgments, and the high levels of variability in teachers' judgments when compared to standardised tests. Martinez, Stecher, and Borko (2009) used data from a large scale, national survey in the USA to investigate teachers' judgments of students' achievement and to explore variability in teachers' judgments when compared to standardised test scores. Their findings suggested that teachers used comparison between students to make judgments of students' achievement relative to that of other students rather than in absolute terms.

Research exploring teachers' summative judgments of students' achievement

Allal (2013) used interviews and analysis of the assessment documents of 10 sixth-grade teachers in Switzerland to explore the role of their professional judgments in assigning end of year grades in mathematics. The study was based in Geneva, where end of year grades are one of the main determinants of whether a student is offered academic or vocationally oriented secondary schooling. All participating teachers were from different schools. Two semi-structured interviews were conducted with each teacher at the end of the second term and the end of the school year respectively. Teachers provided all of their assessment evidence for two students who were at the borderline between an achieved or not achieved grade. These students were the focus of the interviews.

The interviews were recorded and later transcribed, analysed and coded. Information from the assessment artifacts were also analysed and incorporated into the coding system. The primary researcher coded all of the transcripts and artifacts; a second researcher reviewed these. Inconsistencies or anomalies in coding were discussed and recoded until consensus was reached.

The research found that some schools had developed their own policies on making summative judgments, specifying which pieces of evidence to use and how to aggregate that evidence. Teachers followed these policies for all students, with the exception of students at or below the borderline. For borderline students, teachers sought additional information that varied in each case, in terms of the source and nature of the evidence. While they worked collaboratively to prepare assessments, the majority of the teachers did not collaborate when making judgments. This study provides a rich description of this context however the findings cannot be generalised, as they are specific to the teachers in the study. They do however illustrate the complexities of making summative judgments of students' achievement and highlight some of the non-assessment variables that teachers may take into account when making judgments. The high-stakes nature of the consequences of these judgments make the context quite different from the judgments against the National Standards teachers in New Zealand are required to make.

Cooksey et al. (2007) used a quantitative approach to contrast teachers' judgments of students' writing using both teachers' own assessment systems and a national system of benchmarks as mandated by the state of Queensland. The research aimed to answer a series of related questions. How do teachers go about their assessment tasks using either their own native assessment systems or a national system of benchmarks mandated by the state? Are such judgments defensible and reliable? Are such judgments in any way comparable (correlated)? How are those judgments influenced by contextual circumstances surrounding the production of the text? The sample of teachers (n=20) all taught Year 5 students from a range of schools within the Brisbane metropolitan area. Each teacher made judgments on a set of up to 25 pieces of students' writing

from their own class and a common set of 25 pieces of writing supplied by the research team. Teachers made judgments on the two sets of students' writing using their own assessment system, assigning grades from 1 (poor level of achievement) to 5 (excellent level of achievement). This scale was designed to be analogous to the type of informal grading scale used for classroom assessments. They then made a judgment for each piece of work against the national benchmark standards assigning grades from 1 (below benchmark standard) to 3 (above benchmark standard).

To identify the cues that teachers used to influence their judgments a think-aloud protocol was followed. This protocol recorded teachers verbalising their thought processes as they made judgments of students' achievement. From these recordings, the cues each teacher used when making judgments were identified. These cues were coded and grouped together into categories that were combined into macro themes resulting in four feature cues. The cues were log-transformed to normalize their distribution.

All of the samples of writing were independently analysed to identify a set of cues likely to be used to make judgments based on the linguistic and textual features of the samples. The text feature cues were analysed and all cues that were identified as being potentially relevant to the judgment process were coded. Codes for all 520 samples of writing were analysed using principal components analysis with promax oblique rotation. Five components were identified and, with the exception of one virtually independent component, the other components were all moderately correlated. These five text feature cues and four think aloud feature cues were used in the judgment analysis.

Judgment analysis utilized retrospective modeling of judgments using statistical procedures such as multiple regressions to capture the teachers' judgment policy and their consistency in applying this. The teachers' judgments of the writing samples were statistically analysed, the cues associated with each sample were used to predict the teachers' judgments using multiple regression analysis. The regression model represented the teachers' judgment policy. The multiple correlations (R), (the correlation between the actual judgments and the judgments predicted by the policy) reflected the consistency with which each

teacher made judgments according to their judgment policy. Two statistical models were created for each teacher based on the judgments made about writing samples using their own assessment system and the judgments made using the national benchmark system. Once these had been calculated the two judgment systems were compared for each teacher.
The judgments teachers made using their own systems of assessment were analysed. For all 20 teachers, the cues identified in the think aloud protocol explained significant judgment variance for writing samples using their own assessment system. For 55% of the teachers the cues from text feature analysis contributed significantly to the explanation of judgment variance using their own assessment system. The models strongly captured the judgment policy using their own assessment system of 15% of the teachers.

The judgments teachers made using the national benchmark system were analysed. For 55% of the teachers the text feature cues contributed significantly to the explanation of benchmark judgment variance. For 85% of the teachers the think aloud cues explained significant benchmark judgment variance. Comparatively the think aloud feature cues explained less of the variability in the benchmark judgments, with the text feature cues explaining more of the variability for these judgments.

This research identified the importance of teachers using their own judgment processes in the assessment of students' writing for this small, unrepresentative, sample of teachers. This study has high validity and is relevant to the current study even though it is based on judgments of writing samples this is analogous to assessing students' work samples when solving complex problems or using an investigative approach in mathematics. This research is very complex and identifies some of the complexities and competing tensions of teachers' judgment processes.

Both of these research studies identified some of the complexities associated with teachers making summative judgments of students' work. This is especially true when the work being assessed is complex as is the case for primary teachers

in New Zealand when making summative judgments of students' achievement covering the breadth of the mathematics curriculum.

Teachers' judgments against assessment criteria

Teachers are commonly required to make judgments against a specified set of assessment criteria. Recent research by Sadler (2005, 2009) into the use of assessment criteria has been critical of this approach to assessment and offers alternatives. Sadler (2005) explored the use of assessment criteria when making judgments of students' achievement in the context of higher education. He reviewed the nature of criteria-based grading policies and practices from nearly 200 institutes of higher education from across Australia, New Zealand, The USA, The UK, Canada and South Africa. His findings suggest that there is no common understanding of what criteria-based assessment is and argues that standards-based assessment would provide a more suitable alternative. Sadler (2009) discusses the multiple criteria based assessment schemes used throughout tertiary institutes. He identifies anomalies associated with the use explicit grading models and discusses an alternative that is based on a more holistic approach to making judgments of students' achievement.

In their qualitative study, Wyatt-Smith and Klenowski (2013) explored the nature of the assessment criteria used by teachers to make judgments of students' achievement against standards in mathematics and English. The standards are specified by the Queensland Curriculum, Assessment and Reporting Framework, and require teachers to assign a grade (A-E) to indicate students' level of achievement. Wyatt-Smith and Klenowski interviewed 164 teachers of Year 4, Year 6 and Year 9 students from across Queensland and observed moderation meetings (89 teachers from 49 schools). Transcripts of all meetings and interviews were analysed using a constant comparative method to identify emergent themes. This method requires the researcher to identify themes from multiple readings of the transcripts, while constantly comparing and refining the themes until the themes adequately captures the information from the data. When multiple researchers are involved this process continues until consensus is reached. They found that the teachers' approaches to

assessment varied according to the subjects they were making judgments on. Mathematics teachers favoured a rational or analytic approach; that is, a prescriptive, or reductionist approach to assessment. These teachers made judgments of students' achievement solely on the elements specified in the criteria. Contrasted with this was a more global or holistic approach taken by teachers of English. For these teachers, each student's work was judged as a whole. Teachers drew on latent criteria (not specified in the assessment criteria but present in the piece of work) also meta-criteria (knowing when to use or disregard both explicit and latent criteria when making judgments of students' achievement). This approach is likely to increase the validity of the judgments because the teachers respond to the features present in each piece of work, regardless of whether those features were specified in the explicit criteria. This approach is likely to reduce the reliability of the judgments because the judgment becomes subjective, the features that assessors make their judgments on are likely to vary as is the weighting given to the latent criteria.

This study used two methods of data collection and a relatively large sample size, both of which add to the validity of the research. Differences in how teachers use assessment criteria when making judgments were described. From the differences between the teachers of different subjects identified it is likely that the judgments made by the mathematics teachers would be more reliable and consistent than those made by the English teachers.

Crisp's (2013) qualitative research into internally assessed components of the General Certificate of Secondary Education (GCSE) qualification in England explored the extent to which teachers used comparison, the role of assessment criteria and of past experience when making judgments of students' achievement. The research was based on interviews with 13 teachers on coursework in three subject areas (English/English Literature, Information and Communication Technology (ICT), and Geography), followed by a wider survey of teachers (n=378) across a wider range of subjects.

Teachers were interviewed either individually or in small groups using a semi-structured interview format asking teachers questions about five broad themes

associated with their judgment process. The data analysis procedure was not discussed and findings were presented as themes contrasting teachers' judgment practices across the three subject areas. This was supported with anecdotal comments.

The questionnaire collected data using several different methods. Some questions required respondents to select the most appropriate response from a 5-point Likert scale of frequency. Also included were questions requiring a yes or no answer and open-ended questions. The responses to these items were presented in tables showing the percentage of endorsement. These data were also presented as stacked bar graphs, which were used to visually compare the distribution of responses. This was the extent of the analysis of the survey data.

This research suggested that context was an important factor in judgments of students' achievement. The degree to which comparison was also important varied across subjects, with English teachers stating that comparison was an important part of the process more frequently than teachers of other subjects. The extent to which teachers relied solely on explicit assessment criteria when assessing students' work was also explored. The majority of the teachers interviewed reported that their judgments were based on the explicit assessment criteria, however teachers also reported using unspecified, latent criteria when making their judgments.

The final component explored was the extent to which previous assessment experience contributed to the judgment process. Many of the more experienced teachers reported that, over time, they had built up mental representations of the grades, which they consciously referred to when assessing students' work. Teachers identified several factors as important for increasing their confidence and consistency when making judgments of students' achievement. These included the opportunity to practice and build experience of assessing, working collaboratively and being involved in moderation, receiving training and feedback on their assessment practices.

This research provides a rich description of teachers' judgment-making processes in this particular context. Incorporating data from two sources increased the internal validity of the research: however, there is no explanation of the qualitative methods used to analyse the interview data. There was no statistical analysis of the survey data, relying solely on the distribution of the raw data to make comparisons. This calls into question both the reliability and validity of these findings.

The research of Wyatt-Smith and Klenowski (2013) and of Crisp (2013), attempt to describe teachers' use of assessment criteria when making judgments of students' achievement. The concept of the assessment criteria used by teachers being augmented by unspecified latent criteria was a feature of both studies, drawing on the work of Sadler (2009). There is a lack of quantitative analysis of teachers' application of assessment criteria when making summative judgments.

Variability in teachers' summative judgments

Several research studies of teachers' judgments (Allal, 2013; Connolly et al., 2012; Cooksey et al., 2007; Crisp, 2013; Martinez et al., 2009; Wyatt-Smith & Klenowski, 2013) found that context was a factor that contributed to the variability in teachers' judgments. Teachers made judgments about students' achievement within their own educational settings and contexts. Many factors are inherent in these contexts including the schools demographic profile, the makeup of the class, class size, grouping decisions at the school and class level and even the subject that the judgments are being made in.

Martinez et al. (2009) carried out a secondary analysis of the data set generated from the Early Childhood Longitudinal Survey, Kindergarten class of 1998-1999 (ECLS-K) conducted by the National Center for Education Statistics (US Department of Education). They used a combination of descriptive and correlative statistical approaches to explore relationships between teachers' judgments of achievement and students' achievement in mathematics measured by standarised tests. Teachers also reported on the types and frequencies of assessments they used as part of their classroom programmes and the

importance they gave to each of the assessments. Two independent measures of students' achievement were collected, one using standardized tests and the other based on teachers' judgments. The research focused on mathematics achievement in grade 3 (n=10,700) and grade 5 (n=8,600). Item response theory was used to calibrate scores onto a scale location, thus allowing longitudinal interpretations across grades. Teachers rated students' skills and knowledge on a five-point criterion-referenced scale of proficiency. These were scaled (within each grade only) using a Rasch model calibrating a scale location for each score. Using an unconditional hierarchical linear model, patterns of variation of tests and teachers' ratings were calculated across classrooms and schools. Teachers' judgments were found to correlate strongly with standardized test scores. One third of the variability was explained at the classroom level and very little at the school level. The findings suggest that teachers may be judging their students' achievement by adopting a school-specific norm referencing approach, rather than a criterion-referenced approach. Teachers tended to judge the achievement of students with special needs lower than their achievement on standarised tests; conversely teachers' judgments of the achievement of girls, minority groups, students from low socio-economic background and second language learners were judged higher, on average, than their achievement on standardised tests. The researchers suggested that this may be due to teachers attempting to compensate for students' difficulties. This research gained much statistical power from the large numbers of students and teachers involved in the (ECLS-K) giving the findings high reliability. As this was a descriptive and correlative study no causal statements could be made.

Connolly et al. (2012) used a qualitative approach to focus on teachers' beliefs and attitudes to standards based moderation. The relatively large sample of Year 4, Year 6 and Year 9 teachers (n=67), from 24 schools, was representative of the population of schools in Queensland on cultural, socio-economic, school type, rural or urban factors. The teachers in the sample assessed their students' work in English, mathematics and science against standards specified by the Queensland Curriculum, Assessment and Reporting Framework. Teachers assigned grades (A-E) to indicate students' levels of achievement. They were

interviewed pre- and post-moderation to identify the extent to which they believed that standards-based assessment coupled with social moderation would increase the consistency of teachers' judgments of students' achievement. This process was repeated the following year giving a total of 113 interviews. The data were transcribed, then coded, following a formal inductive process using a pre-determined set of descriptors. The data were then checked, compared and analysed to identify themes aligned to the research questions. Participants were categorized as having positive, negative or neutral attitudes towards using standards in moderation to increase consistency. A general trend indicating increasingly positive attitudes towards social moderation was reported after the teachers had experienced the moderation process. None the less teachers expressed a number of caveats: that there needed to be additional support by way of exemplars of student work at different grade levels; the exemplars needed annotations explaining the judgment process in assigning a grade; and teachers required training and opportunities to engage in the social moderation of their judgments.

A small part of the study conducted by Connolly et al. (2012) used a sub-sample of teachers (n=20) from nine schools who were purposively selected to be demographically representative of Queensland schools. These teachers all supplied multiple samples of student's work that they had assessed and graded (A-E) against the standards. The work selected included equal proportions of students' work from three subject areas, mathematics, English and science. The teachers were also randomly assigned the work from each other's students to grade, so that the teacher whose class the work was from and one other teacher graded each sample of work. Considerable variability in teachers' judgments in the three subject areas was shown. Only 35% of grades were totally consistent (both assessors assigned the same grade to the piece of work), 51% of responses differed by one grade, 12% of responses differed by two grades and 2% of responses differed by three grades. With the teachers' judgments of work samples from the subject area English most consistent and those work samples from the science area least consistent. The researchers suggested that the context (subject) in which the judgment was made was a significant factor when

considering variability in teachers' judgments. This small study was within a much larger qualitative study. There was no discussion of how significant the differences in grades were; as such both the validity and reliability of this portion of the research are in question.

Wyatt-Smith and Klenowski (2013) discussed the cumulative nature of judgments, with assessors drawing on previous experience when making judgments of students' achievement. Many research studies suggest differences between novice and expert assessors. Currently in New Zealand there is no formal training to support teachers to make judgments against National Standards and schools are required to develop their own moderation processes. Additionally, tools developed to support teachers to increase the consistency of their judgments are not yet available.

Teachers' beliefs about the primary purpose of assessment in mathematics

Teachers' beliefs about assessment in mathematics has been the focus of a substantial body of research (Brown & Harris, 2010; Brown et al., 2012; Brown et al., 2011; Brown, Askew, Millett, & Rhodes, 2003; Delandshere & Jones, 1999; DuCloux, 2009; Even, 2005; Firestone, Winter, & Fitz, 2000; Hattie & Timperley, 2007; Peček, Zuljan, Čuk, & Lesar, 2008; Suurtamm & Koch, 2014; Suurtamm et al., 2010). Much recent research has been conducted exploring formative assessment (Brown & Harris, 2010; Brown et al., 2011) and feedback (Brown et al., 2012; Hattie & Timperley, 2007).

Suurtamm et al. (2010) explored teachers' approaches to assessment, measured by responses (n = 1,096) to two items from a 44-item questionnaire, and the assessment practices of nine teachers based on case studies, with data collected from classroom observations and interviews. These aspects were part of a larger study, the three year Curriculum Implementation in Intermediate Mathematics project that was designed to identify teachers' understanding and implementation of the mathematics curriculum from Grade 7to Grade 10 in Ontario, Canada. The sample was fairly evenly distributed across the province and across the grades.

Two questionnaire items required teachers to rate the frequency with which they use a range of assessment practices on a 4-point Likert scale (not at all, rarely, somewhat, or a lot) for each strategy or tool. One item asked teachers to rate the tools and strategies they used to get a sense of their students' understanding of mathematics. The other item asked teachers to rate the tools or strategies they used to make summative judgments of students' achievement in mathematics. The percentage of teachers who endorsed the use of each of the strategies or tools was compared for both of the items. This comparison identified that teachers relied most heavily on tests and quizzes to gather assessment information for both of the purposes. This reliance is indicative of traditional conceptions of assessment.

The nine teachers selected for case studies were observed teaching mathematics on four to six occasions with a follow-up de-briefing session after each observation where possible. The teachers were also interviewed prior to, and at the conclusion of the series of observations. All observations and interviews were conducted by a researcher and research assistant and were videotaped. The observations were coded for the presence of reform practices as identified in previous research and included problem solving, mathematical communication, the use of mathematical thinking tools and assessment to support learning. The interviews were transcribed analysed and coded based on teachers' beliefs and goals, reflections on their classroom practices and professional development needs and supports. Three of the case study teachers were described in terms of their effective use of a range of assessment practices that support student learning. The researchers found evidence of teachers using a range of assessment tasks. There was evidence to suggest that the case study teachers were integrating assessment into their instruction. This research uses three sources of data collection and two researchers to interpret the results both of which increase the validity of the research. The analysis of the data from the questionnaire was extremely limited, possibly presenting a threat to statistical conclusion validity. This research provided a description of the teaching practice and views of these nine teachers and cannot to be generalised to other contexts.

Brown et al. (2012) used quantitative methods to explore relationships between New Zealand teachers' beliefs about feedback and the practices they associated with feedback. A nation-wide survey of teachers' beliefs about feedback and checklist of practices associated with feedback was completed by a sample of 518 primary and secondary teachers This sample was representative of New Zealand schools on the basis of school size, region, and socio-economic status and representative of New Zealand teachers on the basis of gender and ethnicity.

Participants responded to each item on the survey by indicating their level of agreement on a 6-point Likert scale. They also selected from a list of 17 different feedback practices commonly used in New Zealand, indicating all of the practices that they thought of as feedback. Confirmatory factor analysis was used to recover 10 factors. Exploratory factor analysis was used to identify the dimensionality of the feedback practices instrument. Structural equation modeling was used to evaluate the relationships between teachers' beliefs and their feedback practices. From this analysis two factors were merged and items with strong modification indexes were removed. The measurement model consisting of 37 items within nine factors with acceptable goodness of fit was found for teachers' conceptions of feedback. The factors fell into three groups based on how strongly teachers endorsed them. The factors strongly endorsed by the teachers were, process (feedback about the process or strategies used by the student), self-regulation (reminding students about strategies they can use to improve their work), improvement (students use feedback to improve their work), and reporting and compliance (school expectation of the nature of the feedback given). The factors moderately endorsed by the teachers were, peer and self-feedback (students are able to give themselves and others accurate and useful feedback), task (whether the work was correct or incorrect), and timeliness (students should not have to wait for feedback). Encouragement (feedback is to make students feel good about themselves) was weakly predicted and irrelevance (feedback is pointless because students ignore it) was negatively predicted. The structural paths were generally weak but showed conceptually meaningful relationships between teachers' conceptions of feedback and the

practices associated with it. Teachers' conceptions of feedback predicted four feedback practices. Feedback for improvement predicted teacher formative feedback (β = .28), feedback for reporting and compliance predicted using feedback for reporting (β = .16), Feedback for encouragement predicted the practice of using feedback to protect students' self esteem (β = .22) and Both the conception of process (β = .21) and peer and self assessment (β = .27) predicted the practice of non-teacher feedback.

The work of Brown et al. (20012) suggests that New Zealand teachers endorse feedback factors associated with feedback to improve learning. The reasonably sized, representative sample increased the reliability of the research. One threat to the validity of the statistical conclusion was the inappropriate use of mean and standard deviation to compare the two groups of teachers.

Delandshere and Jones (1999) used case studies of three elementary teachers to identify the assessment beliefs held by these teachers and to examine the connections made between the assessment, teaching and learning cycle. A series of extended interviews with each of the participants were used to collect data. These interviews were recorded, transcribed then analysed using an analytic induction process to formulate a set of assertions. These were revised until the researchers were satisfied that they adequately captured the data, yielding three assertions. These assertions each stated an element that teachers' beliefs about assessment were shaped by: it's externally defined functions and purposes, what they perceive as the official curriculum within the school structure and where they position themselves with regard to the subject matter, and how they understand learning and learners.

The researchers classified their findings as "assessment paralysis"(Delandshere & Jones, 1999, p. 238), with the participants mirroring the content and style of the mandated tests to produce summative assessments of their students' achievement in mathematics. There was little evidence of formative assessment and little evidence of developing students' conceptual understanding of

mathematics demonstrated by the participants. The research suggested that all three of the teachers were teaching mathematical concepts as sets of "disconnected procedures" with little understanding of the underlying concepts or the pedagogical content knowledge related to developing students' conceptual understanding. This research explored teachers' practice on a very small sample, and while it provided a rich description of the beliefs of these three teachers the findings cannot be generalised in any way. Using two methods of data collection and two researchers to analyse the data enhanced the validity of the research. The final two research studies included both explore teachers' beliefs about assessment and either correlate those with their assessment practice, or compare the beliefs of different groups of teachers.

Using the four purposes of assessment as identified by Brown (2008): improved learning and teaching; school evaluation; students evaluation and irrelevance, Brown and Harris (2010) explored the relationship between teachers' beliefs about the purposes of assessment and their classroom assessment practices. A survey was used to measure teachers' beliefs about the purposes of assessment completed by over 160 teachers. Of these, 26 teachers agreed to be interviewed and a small sample of teachers (n=9) provided 32 self-selected samples of assessments they had used in English. From these samples 339 separate items, tasks or questions were identified for analysis. These items were coded using the American Surveys of Enacted Curriculum project taxonomy and then rated for their content and cognitive demand.

The survey consisted of 27 items, which aggregated into the four purposes of assessment. Each teacher's mean value for the four purposes of assessment were calculated. The mean values were correlated. There is very little information about the analysis of the survey data. From the analysis of the assessment items it is worth noting the relatively low level of the cognitive demand of the majority of the assessment items provided. The majority of items (73%) focused on lower-level cognitive tasks like memory, recall, explaining and following procedures compared to the 10% of items that required higher-level cognitive tasks like analysis and evaluation. The researchers also noted difficulties and

inconsistencies with using this tool within a New Zealand education setting because the tool did not map easily onto the New Zealand curriculum. There was a lack of clarity on how to differentiate between similar or overlapping categories. Brown and Harris (2010) suggested that the tool would need to be adapted further to adequately fit the New Zealand curriculum. The nine teachers agreed that assessment serves the evaluation of school quality more than any other purpose; this represented a shift in focus from previous studies of assessment beliefs held by teachers in New Zealand. From the statistical analysis of the data no statistically significant relationship was found between the teachers' beliefs about the purposes of assessment and their assessment practice. The small size of the sample limited the statistical power of the analysis. The inadequacy of the instrument used to rate the cognitive demand of the tasks presented a threat to the validity and reliability of this research.

Brown et al. (2011) compared primary and secondary teachers' conceptions of assessment in Queensland. They compared these data to data obtained from similar studies in New Zealand and Hong Kong, to explore the association between teachers' conceptions of assessment and differences in the education settings. A sample of 1,398 teachers from across Queensland completed a questionnaire (a response rate of 52.8%). The respondents were predominantly teaching in the range of Year 1 to Year 10, and were representative of the population of teachers employed in Queensland state schools. The conceptions of assessment questionnaire consisted of 27 items based on four conceptions of assessment. Assessment is to improve student learning, assessment is to increase student accountability, assessment is to increase school accountability and assessment is irrelevant. Respondents indicated their level of agreement to each item by selecting from a positively packed (two disagreement options and four agreement options) 6-point Likert scale.

The data were compared with the data set from an equivalent study of 525 New Zealand primary school teachers. From the New Zealand study a 27- item, nine factor, hierarchical model was found to have goodness of fit within acceptable levels.

Confirmatory factor analysis was used to test the fit of the Queensland data set to the existing model. The goodness of fit was adequate for the subsample of primary teachers but the model was rejected for the secondary teachers. The model was re-specified based on the original nine factor structure with two additional paths. This provided adequate goodness of fit for the whole data set. The fit was improved when the data was split into primary and secondary groups and the model was allowed to have different parameter values for each group. Examination of the model parameters for the two groups of teachers (primary and secondary) showed that there were significant differences in the interfactor correlations and the pathway regressions. The two groups had different interpretations of the conceptions of assessment and the model needed to be interpreted separately for each group. The means and standard deviations were calculated to compare the responses of the two groups.

The research found that Queensland primary and secondary teachers had similar conceptions of assessment. The small differences found in the conception of assessment for student accountability (greater for secondary teachers, as predicted) and assessment to improve teaching (greater for primary teachers) between the two groups, were consistent with the differences in assessment policy at the different levels of schooling. Comparison of the results with those found from the New Zealand and Hong Kong studies showed that Queensland teachers held conceptions of assessment very similar to New Zealand teachers, but very different from Hong Kong teachers, which is consistent with differences in the assessment policy in the three educational settings. The researchers suggested that the differences in policy framework may be associated with differences in how assessment is conceived by teachers. Differences in both policy and assessment conceptions may also be related to cultural factors.

Teachers' beliefs about the effectiveness of a more connectionist or discrete pedagogical approach to the teaching of mathematics

Askew et al. (1997) used a mainly qualitative approach to identify the key factors that enable effective teaching of numeracy in primary schools. The research was

based on a sample of 90 teachers from 11 schools in the south east of England. Thirty-three teachers were selected to interview and observe their classroom practice. Eighteen of these teachers provided the data for the case study. The remaining 15 teachers were used to validate the findings of the study and for supplementary data when necessary. A range of data collection methods was used including student assessments, questionnaires, interviews and observations. To assess students' understanding of the number system and ability to solve numerical problems, three variations of a test were developed to cater for the range of primary school students. Test one had 87 items and was designed for Year 1 and Year 2 students. Test two had 127 items and was designed for Year 3 and Year 4 students. Test three had 144 items and was designed for Year 5 and Year 6 students. These tiered-tests were administered at two testing points six months apart. Using the same test presents a threat to testing validity as it is unknown how much students' increase in achievement is attributable to familiarity with the test. This is particularly relevant as the researchers reported that the teachers found the second administration much easier than the first as teachers and students were more familiar with the format. Students' scores were adjusted to account for reduced gain due to possible ceiling effects of the test. The mean gain of each class was calculated and used to put classes in rank order within year groups to identify the most effective teachers of numeracy. As the students' scores were ordinal data, there were two issues to address with the use of these data. Calculating the gain in achievement by finding the difference between the raw scores and using the mean gain in students' achievement to compare classes are both inappropriate for ordinal data. Scale scores needed to be calibrated for this data, which in effect places each students' achievement onto the same scale, once this is done gains in students' achievement and class mean can be meaningfully calculated and compared.

Transcripts of observations and interviews were analysed using a constant comparative method. Each set of data was reviewed by at least three members of the research team. The analysis of teachers' interviews explored three aspects of beliefs that had been identified as being important. The beliefs that were

examined were beliefs about the nature of numeracy, beliefs about pupils and how they learn to become numerate, and beliefs about how best to teach pupils to become numerate. Three areas of teachers' knowledge were also explored: numeracy subject knowledge, knowledge of their pupils, and knowledge of numeracy approaches and representations. The beliefs and knowledge of teachers found to be highly effective were compared and contrasted with those of teachers who were less effective teachers of numeracy. The education settings that the teachers were operating within were also taken into consideration to identify aspects of the setting that appeared to impact on teachers' effectiveness. From the analysis of teachers' beliefs, three orientations towards the teaching of numeracy were identified: *connectionist*, *transmission* and d*iscovery*. These teaching orientations were compared and contrasted in relation to the three aspects of beliefs being explored. Teachers' practice was then analysed in light of these three orientations. Teachers' content knowledge was analysed from the questionnaire data, teachers' profiles and classroom observations. These analyses compared and contrasted teachers' content knowledge through the lens of the teaching orientations. Six factors were identified as important from the analysis of teachers' interviews. These factors were teachers' fluency with numeracy, the scope of teachers' conceptions of numeracy, the links made between concepts, the explanation of the way concepts were linked, the depth of the explanation, and the level of understanding shown by the explanation. Observations of teachers' practice were reviewed in light of these factors. The frequency with which each factor was observed was analysed in respect of its correlation with adjusted mean class gain score. The correlation coefficients and the regression analyses are not reported, but they are mentioned as showing no relationships except for the factor depth. Teachers with a low proportion of conceptual links tended to have low gains in students' achievement. Age group was identified as a possible interacting factor. Analysis of teachers' interviews identified eight factors in relation to teachers' knowledge of their students which were compared with pupils gain scores to identify associations between the factors and students' achievement. Students' attitudes towards mathematics and students' approaches to mathematics were both associated with their achievement. The greater the

percentage of statements made by the teachers' about either of these factors, the greater the students' gain.

The findings of the research suggested that teachers identified as being highly effective in the pedagogy of numeracy held a particular set of coherent beliefs that the researchers described as connectionist.

The validity of the study was increased by the multiple sources of data and the analysis by at least three members of the research team. Threats to validity included the use of the same test at both testing points and using inappropriate methods to compare data. The statistical analyses used throughout this research were either inappropriate (using the mean to describe ordinal data) or not adequately reported (correlation and regression analysis), which are threats to both the validity and reliability of the research. Throughout the research the findings are reported as traits indicative of all effective teachers of numeracy, which is inappropriate as this was an in-depth study of a relatively small group of teachers from one general geographic area of the UK and as such cannot be genaralised. This research was included in the review as this research was the basis of the connectionist pedagogical approach. As such this was an important research study to include when exploring teachers' connectionist pedagogical beliefs.

Stipek et al. (2001) examined teachers' beliefs and practices to identify the coherence of these beliefs and their association with classroom practices. The sample comprised 21 teachers of grade 4 to grade 6 students from schools throughout Los Angeles County, California. Data were collected from a teachers' beliefs survey, a teachers' questionnaire and a students' questionnaire at the beginning and end of the teaching year. Each teacher was also observed teaching two mathematics lessons. Students (n = 437) were predominantly from low socio economic backgrounds. Teachers' beliefs were measured using a 57-item survey, respondents select from a 6-point Likert scale of agreement. The statements contrasted traditional pedagogical beliefs with inquiry oriented pedagogical beliefs over five constructs. Two additional sets of items were focused on

teachers' enjoyment of and confidence in their teaching of mathematics. The five constructs of teachers' beliefs were: mathematics as a set of operations versus a tool for thought, correct answers versus understanding as the primary goal of teaching, teacher control versus child autonomy in classroom lessons, entity versus incremental view of intellectual ability and extrinsic versus intrinsic motivation of students. The beliefs theoretically aligned with inquiry-oriented practice, as well as decreased enjoyment and confidence, were reverse coded. The mean was calculated for each of the seven constructs. The data were not analysed to identify how each of the items within each construct loaded onto the factors identified. Cronbach's alpha was reported as a measure of the coherence of the items within each construct, for both administrations of the survey. The Cronbach's alpha for the first administration ranged between .67 and .84 and for the second administration from .44 to .80. Additionally, the stability of the teachers' beliefs over the two administration points was estimated using the correlation between the two administrations of the survey. Correlations ranged from r = .47 for teachers' enjoyment to r = .81 for a focus on correctness versus understanding and teacher control versus child initiated. Principal components analyses were carried out using the mean agreement level for each construct at both data collection points. Two factors were identified; the first factor with an eigenvalue of 4.01, accounted for 57% of the variance and the second factor with an eigenvalue of 1.44 accounted for 21% of the variance. The five constructs of teachers' beliefs loaded onto the first factor with loadings ranging from .72 to .89. Teachers' confidence and enjoyment both loaded onto the second factor with loadings of .76 and .68 respectively. Because of the inappropriate use of teachers' mean agreement for each construct, the validity of the principal components analyses and the correlation between the two administrations of the survey are in questionable.

To compare teachers' classroom practices, each teacher was observed and videotaped while teaching two lessons on adding and comparing fractions. A rating system was developed (from 1 "not at all like this teacher" to 5 "very much like this teacher"), to assess the degree to which teachers used each of seven identified practices. The practices of interest were: the teachers' emphasis on

performance outcomes, the teachers' emphasis on speed in completing tasks, the type of environment the teacher fostered, the teachers' facilitation of students working autonomously, the teachers' emphasis on the value of effort, the teachers' emphasis of students developing understanding and mastery, and the teachers level of enthusiasm. The recordings were watched multiple times with 29% of the lessons rated independently by two raters (agreement ranging from 70% to 100%). The mean rating for each of the observed practices was calculated. Each of the mean values for the five constructs of teachers' beliefs was correlated with each of the mean values for classroom practices. These identified relationships between more traditional beliefs with more traditional practices.

An entity belief of students' mathematical ability was related to teachers emphasising performance. The elements that teachers based their summative assessments on were measured by selecting from a 5-point Likert scale (from 1, "not at all" to 5, "mostly") for each of four dimensions of students' achievement. The dimensions were, effort, relative scores, creativity, and independence. These dimensions were correlated with teachers mean beliefs. Students completed a 6-item questionnaire, rating their responses on a 6-point Likert scale. Three items were focused on students' self-perceptions of their mathematical ability and three items were focused on their enjoyment of mathematics. Two partial correlations were calculated. Teachers' confidence teaching mathematics and the mean of students' confidence in their mathematics ability at the end of the year, with the mean of students' confidence in their mathematics ability at the start of the year covaried (n = 18, r = .54). The other partial correlation was between teachers' and students' enjoyment of mathematics activities at the end of the year, with students' enjoyment of mathematics activities at the beginning of the year covaried (r = .05). The findings of the research suggest that teachers tend to have coherent sets of beliefs that are related to their classroom practices.

This research used multiple instruments to collect data and used two data collection points and two groups of participants, all of which increased the validity and reliability of the research. There were however several serious threats to both the validity and reliability. The small sample that was skewed

towards students from lower socio economic and Latin American backgrounds were both threats to the reliability of the research.

Gill and Hoffman (2009) analysed teachers' discourse during mathematics planning meetings to identify the beliefs underlying planning decisions of a team of four 8th grade teachers. The teachers all taught at a suburban middle school in Florida. Their semester-long case study was based on observations and recordings of the team's weekly meetings (n = 9) to discuss their lessons, get feedback and plan the following weeks' mathematics lessons. The recordings were selectively transcribed based on the scenes of rich dialogue indicated from the observations. The transcripts were initially coded and classified into relevant domains by the primary researcher, yielding 57 domains. The researchers reanalyzed the data looking for patterns. All of the domains that were associated with teachers' thinking about learning and instruction and teachers' rationales were organised into a taxonomy supported by examples from the transcripts. Six categories of teachers' beliefs were created, based on the researchers' review of literature. These categories were based on beliefs about: pedagogical content, general pedagogy, subject matter, curricular choices, resources/textbooks, and students' thinking. Teachers' discourse was mapped onto the beliefs; these were then reviewed to identify themes. Seven themes were identified; these themes were consistent with traditional, instrumentalist beliefs such as a procedural over a more conceptual emphasis and entity theories of students' mathematical ability. Of note was the teachers' deliberate decomposition of the curriculum into increasingly discrete, de-contextualised components as their perception of student's mathematical ability decreased.

Transcripts from one meeting were coded and categorized by two reviewers asked to focus on rationales for teachers' behavior or beliefs about teaching and learning mathematics. Both coders identified patterns of rationales and behavior with the same underlying belief as the primary researcher. A measure of inter-rater reliability was required to indicate the level of agreement between the various coders for this one set of transcripts. Data were triangulated through comparison with interviews with the principal and assistant principal and

interviews and classroom observations of the lead teacher in the team that had been carried out the previous semester as part of another study.
Of particular interest was the choice of participants for this study.

One of the participants had been a participant in previous research focusing on reform-based teaching of mathematics; the teacher was identified as teaching in a traditional (instrumentalist) style. So, the researchers already knew the likely orientation of beliefs of at least one of the participants. This was a particularly powerful influence as this teacher took the dominant role in planning meetings. The primary researcher completed all of the observations and initial coding. The one set of transcripts that were coded by two other teachers as a blind review were not supported by a measure of inter-rater agreement. All of these factors present threats to the validity of the research. The research described the planning meetings of a very small sample of teachers working in the same school and cannot be generalised to other contexts.

Bahr et al. (2009) studied the effects of combining in-service professional learning with a pre-service mathematics methods course with a focus on reform pedagogy. The study involved 21 in-service teachers and more than double that number of pre-service teachers (n = 52). The study was described as an experimental design used to measure the effects of the programme, comparing the students in the course with students in two equivalent courses at Utah institutions of higher education. The students were not randomly assigned to their classes, they were intact class groups, and as such this research is more accurately described as a quasi-experimental design. The three courses were taught by the same teachers and had equivalent proportions of instruction and classroom practice. The intervention was a pre-service mathematics methods course taught at a local school with 11 staff members released to attend the course. They were joined by 10 teachers from neighbouring schools. Each in-service teacher was grouped with two or three pre-service teachers. They attended a two-hour lesson on methods instruction, a component of which was for each group to plan a reform based mathematics lesson. Each group then implemented the planned lesson in the in-service teachers' class. Over the

semester, the pre-service teachers took increased responsibility for planning and implementing the lessons.

The instrument used as the pre- and post -intervention assessment was an on-line survey designed to measure teachers' beliefs associated with reform-based classroom practice. The survey presented teachers with video or written vignettes, which the teachers responded to using a series of open-ended questions. These responses were analysed using rubrics. The means for pre and post assessment were calculated and compared for each group of teachers (pre-service intervention, comparison group A, comparison group B and in-service teachers). A series of t-tests were completed to examine changes in teachers' beliefs. This analysis identified significant changes in five of the seven beliefs measured for the intervention pre-service teachers. The belief changes for the three pre-service groups were compared using an analysis of covariance, followed by the Tukey-Kramer post-hoc procedure to test the degree of significance associated with specific group comparisons. No significant differences in the belief changes were found, with two exceptions. The two comparison groups differed on the belief that children can solve problems without being taught how to solve them. The intervention group differed from comparison group A on the belief that children should do as much of the thinking as possible, with the intervention group having the greater change towards agreeing with this belief.

Students' achievement was measured using three instruments. The first measure was a standardized test that was administered at the beginning and end of the year and had also been administered in the previous year. The gain score was calculated for students in grade 3 to grade 5; these were compared to the previous year's gain scores. These scores were analysed using a t-test for paired samples. Both the grade 3 and grade 4 students showed significant gains in comparison to their gains the previous year. The grade 5 students showed significantly less gain than the previous year, this was in part attributable to a ceiling effect within the test. The second measure of students' achievement was the statewide, standardised assessment. The mean student achievement (%) by

grade level was compared to the previous years' achievement. This comparison showed increased achievement for all levels from grade 1 to grade 5. With both of these standardized assessments there is no mention of whether the students' raw scores were converted to scale scores, prior to calculating the mean. The final assessment was designed by the teachers to assess students' understanding of mathematical concepts and ability to solve problems. A rubric was designed to quantify the level of support the student needed to solve the problems and to allocate scores to students based on their level of independence when solving problems. The teachers were trained in the rating students' achievement using this assessment. Inter-rater reliability was calculated using Cronbach's alpha, with high levels of consistency across all grades, the one grade with lower consistency was when assessing Grade 1 students.

A weakness of the research was that in-service participants might have had a predisposition to adopting reform pedagogy in mathematics, because that is what they had chosen to study. As such the in-service participants could not be considered representative of teachers in general.

Bolden and Newton (2008), in their qualitative study, sought to identify the epistemological beliefs of three teachers of Year 5 and Year 6 students. All teachers were from schools considered to be successful in raising students' achievement as measured by standardised test results. Interviews and observations of practice were used to infer their beliefs. Three epistemological worldviews held by teachers were discussed: the realist, the contextualist, and the relativist world-views. The realist world-view assumes that there is an objective and unchanging body of knowledge, that knowledge is absolute. This is the world-view associated with the behavioural model of teaching. The contextualist world-view assumes that knowledge is consensually agreed and shared within communities, and is associated with social constructivist theories of teaching and learning. The relativist world-view assumes that each learner constructs his or her own unique version of knowledge. This world-view is associated with a radical constructivist model of teaching and learning.

The teachers appeared to hold hybrid epistemological world-views, although all appeared to hold more relativistic views than realist views. All teachers expressed a desire to teach using an investigative approach, and associated with children gaining a greater conceptual understanding of mathematics. This was at odds with the more transmission style of teaching evident during observations of mathematics teaching. Teachers all expressed concerns that teacher accountability practices and the standardised testing regime used to measure and compare students' achievement and schools' effectiveness in the UK were impeding their pedagogical decisions.

Relationships between teachers' pedagogical content beliefs, pedagogical content knowledge (measured by structured questionnaires, interviews and cognitive tasks), and students' achievement in mathematics (measured by a test using word problems and a test of addition and subtraction number facts) were identified in the research by Peterson et al. (1989). First grade teachers (n=39) from 27 schools participated in the study. Two instruments were used to measure teachers' pedagogical content beliefs: a 48-item questionnaire, rating their level agreement to each item on a 5-point Likert scale and structured interviews. Responses to the interview were given a rating from one to five. Means and standard deviations were calculated for each construct, Teachers' pedagogical content knowledge was assessed by the completion of three sets of tasks: discrimination between pairs of addition tasks based on difficulty; identification of strategies students used to solve addition problems; and identification of the strategies specific students in their class use to solve addition problems. Students' achievement (n=710) in addition and subtraction was measured using two assessments: a timed addition and subtraction number fact test and a test using word problems. The data from the questionnaire and the interview were analysed by calculating the mean and standard deviation for each of the constructs, as the data are ordinal these are not appropriate measures of central tendency or dispersion. Cronbach's alpha was calculated to measure the internal consistency of each of the constructs, the correlation between each of the constructs showed positive correlation between each of the

constructs.

Two groups of teachers were identified as consistently scoring either high or low on each of the constructs. Each group consisted of seven teachers: one group was categorized as having cognitively based perspectives and the other group as having less cognitively based perspectives. Associated with the cognitively based teachers were a strong focus on developing students' conceptual understanding, and an informal, formative approach to assessment. The less cognitively based teachers were associated with a strong focus on developing students' procedural competence and a formal assessment approach, focusing on computational accuracy. The number of years teaching was a statistically significant difference between the two groups. There was a significant positive correlation between teachers' beliefs and students' achievement when solving word problems, however there was no correlation between teachers' beliefs and students' achievement in the calculation of addition and subtraction facts.

Research findings about teachers' procedural pedagogical beliefs

Identification of the most effective sequencing of addition and subtraction procedural lessons with decimal place value conceptual lessons on student achievement was the aim of the research carried out by Rittle-Johnson and Koedinger (2009). They conducted two studies, both aimed at identifying the effect that the order of deliver of lessons had on students' achievement. The initial study used a quasi-experimental design, using a pre-test, an intervention, and a post-test. The participants (n=77) were from four sixth grade classes from two schools. Two classes at each school were randomly assigned either the intervention or the control. The pre-test and post-tests were randomly assigned to students from two tests using the same format and question layout with different numbers. The tests assessed students' knowledge of decimal place value and decimal arithmetic using both familiar items (similar to those used in the lessons) and novel items (requiring an extension of the material covered in the lessons). A series of six computer-based lessons were created for the students to work through independently. Three of the lessons were developing students' decimal place value knowledge, which the researchers referred to as

the conceptual lessons. The remaining three lessons were teaching the algorithmic procedure for decimal arithmetic, these were the procedural lessons. The intervention groups were presented the lessons in an iterative fashion, sequencing first a conceptual lesson followed by a procedural lesson and so on until students had completed all six lessons. The control groups were presented the series of three conceptual lessons first followed by the series of three procedural lessons. Students' gain score was calculated (post-test score – pre-test score). Mean gain scores were calculated and compared across all groups and variables. A mixed-measures ANCOVA was conducted to confirm the effects were significant. The gains made by students in the iterative group were greater than those made by the control group. There was an interaction between sequence and knowledge type, with the iterative group having greater gains in addition than place value. A regression of pre-test place value and arithmetic knowledge on gains in place value and arithmetic knowledge found that knowledge of place value at pre-test predicted gains in arithmetic.

The second experiment differed from the first on a couple of important factors. Firstly, the sample of students was drawn from two classrooms at one school and was much smaller (n=26). The lessons and assessments were slightly modified from feedback from the first experiment. Also, the treatment was randomly assigned to students. The same analyses were carried out on these data, with similar results. The students in the treatment group made substantially greater gains in arithmetic knowledge than the students in the control group. The students in the control group made greater gains in place value knowledge within familiar contexts, but the treatment group made greater gains in place value within novel contexts. The results from the regression showed place value knowledge at pre-test predicted gains in arithmetic knowledge and arithmetic knowledge at pre-test predicted gains in place value. The approach of iterating between concepts and procedures is similar to that advocated in the New Zealand Numeracy Development, with number knowledge being developed alongside strategy development. There is however questionable construct validity in this study, with the conceptual lesson structure and test sample appearing more procedural than conceptual in the approach to teaching place

value and re-grouping.

Research findings from the study of implicit beliefs of ability as either incremental or entity

Much of the research reviewed is studying intelligence, for the purposes of this research these are treated as synonymous. Also, implicit theories, beliefs and views are treated as synonymous in this study.

In their review of the findings from their research into implicit theories held by people, Dweck et al. (1995) argued that the implicit theories held by people influenced their inferences, judgments and reactions. They drew on the results from six studies exploring the relationships between participants' implicit theories in the domains of intelligence and morality and their judgments and actions in these domains. In each of the studies participants responded to a 3-item survey with each item stating an entity view of intelligence. Respondents rated their agreement to each item on a 6-point Likert scale. Scores were averaged to form an overall implicit theory score. The measure of implicit theories of morality had the same format and scoring method. An additional general concept was also measured, the implicit theories about the nature of people. The structure and scoring format was the same used for measuring the implicit theories of intelligence and morality.

Jones et al. (2012) explored the implicit beliefs about intelligence held by pre-service and in-service teachers enrolled in educational psychology courses at three universities in the USA. The students (n = 270) completed the implicit theory of intelligence scale designed by Dweck et al. (1995) and an open-ended definition of intelligence item, designed to identify the students' structure of intelligence. To investigate preservice and in-service teachers' definition of intelligence, a thematic whole text analysis was used to develop a grounded theory from the responses to the open-ended items. The responses were coded to identify seven themes and 55 coding categories within the themes by one of the authors. The remaining three authors compared codes after independently coding 10% of the responses. As a result of the reanalysis the codes were re-categorized, eliminating 13 of the original codes leaving seven themes containing

42 categories, with the inter-rater reliability of 86.7%. The themes that respondents identified as elements of intelligence were achievement, declarative knowledge, procedural skills, self-regulation, cognitive processes, motivation, and personal characteristics such as open-mindedness.

The Implicit Theory of Intelligence Scale comprised three items, with students choosing from a 6-point Likert scale of agreement for each item. Responses to the three items were averaged to give the mean score for each respondent's theory of intelligence. The internal consistency of the three items in this study, Cronbach's alpha, $\alpha = .92$, which the researchers interpreted as an indication that internal consistency was high, however another interpretation is that the items were mutually redundant. There was found to be no statistical difference between the means for the in-service teachers ($n = 33$) compared to the preservice teachers ($n = 237$). From the combined sample 77.9% of respondents endorsed an incremental view of intelligence. The correlation between years of experience and implicit theory of intelligence gave a correlation coefficient of $r = .15$, which indicates no relationship between the variables. This is counter to findings from similar studies, however the number of in-service teachers in this sample is small.

In their study exploring the relationship between educators' implicit theories of intelligence and their beliefs about the identification of gifted students Garcia-Cepero and McCoach (2009) used four surveys to measure respondents' beliefs. The surveys were sent out to a nationally representative sample of 1,000 teachers (respondents, n= 168) and 1,000 teacher educators (respondents, n=204) from across the country (unspecified). The respondents (less than 20% of the selected sample, so unlikely to be representative) completed four surveys described below. The implicit theory of intelligence survey (constructed by the authors) was developed to identify the beliefs about the structure of intelligence. Respondents rated items on a 7-point Likert scale of agreement that each item contributes to the prototype of intelligence. The items belonged to four subgroups, analytic, practical, creative and inter-intra personal. The survey of implicit theories of intelligence, developed by (Dweck et al., 1995), identified respondents' beliefs about the malleability of intelligence. Respondents rated

each of the eight items on an 8-point Likert scale of agreement, with half of the items stating incremental views and half of the items stating fixed views of intelligence. The survey to measure participants' beliefs about the identification of gifted students was adapted from Brown et al. (2005) and respondents rated items on a 7-point Likert scale of agreement. This survey had ten items, half of the items stating views consistent with using IQ tests and the remaining half advocating the use of multiple criteria for the identification of gifted students. The fourth instrument asked respondents to rate themselves on a 7-point Likert scale about their own abilities in each of the following areas, inter personal skills, analytic ability, practical abilities, inter-intra personal abilities, social conscience and creativity.

Structural equation modeling methods were used to analyse the relationships among the scales and the differences between teachers' and teacher educators' responses. The data from each survey was fitted to a model. Exploratory factor analysis was used on data from two of the surveys, the remaining two sets of data used findings from factor analysis from previous studies.

To analyse the relationships among the constructs, structural equation modeling was used, and a model was developed to illustrate the relationship between each of the components.

They did not find any clear relationship between the structure of intelligence and the beliefs about the malleability of intelligence, though educators who endorse practical abilities and inter-intrapersonal skills as attributes of intelligence are also more likely to hold an incremental view of intelligence. There are several inconsistencies, errors and omissions throughout this article. The country in which the research was carried out was not specified. `The authors state that the response rate for their surveys was approximately 25%, 372 out of 2000 is less than 20%. Throughout the article, there are similar inconsistencies, stating that there are five factors, then listing six, stating that seven items were omitted then giving an explanation for six of the items.

Research exploring differences in teachers' implicit beliefs about intelligence based on the subject areas in which they teach are considered. A quasi-experimental study by Jonsson et al. (2012) was designed to find if teachers within different subject areas, held different implicit beliefs about intelligence, or favoured different scientific theories of intelligence. They also explored if the age or experience of the teachers were related to their implicit theories of intelligence. A purposive sample of teachers (n=226) from 4 Swedish high schools was included in the research. The teachers were selected to ensure that a full range of subject areas and school types, suburban/rural, theoretical/practical were included in the sample. Secondary schools in Sweden offer either a theoretical or a practical programme. Dweck's (1999) Theories of Intelligence Scale was used to measure teachers' beliefs about intelligence. The scale consisted of eight items, with four of the items stating entity views and four of the items stating incremental views of intelligence. Teachers rated their agreement with each statement by selecting from a 10-point Likert scale. Teachers then rated four scientific theories of intelligence using a 10-point Likert scale of credibility. The theories described were Sternberg's triarchic theory of intelligence, Gardiner's theory of multiple intelligences, Soviet sociocultural theory and Cattell-Horn-Carroll's theory of cognitive abilities.

To answer the first question a mixed 2 x 4 ANOVA was carried out, with entity and incremental beliefs within subject variables and discipline (mathematics and science, language, social science/humanities and practical) as between subject variables. A main effect was found that the teachers favoured incremental theories of intelligence over entity theories of intelligence for all subject areas except mathematics and science. The teachers of mathematics and science did not significantly prefer either entity or incremental theories of intelligence.
To answer the second question a mixed 4 x 4 ANOVA was carried out, with scientific theory of intelligence as within subject variable and subject area as between subject variable. A main effect was found on the within subject factor, scientific theory of intelligence F (3, 186) = 35.74. Pair-wise comparisons with adjustments for multiple comparisons found that Cattell-Horn-Carroll's theory (M = 4.26) differed from the three other scientific theories, triarchic theory (M =

5.78) differed from sociocultural theory, and multiple intelligences differed from sociocultural theory. The teachers' endorsement of sociocultural theory was significantly higher than the other theories. Pearson's correlations were used to analyse the relationships between entity and incremental theories of intelligence and the four scientific theories. Both measures are examples of ordinal data, Spearman's rho is designed to calculate the correlation coefficients for ordinal data, Pearson's should only be used with scale data. A strong negative correlation ($r = -.704$) was found between the two implicit theories of intelligence. A moderate correlation ($r = .398$) was found between Cattell-Horn-Carroll's theory and an entity theory of intelligence, with a slightly smaller negative correlation ($r = -.275$) between Cattell-Horn-Carroll's theory and an incremental theory. The opposite pattern was found with sociocultural theory, with a small negative correlation ($r=-.253$) between an entity theory and a small correlation ($r = .228$) between an incremental theory. As the correlations were calculated using an inappropriate method the validity of these findings is called into question.

To answer the third question a 2 x 2 ANOVA was performed with age and experience the between subject factors, and entity and incremental beliefs of intelligence the dependent variables. A split half method was used to divide the teachers into two groups based on the median for each of the independent variables, age and experience. The younger group of teachers was ≤ 48 years old and the older group of teachers was ≥48 years old, this means that teachers who were 48 years old were included in both groups. Likewise, less experienced teachers were those with ≤ 13 years' experience and more experienced teachers were those with ≥ 13 years experience, so teachers with 13 years experience were included in both groups. No main effects were found, but an interactive effect for the dependent variable entity theory was found between age and experience. Older teachers with more experience and younger teachers with less experience both showed a greater preference for entity theories of intelligence. However, the discrepancy with the groupings calls into question the validity of these findings.

The findings from this study support the assertion that teachers of mathematics and science are less likely to hold an incremental theory of intelligence than teachers of languages, humanities and practical subjects.

Ilhan and Cetin (2013) developed an instrument to measure students' theories about intelligence in mathematics. They adapted the implicit theory of intelligence instrument developed by Dweck et al. (1995) to be mathematics specific and added items to those designed to measure entity beliefs an included items designed to measure students' incremental theories of fgbij8 v high schools in Diyarbakir in Turkey. Data from the first sample of students (n = 304) was used to understand the psychometric characteristics of the scale. The scales' construct validity was examined by first performing exploratory factor analysis, using basic components method and direct oblimin rotation. A two-factor structure explaining 48% of the total variance was obtained. The first factor was named entity theory; this subscale comprised six items (with factor loads ranging between .39 and .86) and explained 31% of the variance. The second factor was named incremental theory; this subscale explained 18% of the variance and comprised five items (with factor loads ranging between .49 and .75). Confirmatory factor analysis was performed to determine if the 11 item, two factor structure identified had satisfactory goodness of fit. Twelve different measures of goodness of fit were calculated, with all measures meeting the criteria for either perfect goodness of fit or acceptable goodness of fit.

To establish the criterion-related validity of the mathematics oriented implicit theory of intelligence scale, students' end of year grades were correlated with their entity and incremental beliefs measured in this study. Based on findings published previously in literature, it was argued that a positive relationship between mathematics achievement and students' incremental theories of mathematical intelligence, and a negative relationship between students' achievement in mathematics and their entity theories of mathematical intelligence would be evidence of the criterion-related validity of the instrument. The correlations confirmed these findings with a negative correlation (r = -.36) between students' entity theory of mathematics intelligence and their

mathematics achievement. The correlation between students' incremental theory of mathematical intelligence and their mathematics achievement was described as positive but reported as negative (r = -.45). The authors stated that these results could be evaluated as proof of criterion-related validity. There is no discussion of the criteria used for the assignment of students' end of year grades in mathematics at either of these high schools the validity and reliability of the grades are unknown. Additionally, correlations in the hypothesized direction between these variables may be evaluated as supporting, rather than proof of, the criterion-related validity of the instrument.

The internal consistency of the entity theory subscale was found to be .75, and the incremental theory subscale was found to be .76. To determine the test-retest reliability coefficient, the instrument was administered to a small sample of students (n = 91), and then repeated two weeks later. The test-retest reliability coefficient was .96 for the entity theory subscale and .93 for the incremental theory subscale. Both of these measures provide reliability coefficients within the guideline of .7 and over being considered to be reliable.

This study supports the view that students' incremental views of intelligence are associated with increased achievement in mathematics. Of interest to the current study is if there is a similar relationship between teachers' views of students' ability and their increased achievement in mathematics.

This review of the literature has informed the research questions that were explored by this study.

Methods

Three measurement variables were collected and analysed in this research. Teachers' beliefs about the teaching, learning and assessment of mathematics were collected through their responses to a questionnaire. Students' achievement data were collected from each teacher involved in the research, at three testing points, using three, randomly assigned, parallel tests. The final judgments teachers made about students' achievement against the mathematics standards were also collected. These three sets of data were all examples of ordinal data. The main characteristic of ordinal data is that it is possible to place respondents in a rank order with those showing the greatest amount of the attribute being measured assigned the highest value and those with the least being assigned the lowest value. Students' scores on a mathematics test are an example of ordinal data because students are place in order from the lowest score to the highest score. Ordinal data cannot however be used to estimate change such as progress over time.

Participants

Two groups of participants took part in this research: The main group comprised a sample (n=14) drawn from Wellington primary school teachers, teaching Year 5 and/or Year 6 students in 2014. This group of teachers collected the student assessment data analysed in the research. A supplementary group of participants (n=68) drawn from the total population of primary school teachers completed a questionnaire designed to measure teachers' beliefs about effective pedagogy, the nature of students' ability and the purpose of assessment in mathematics, in order to provide sufficient statistical power to calibrate the questionnaire items to measurement scales.

Participants for the main focus of the research

The main sample of teachers was drawn from a range of schools throughout the Wellington region. All participating teachers were teaching Year 5 and/or Year 6 students in 2014. Principals and lead teachers in mathematics from the majority of primary schools within the greater Wellington region were approached via e-

mail and a follow-up telephone call. They were asked to inform eligible teachers in their schools about the research and to refer any interested teachers to the researcher. Contact was made with those interested in participating via e-mail. Seventeen teachers agreed to participate, one of who withdrew part way through the research. Of the 16 remaining teachers, four provided incomplete sets of data for their students and, of these, two were missing data that was essential for analysis of students' achievement. This left 12 teachers providing complete sets of data for their students and an additional two teachers with enough data to be included in the sample. These 14 teachers comprised the focus sample for the research. From these 14 classrooms, complete data sets (all 3 assessments and overall teachers' judgments of achievement in mathematics) were provided for a total of 231 students, and reduced data sets (2 assessments and overall teachers' judgments of achievement in mathematics) were provided for another 39 students.

The 14 teachers were from eight central Wellington, state, co-educational schools. The characteristics of the schools are outlined in table 2.1. Demographic information about the schools was taken from the Schools Directory (Ministry of Education, 2015a). The final sample is heavily biased towards high decile schools.

Table 2.1: Participating schools' demographic information

School	Number of Participating teachers	Roll	Decile	School type
A	5	381	10	contributing
B	2	390	10	full primary
C	2	202	10	full primary
D	1	164	10	Contributing
E	1	212	7	Contributing
F	1	116	9	Contributing
G	1	210	9	full primary
H	1	260	10	Full primary

Participants who completed the questionnaire of teachers' beliefs

The supplementary sample of teachers was contacted either by e-mail or in person and was provided with a copy of the questionnaire to complete on a voluntary basis.

Instrument to measure students' achievement in mathematics

An independent measure of students' achievement in mathematics was developed using items from e-Asstle. Initially 70 items were selected to broadly reflect the achievement objectives at level 3 of the New Zealand Curriculum (Ministry of Education, 2007) in mathematics. A small number of level 2 and level 4 items were also selected to minimise floor and ceiling effects. All 70 items were trialed on 91 Year 5 and Year 6 students, late in 2013. Following this trial four items were rejected on the grounds that they did not adequately discriminate; they were either answered correctly or incorrectly by almost all of the examinees. The remaining 66 items were distributed into three, approximately parallel, tests of 22 items each (see appendix 1). The three tests were administered to the students in each of the 14 participating teachers' mathematics classes, on three separate occasions over the period of the research. The teachers also provided their summative judgments for each of their students against the National Standards in mathematics, at the end of Year 5 and at the end of Year 6, as appropriate.

Instrument to measure teachers' beliefs about effective pedagogy, students' ability, and the purpose of assessment in mathematics

A questionnaire was constructed to measure teachers' beliefs about each of four constructs, namely: the extent to which the teachers believe that mathematics is best taught by explicitly connecting concepts for the students; the extent to which the teachers believe that mathematics is best taught by focusing on developing a conceptual understanding, as opposed to a procedural competence; the extent to which the teachers believe that students have an inherent mathematical ability (or lack of ability) largely independent of the quality of the teaching they receive; and the extent to which the teachers believe that the

purpose of assessment in mathematics is to inform the teaching and learning as opposed to assessment being primarily for reporting purposes.
The questionnaire consisted of 40 items, four items for each of the four constructs, two stating the positive view of the construct, two stating the contrary view.

The statements were then compiled into a questionnaire asking respondents to rate their level of agreement with the statement from a 6-point Likert type scale. One open-ended question was asked giving respondents the opportunity to express their views about effective pedagogy, students' ability and the purpose of assessment in mathematics. The final part of the questionnaire asked respondents for demographic information. The questionnaire was disseminated initially on paper copy and later, electronically using Qualtrics (see appendix 2).

Procedure

The data were collected between April 2014 and April 2015. Participating teachers completed the questionnaire upon agreeing to participate. The student assessment element of the research took place over an 18-week teaching period (two holiday periods of two weeks each occurred over the time of the study, these were not counted) between Term 2 and Term 4, 2014. Students were assessed at three different times over this period: the first assessment was administered in Term 2; the second was administered 11 (teaching) weeks later; and the third after seven further (teaching) weeks. The assessments were constructed to broadly cover the mathematics curriculum at level 3, Each assessment comprised 11 items focusing on number and algebra problems, and the remaining 11 items, a mix of measurement, geometry and statistics problems. Students were assessed on all three assessments over the 18-week period, thus allowing a measurement of students' progress in mathematics over this time.

Overall teacher judgments against the National Standards in mathematics were collected from the 12 teachers in the study. These judgments gave the teachers' on balance judgment for each student's achievement in mathematics. The

judgments rated each student's achievement in mathematics on a four-point scale as being either *above*, *at*, *below* or *well below* the standard at each year level

The questionnaire was completed by all of the teachers in both samples. The questionnaire was used to infer the teachers' beliefs around the four constructs.

Design and Analysis

A nonexperimental, quantitative research design was used to answer the central question: are there any relationships between teachers' application of assessment criteria against the mathematics standards, either conservative or liberal, and each of the four dimensions of teachers' beliefs?

There were three elements to the research: measuring students' achievement and tracking their learning trajectories over an 18 teaching-week period; collecting teachers' judgments against the National Standards in mathematics for the students in their classes; and inferring teachers' beliefs against the four constructs of effective pedagogy, against students' achievement and the purpose of assessment in mathematics

Assessments were randomly assigned to each class using a Latin square design; this design ameliorates any differences between the three assessments. This design requires all versions of the assessment to be used at each point of assessment, in equal numbers, and that each class is assessed on all versions of the assessment over three assessment points. The items used in the assessment were either multi-choice or short answer and were marked using a dichotomous marking schedule, allocating 1 mark for a correct response and 0 marks for an incorrect response.

Item response theory was used to analyse the student assessment data. A Rasch model was used to calibrate these ordinal data onto a scale location for each respondent. One condition of entering data into a Rasch model equation is that the data is unidimensional, which mean that the data was based on one underlying component. Principal components analyses were used to identify the

components underlying the data sets. The data were then regrouped until the condition of unidimensionality was met. The regrouped data were then entered into Rasch model equations and scale locations were calibrated for each dimension identified. Once scale locations were calibrated the data from different sources could be compared as they were on the same scale

Teachers' judgments were correlated with each student's achievement on the independent assessments, these were compared across the sample of teachers allowing a comparison of each teacher's application of success criteria and hence giving an indication of consistency of judgments across the sample of teachers.

The questionnaire asked respondents to choose their level of agreement from a 6-point Likert-type scale. The responses were scored using a polychotomous marking schedule, allocating between 1 to 6 points depending upon the level of agreement with the statement. A measure of teachers' beliefs for each of the four constructs was calculated from their answers to the questionnaire. The measure of teachers' beliefs for each of the four constructs was correlated with the teachers' measure of conservatism of application of success criteria from within the sample. Thus, identifying any relationship between teachers' beliefs about each of the four constructs and their application of success criteria. These were then correlated with the measure of effectiveness for each teacher to identify any possible relationship between beliefs and effectiveness of teaching over the period of the study.

Ethical considerations

Ethical approval to conduct this research was granted by the Human Ethics Committee of Victoria University of Wellington. A letter of introduction was e-mailed to all prospective participants to outline the purpose of the research, the timeline and the requirements of the participants throughout the research (see appendix 3).

Participation in the study was voluntary and confidentiality of participants' identities was ensured. No identifying features of the schools or teachers were

included in the write up or any subsequent dissemination of the research. Informed consent was given by all of the principals and teachers in the research. The consent included the right to withdraw from the research prior to the analysis of data (see appendix 4).

As the researcher was a practicing teacher and parent of primary school aged children, neither the school in which she worked, nor the school her children attended were included in the study. The researcher ensured that teachers were informed that the tool developed to measure student achievement was for research purposes only and was not intended to be suitable as part of classroom assessment regime. All students' assessments were either returned to participants or destroyed upon completion of the research. All teachers involved in the research were given a copy of the main findings of the research

Results

Analysis of the questionnaire data, designed to measure teachers' beliefs about effective teaching methods, the nature of students' ability when learning and the purpose of assessment in mathematics

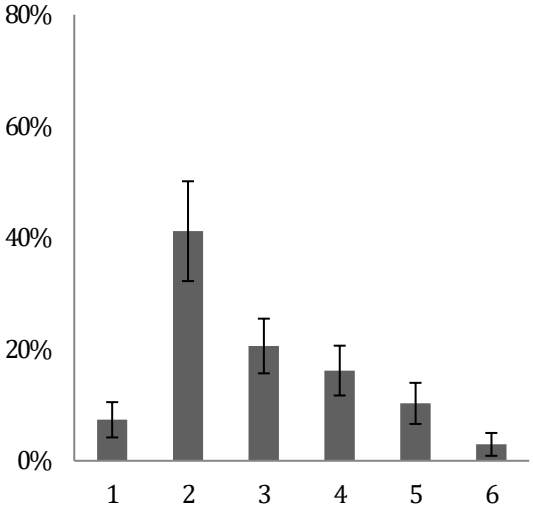
Data from the questionnaire (n= 68) designed to measure teachers' beliefs about the purpose of assessment in mathematics, effective pedagogy in mathematics, and the nature of students' ability when learning mathematics, were analysed using principal components analysis. An initial analysis of the complete data set of 40 questions identified 12 components with eigenvalues greater than one. Component one had an eigenvalue of 7.85 and explained 20% of the variance. Twenty four of the questions were loaded onto this component, including the complete set of eight questions designed to measure teachers' beliefs about students having a fixed mathematical ability and six of the questions designed to measure teachers' beliefs that a procedural approach is effective when teaching mathematics. The remaining dimensions loaded to a lesser degree onto this component. A second component had an eigenvalue of 5.54 and explained an additional 14% of the variance. Seventeen questions loaded this component. The third component in the data had an eigenvalue of 3.70, with 16 questions loading it, and explained 9% of the variance. The next three components each had eigenvalues close to 3 and each explained between 5% and 6% of the variance. Analysis of the questions did not provide any identifiable patterns to explain these loadings. For these reasons, it was decided to group the 16 questions of interest based on the dimension of teachers' beliefs they were designed to measure. The analyses for each of these groups are outlined below

Analysis of teachers' responses to the questions designed to measure the degree to which discrete and connectionist pedagogical beliefs

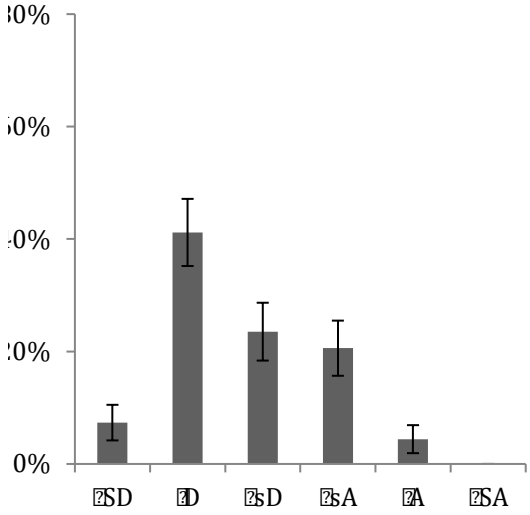
Table 3.1 shows the distribution of responses to each of the four questions designed to measure the teachers' agreement with the discrete pedagogical views and the connectionist pedagogical views expressed. Questions 1 and 31 were stating discrete pedagogical views, both suggesting that it was more effective to teach mathematical concepts and solution method separately. These questions both discriminate between respondents, with the majority tending towards disagreement with these views. Questions 11 and 21were designed to

measure teachers' connectionist pedagogical views; both stating that it was important to focus on the links and connections between mathematical concepts and solution methods. These two questions showed very small levels of discrimination between respondents. Both questions elicited very high rates of agreement from teachers.

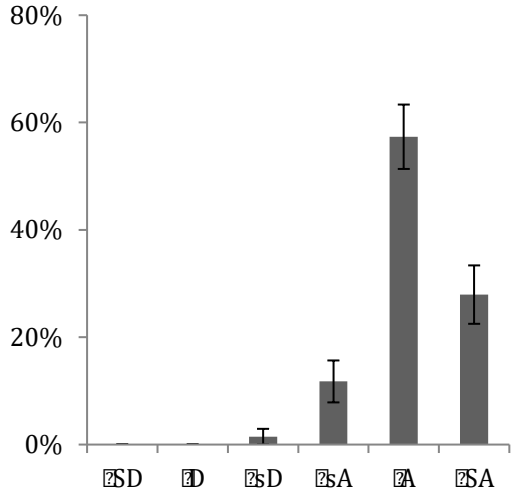
Table 3.1 The distribution of responses to the two statements (1 and 31) expressing discrete pedagogical views, showing teachers' levels of agreement. The distribution of responses to the two statements (11 and 21) expressing connectionist pedagogical views, showing teachers' levels of agreement. Error bars denote standard errors of the percentages. Legend: SD= strongly disagree, D= disagree, sD= slightly disagree, sA= slightly agree, A= agree, SA=strongly agree.



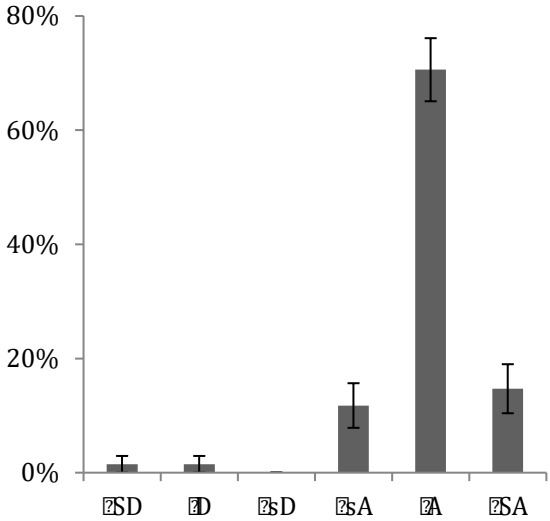
The different mathematical concepts and methods that the mathematics curriculum requires students to understand, are best taught separately.



Students learn mathematical concepts and methods best when they are taught separately.



It is important that students explore different solution methods for the same type of...



Making explicit, the links between different mathematical concepts and solution methods, is important for students' learning.

Table 3.2 shows the results from the principal components analysis conducted on this set of four questions. Questions 1 and 31 were designed to measure teachers' discrete pedagogical beliefs and questions 11 and 21, to measure teachers' connectionist pedagogical beliefs. Varimax rotation was used to give maximum loading values onto the components identified. Questions 1 and 31 both loaded onto the first component with an eigenvalue of 1.59 and explained 40% of the variance in the responses to this set of questions. Both questions have quite similar, high loadings. Questions 11 and 21 both loaded onto the second component with an eigenvalue of 1.18 and explained 29% of the variance. The loadings for these two questions were also similarly high.

This analysis identified that discrete pedagogical beliefs and connectionist pedagogical beliefs were loading onto separate components or factors. These two components explained 69% of the variance of the responses to the four questions. This result was unexpected. Research in this field (Askew et al., 1997; Baturo, 2004; Gill & Hoffman, 2009) supported the expectation that the four questions would predominantly load onto the same component with negative loadings for the two questions designed to measure teachers' connectionist pedagogical beliefs. Further analysing the questions in light of these results led to the conclusion that holding connectionist pedagogical beliefs does not preclude a teacher from also holding discrete pedagogical beliefs. It is possible to agree that mathematics concepts and methods are best taught separately whilst also agreeing that students need to explore different solution methods and that the links between concepts are made explicit. These four questions were therefore divided into two separate dimensions of beliefs. The questionnaire data from questions 1 and 31 were calibrated to a scale relating to discrete pedagogical views using a Rasch model and data from questions 11 and 21 were similarly calibrated to a scale relating to connectionist pedagogical views.

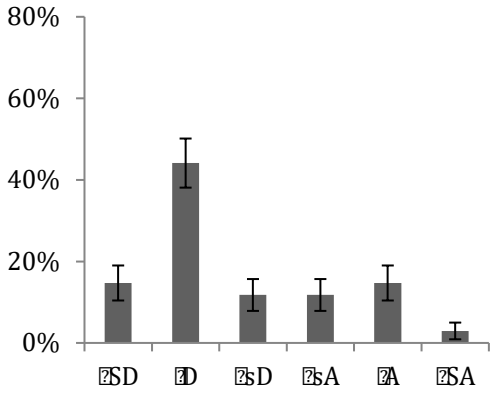
Table 3.2 Eigen values for each of the components identified through principal components analysis. Each question's relative loading onto the components identified. Component loadings < .3 are omitted.

	Component	
	1	2
(eigenvalues)	(1.59)	(1.18)
31. Students learn mathematical concepts and methods best when they are taught separately.	.841	
1. The different mathematical concepts and methods that the mathematics curriculum requires students to understand are best taught separately.	.829	
21. Making explicit, the links between different mathematical concepts and solution methods, is important for students' learning.		.828
11. It is important that students explore different solution methods for the same type of mathematical problem.		.819

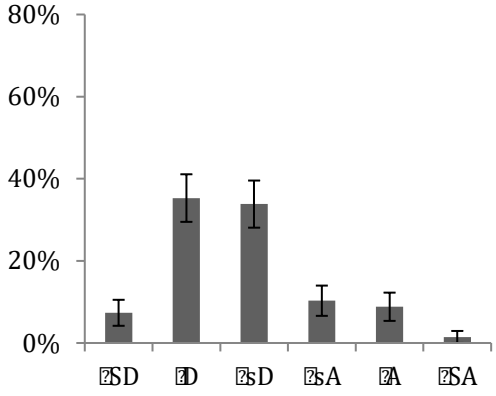
Analysis of teachers' responses to the questions designed to measure their procedural and conceptual pedagogical beliefs

Table 3.3 shows the response patterns to the set of four questions designed to measure teachers' agreement with the procedural pedagogical views and the conceptual pedagogical views expressed. The procedural pedagogical approach expressed by questions 2 and 32 stressed the importance of students learning facts and solution methods. These questions discriminate well between respondents with a range of responses being selected by the sample of teachers. The overall trend is towards disagreement with both questions. Questions 12 and 21 express the more conceptual view that stressed the importance of developing students' understanding and ability to reason mathematically. Question 12, while being strongly endorsed by the teachers in the sample, showed a range of responses. Question 22 however, was much less discriminating. The more moderate view expressed, elicited high levels of agreement from respondents.

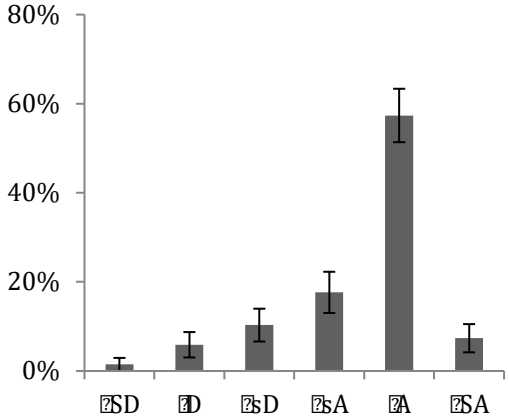
Table 3.3 The distribution of responses to the two statements (2 and 32) expressing procedural pedagogical views, showing teachers' levels of agreement. The distribution of responses to the two statements (11 and 21) expressing connectionist pedagogical views, showing teachers' levels of agreement. Error bars denote standard errors of the percentages. Legend: SD= strongly disagree, D= disagree, sD= slightly disagree, sA= slightly agree, A= agree, SA=strongly agree.



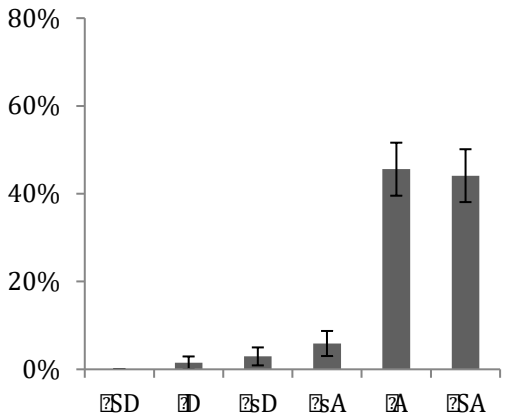
Educational progress in mathematics mostly involves learning facts and solution methods.



It is more important that teachers focus on mathematical facts and solution methods than on abstract concepts.



Students' demonstration of sound mathematical reasoning is more important than whether or not they produce correct answers.



It is more important that students understand why mathematical solution methods work than it is for them to memorise formulas.

Table 3.4 shows the results from the principal component analysis on the set of four questions designed to measuring teachers' procedural pedagogical beliefs and conceptual pedagogical beliefs. Two components were identified, with questions 2 and 23 loaded strongly onto component one. Questions 12 and 22 loaded weakly onto the first component, with question 12 having a small positive loading and question 22 having a small negative loading. Component one had an eigenvalue of 1.60. This component explained 40% of the variance in teachers' responses to this set of questions. Both questions 12 and 22 loaded much more strongly onto component two with an eigenvalue of 1.16, explaining 29% of the variance. These results were again unexpected, rather, one component was expected, with the questions designed to measure conceptual beliefs negatively loading. Having established that two, unidimensional components explained most of the variance in this set of questions, a Rasch model was used to fit the data for each of these components. The data from questions 2 and 32 were calibrated to a scale relating to procedural pedagogical beliefs using a Rasch model and the data from questions 12 and 22 were similarly calibrated to a scale relating to procedural pedagogical views.

Table 3.4 Eigenvalues for each of the components identified through principal components analysis. Each question's relative loading onto the components identified. Component loadings < .3 are omitted.

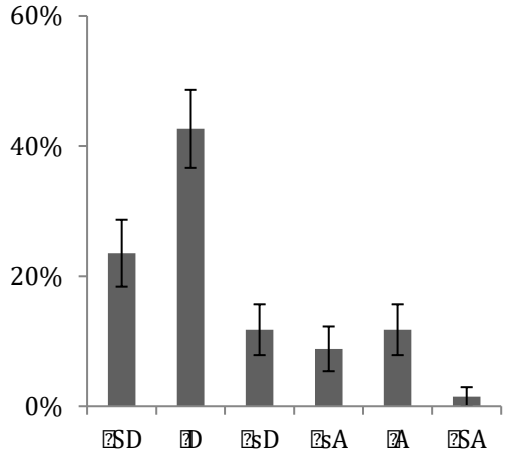
	Component	
	1	2
(eigenvalues)	(1.61)	(1.16)
2. Educational progress in mathematics mostly involves learning facts and solution methods.	.830	
32. It is more important that teachers focus on mathematical facts and solution methods than on abstract concepts.	.812	
12. Students' demonstration of sound mathematical reasoning is more important than whether or not they produce correct answers.	.334	.780
22. It is more important that students understand why mathematical solution methods work than it is for them to memorise formulas.	-.382	.730

Analysis of teachers' responses to the questions designed to measure their incremental and entity ability beliefs

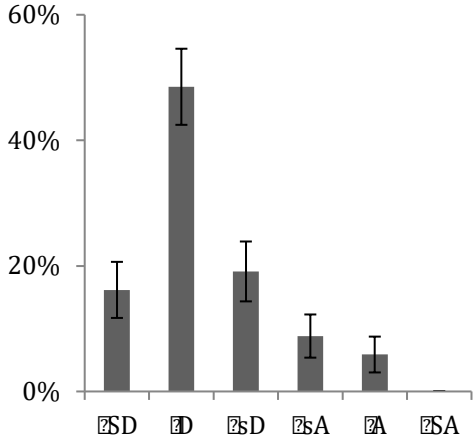
Table 3.5 shows the distributions of responses to the set of questions designed to measure teachers' beliefs about the nature of students' ability to learn mathematics. Questions 3 and 33 express the entity view of ability, that students have a relatively fixed, or innate level of mathematical ability. These two questions discriminate quite well, with an overall trend towards disagreement. Questions 13 and 23 express the incremental view of ability that most students who work hard are able to achieve in mathematics. These two statements have very different response patterns. Question 13 discriminates well between respondents, with a strong central tendency. Question 23 discriminates well with a strong tendency towards disagreement. The response pattern is quite similar to Question 33. There is a similarity in the phrasing of these two statements; both attribute students' "lack of achievement" in mathematics to either lack of effort or lack of ability. Teachers seem to be responding to these questions

similarly even though they are attributing the lack of achievement to very different causes.

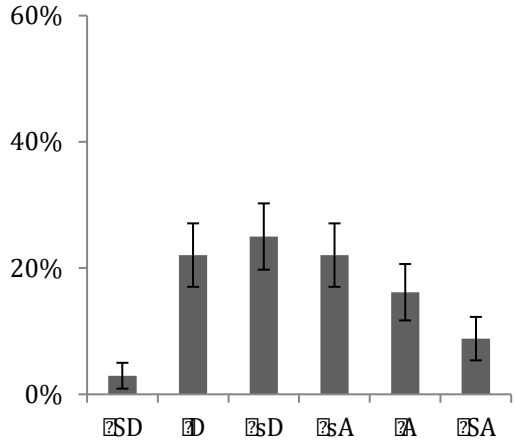
Table 3.5 The distribution of responses to the two statements (3 and 33) expressing entity views of students' ability. The distribution of responses to the two statements (13 and 23) expressing incremental views of students' ability. Error bars denote standard errors of the percentages. Legend: SD= strongly disagree, D= disagree, sD= slightly disagree, sA= slightly agree, A= agree, SA=strongly agree.



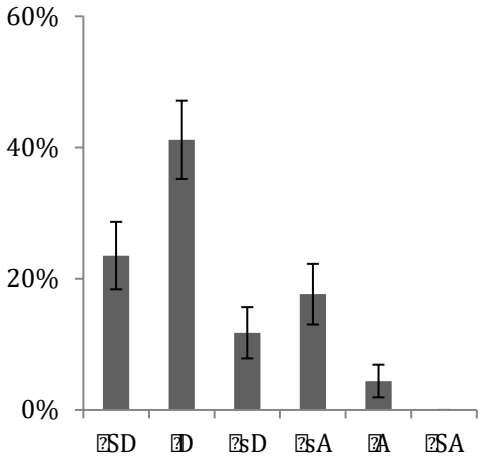
How good at mathematics a person is, is something that remains relatively fixed throughout life.



A students' lack of achievement in mathematics is usually attributable to the student not being good at mathematics.



All students can be excellent at mathematics if they work hard at it.



Students' lack of achievement in mathematics is usually attributable to lack of effort.

Table 3.6 shows the results of the principal components analysis for the set of four questions designed to measure teachers' beliefs about the nature of

students' ability to learn mathematics. Two components were identified. Component one explained 39% of the variance of this set of four questions, with an eigenvalue of 1.54. Question 33, designed to measure teachers' entity beliefs about students' ability, loaded strongly onto this component. Question 3 (see table 3.7) had a moderate loading and question 13 had a moderate, negative loading onto this component also. The second component had an eigenvalue of 1.04 (which is very close to 1) and explained 26% of the variance. Questions 13 and 23, designed to measure teachers' incremental beliefs about students' ability, were mixed in their loading onto this component; question 23 loaded very strongly and question 13 only moderately loaded.

Table 3.6 Eigen values for each of the components identified through principal components analysis. Each question's relative loading onto the components identified. Component loadings < .3 are omitted.

	Component	
	1	2
(eigenvalues)	(1.54)	(1.04)
33. A students' lack of achievement in mathematics is usually attributable to the student not being good at mathematics.	.857	
3. How good at mathematics a person is, is something that remains relatively fixed throughout life.	.619	
23. Students' lack of achievement in mathematics is usually attributable to lack of effort.		.906
13. All students can be excellent at mathematics if they work hard at it	-.493	.619

Another principal component analysis was performed with question 23 removed because the condition for unidimensionality was not met. Table 3.7 shows the results from the revised principal components analysis, which confirms the decision to remove question 23 and complete a second analysis. One dichotomous component was identified. This component explained 50% of the variance in this reduced set of three questions, with an eigenvalue of 1.49.

Question (33) loaded strongly onto this component. Question (13) now had a strong negative loading onto this component and question (3) had a moderate loading. From these results the data from questions 3, 13 and 33 were calibrated to a scale relating to entity beliefs using a Rasch model, with the data from question 13 reverse coded because it was negatively loaded.

Table 3.7 Revised eigenvalues for each of the components identified through principal components analysis of the data from three questions designed to measure teachers' entity and incremental ability beliefs. Components < .3 are omitted.

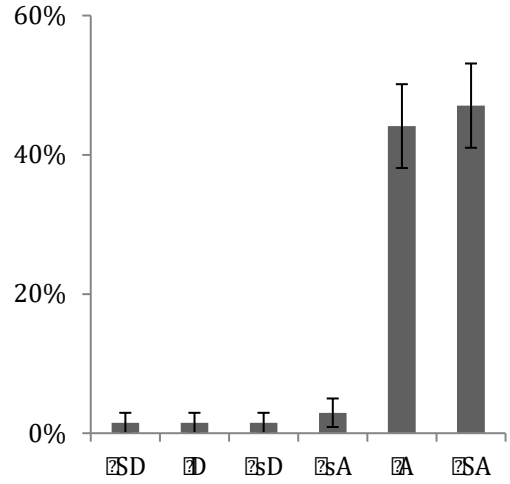
	Component 1
(eigenvalue)	(1.49)
33. A students' lack of achievement in mathematics is usually attributable to the student not being good at mathematics.	.821
13. All students can be excellent at mathematics if they work hard at it.	-.724
3. How good at mathematics a person is, is something that remains relatively fixed throughout life.	.544

Analysis of teachers' responses to the questions designed to measure their formative and summative assessment beliefs

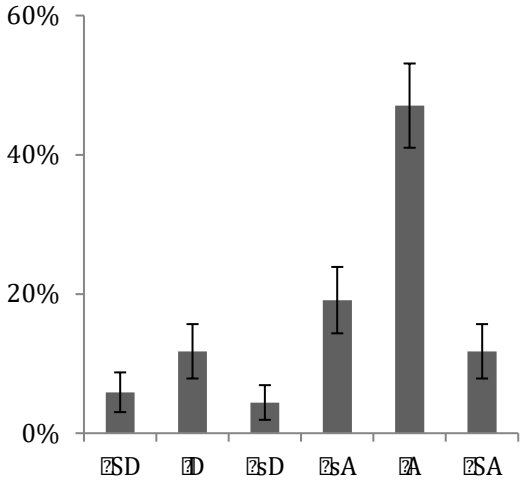
Table 3.8 shows the response patterns to the set of four questions designed to measure teachers' agreement with formative and summative assessment views. Questions 4 and 34 express the formative view that the primary purpose of assessment is to inform teaching and learning. This view is very strongly endorsed by the teachers in this sample. There is very little variance in responses to question 4. In comparison to the distribution for question 34, the response pattern shows much greater variability. Questions 14 and 24 are expressing the summative view of assessment that the primary purpose of assessment is for reporting. The response patterns from these two questions are quite similar; both are tending to disagree with the summative views expressed.

Both questions discriminate between respondents however; the responses to question 24 have less variance than those to question 14

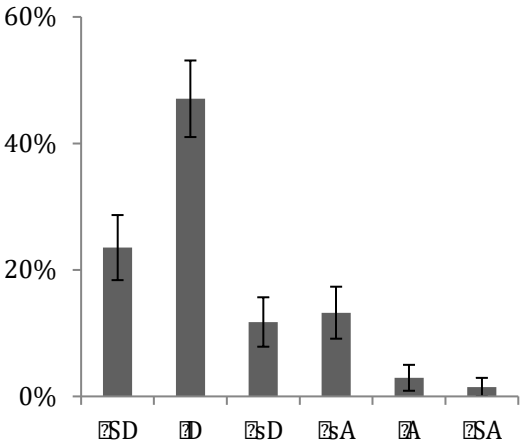
Table 3.8 The distribution of responses to the two statements (4 and 34) expressing formative views of assessment. The distribution of responses to the two statements (14 and 24) expressing summative views of assessment. Error bars denote standard errors of the percentages. Legend: SD= strongly disagree, D= disagree, sD= slightly disagree, sA= slightly agree, A= agree, SA=strongly agree.



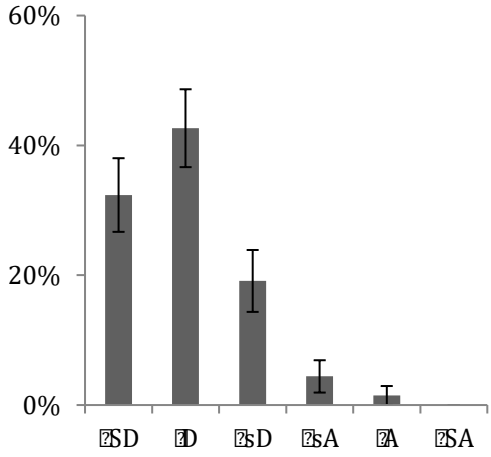
The primary purpose of assessment in mathematics is to inform teaching and learning.



The most valuable assessment in mathematics is that carried out prior to, or during a unit of work.



The primary purpose of assessment in mathematics is to report on students' achievement.



The most valuable assessment in mathematics is that carried out at the completion of a unit of work.

Table 3.9 shows the results from the principal component analysis on this set of four questions designed to measure teachers' beliefs about the primary purpose

of assessment. Two components were identified. Component one explained 36% of the variance with an eigenvalue of 1.44. Both of the questions designed to measure teachers' agreement with a summative view of assessment loaded onto this component. Question 14 loaded strongly and question 24 loaded moderately. Question 4, designed to measure formative assessment beliefs, had a moderate negative loading onto this component. Component two explained 28% of the variance, with an eigenvalue of 1.11. Question 34, which was designed to measure formative assessment beliefs was strongly loaded onto this component, question 24 was moderately negatively loaded onto this component. These results were unexpected, questions 4 and 34 were designed to measure the same view of assessment, and they did not load onto the same components. This shows that the two questions designed to measure teachers' formative views of assessment were not measuring the same component.

Table 3.9 Eigen values for each of the components identified through principal components analysis. Each question's relative loading onto the components identified. Components < .3 are omitted.

	Component	
	1	2
(eigenvalues)	(1.44)	(1.11)
14. The primary purpose of assessment in mathematics is to report on students' achievement.	.780	
24. The most valuable assessment in mathematics is that carried out at the completion of a unit of work**.**	.621	-.586
4. The primary purpose of assessment in mathematics is to inform teaching and learning.	-.593	
34. The most valuable assessment in mathematics is that carried out prior to, or during a unit of work.		.894

Another principal component analysis was performed with the data from the questions 14 and 24 only. Data from question 34 were not included as it loaded strongly onto component 2 and data from question 4 were not included as the distribution of responses shown in table 3.8 showed that it did not adequately

discriminate between respondents. Table 3.10 shows the results from the second principal components analysis. One component with an eigenvalue of 1.35 was identified, which explained 67% of the variance in the responses to these two questions. Both of the questions had strong, equivalent loadings onto this component. Based on these analyses, the data from questions 14 and 24 were calibrated to a scale relating to the degree that teachers believe that the primary purpose of assessment is for reporting, using a Rasch model.

Table 3.10 Revised eigenvalues for each of the components identified through principal components analysis of the data from two questions designed to measure teachers' summative assessment beliefs. Component < .3 are omitted.

	Component 1
(eigenvalue)	(1.35)
14. The primary purpose of assessment in mathematics is to report on students' achievement.	.820
24. The most valuable assessment in mathematics is that carried out at the completion of a unit of work**.**	.820

Six dimensions of beliefs were identified as unidimensional from the principal components analyses. The dimensions measured by the questionnaire included the degree to which teachers believe that the following four pedagogical approaches are effective when teaching mathematics; a discrete approach, a connectionist approach, a procedural approach and a conceptual approach. One dichotomous dimension of teachers' beliefs about the nature of students' ability to learn mathematics was identified, and one dimension of teachers' beliefs about the purpose of assessment in mathematics was identified. Teachers' scale locations for each of these dimensions were calibrated using Rasch models as described above. These calibrated data were used to calculate the relationships between the six dimensions of teachers' beliefs and their application of assessment criteria when making judgments about students' achievement against the mathematics standards.

Placing teachers on a scale from conservative to liberal based on their application of assessment criteria against the mathematics standards

Students' achievement data were analysed using principal components analysis. One dominant component was identified with an eigenvalue of 9.33, explaining 14 % of the variance in students' responses to the 66 items that made up the students' assessments. Forty-two items loaded onto this component. The analysis identified 23 additional components with eigenvalues greater than one. Each of these components explained very little of the variance in the data, the largest of these had an eigenvalue almost four times smaller than component one (component 2 had an eigenvalue of 2.5 and explained only 4% of the variance). The remaining components were also very closely grouped on the scree plot with a very shallow downward trajectory. For the purposes of calibrating measurement variables, it is permissible to assume unidimensionality under these conditions.

A Rasch model was used to calibrate these data as the requirement for unidimensionality was met. Data from each testing point and each of the three assessments were entered into separate Rasch model equations. These models calibrated scale locations for students at each testing point and for each assessment. This allowed comparison of students' achievement over time and across the three tests.

Correlations between teachers OTJ's and students' achievement

To calculate teachers' application of assessment criteria, their judgments were correlated with each student's achievement scale location at testing point 3. Spearman's rho (ρ) was used to calculate the correlation because teachers' judgments are ordinal data. The greater the value of ρ, the more aligned the teachers' judgment were with the independent measure of students' achievement (see table 3.14). Pseudonyms are used for all of the teachers involved in this research.

Table 3.14 shows the correlations between teachers' summative judgments and students' scale locations at each of the three testing points (TP's). Vivian, Niamh and Ken's judgments were most strongly correlated with the scale locations of

their students at testing point three. Janine and James' judgments were the least consistent. Janine had the lowest correlation between her judgments of students' achievement and their scale location, with virtually no correlation at testing point three. Correlations could not be calculated for Claire or Jacob because their school cross-grouped their students for mathematics on the basis of perceived mathematical ability at the beginning of the year. Cross-grouping refers to the practice of regrouping students within a cluster of classrooms all at the same level of the curriculum for specific subjects. The purpose of this practice is to reduce the range of perceived abilities within each class. This is a form of streaming for specific subjects, often mathematics. Claire taught all of the students identified as being of high mathematical ability. She judged all of these students to be above the standards in mathematics at the end of 2014. Jacob taught the middle group of students, those identified as being of average mathematical ability. His judgments were that each of the students was at the standard in mathematics at the end of 2014. This means that there was no variability in judgment in either of these classrooms, making it impossible to identify any correlation between students' achievement and these teachers' judgments.

Table 3.11 shows how stable the correlations are between teachers' summative judgments and students' scale locations at the three testing points. Three main patterns are evident from this analysis, teachers for whom the correlations increase over time, those for whom the correlations decrease over time and those for whom the correlations remain relatively stable over time.

Table 3.11 Correlations between teachers' summative judgments and students' achievement by scale location at each testing point.

School	Teacher	Correlation with teachers' summative judgments (ρ) Scale location at testing point 1 (N)	Scale location at testing point 2 (N)	Scale location at testing point 3 (N)
F	Vivian	.639 (14)	.853 (11)	.700 (15)
D	Niamh	.702 (24)	.792 (23)	.634 (26)
C	Ken	.331 (26)	.382 (22)	.610 (14)
A	Tamara	.627 (14)	.658 (19)	.519 (16)
C	Sabine	.587 (30)	.479 (30)	.417 (32)
B	Daria	.494 (22)	.462 (19)	.398 (22)
A	Marcy	.383 (26)	.325 (24)	.356 (26)
G	Paula	.401 (18)		.314 (20)
A	Bridie	-.011 (24)	.035 (23)	.288 (22)
E	Dillon	.491 (11)		.282 (22)
H	James	.625 (25)	.562 (24)	.110 (26)
B	Janine	.245 (27)	.153 (32)	.017 (36)
A	Claire			
A	Jacob			

Note: blank cells indicate missing data for Dillon and Paula. Correlations were unable to be calculated for Claire and Jacob as summative judgments were a single value.

Teachers for whom the correlation between their judgments and students' scale locations increased with time are Bridie and Ken. This suggests that these two teachers might have been weighting their judgments more on students' recent achievement, than on past achievement. There is a marked increase in the strength of the correlations between Ken's summative judgments and students' scale locations. The correlations increase from moderate to strong over the three testing points. Bridie's judgments go from no correlation with students' scale location at testing point one, to being weakly correlated by testing point 3.

Teachers whose correlation between judgments with achievement decreases with time include Janine, James, and to a lesser degree, Dillon. This suggests that these teachers may be giving greater weight to students' achievement measured earlier in the year than at the end of the year. It should be noted that due to incomplete data sets for Dillon and Paula it was only possible to compare the scale locations at testing point 1 and testing point 3. One possible reason for teachers to give greater weight to assessments of students' achievement from earlier in the year is they may be placing increased weight on data provided by the Progressive Achievement Tests. Traditionally in New Zealand, most primary schools administer the Progressive Achievement Tests between February and March. These are norm referenced, standardised assessments of students' achievement based on the New Zealand curriculum. Teachers may give substantial weight to these assessments when making judgments of students' achievement, even though they were administered at the beginning of the year.

The remaining teachers, Vivian, Niamh, Tamara, Sabine, Daria, Marcy and Paula, have correlations between summative judgments and students' achievement that are relatively consistent across the three testing points. Of particular note is the very high correlation between summative judgments and students' achievement for Vivian and to a lesser extent Niamh. Such high correlations are consistent with teachers basing their judgments solely on tests rather than using the students' achievement as assessed formatively as part of the classroom programme.

Placing teachers on a scale from conservative to liberal based on their application of assessment criteria against the mathematics standards

Table 3.12 shows comparison of the mean scale locations for each level of judgment made by each teacher. This allows an identification of the teachers in the sample who were conservative, and those who were liberal, in their application of assessment criteria. Some teachers were inconsistent in their application of assessment criteria, see Ken for example; he was liberal in his application of assessment criteria when making "below" and "above" judgments, but not when making "at" judgments. For this reason, it was decided to focus

solely on the mean scale location at one level of judgment to determine the degree to which teachers were conservative or liberal in their application of assessment criteria.

There were three main reasons for choosing the mean scale location for teachers' judgment of "at the standard" as the point at which to determine the degree to which teachers are conservative or liberal in their application of assessment criteria. First, with the exception of Claire, "at" judgments were made by all of the teachers in the sample. Second, this was the judgment level made about the largest group of students (n = 152). Third, there was less likelihood of students' assessment data being affected by floor or ceiling effects at this level of judgment.

Janine was the most conservative in her application of assessment criteria when making "at the standard" judgments. Of note is that the mean scale location for students she judged as being "at the standard" is above mean location for students she judged "above the standard". Sabine was conservative at all judgment levels compared with the other teachers. Marcy appeared conservative when making judgments at the standard, however this may be a function of the ability grouping for mathematics at her school. Marcy's class comprised students whose achievement in mathematics was perceived to be above average. There are likely to be less students achieving in the lower range so it is likely that the mean scale location of "at the standard" may be higher because of the selection of students. Dillon was also relatively conservative when making "at the standard" judgments. Niamh was the most liberal in her application of assessment criteria at all judgment levels when compared to the other teachers. Tamara and Vivian were both more liberal when making "at the standard" judgments.

Table 3.12 Comparison of mean scale location at testing point three for each judgment level made by teachers. Sorted by teachers' "at the standard" summative judgments to give a scale location for their application of assessment criteria from conservative to liberal in bold.

School	Teacher	Mean scale location for summative judgments (logits)				Conservative/ liberal
		Well below	Below	**At**	Above	
B	Janine			**1.09**	1.05	conservative
C	Sabine	-.65	.29	**.70**	2.36	conservative
A	Marcy			**.61**	1.15	conservative
E	Dillon			**.58**	1.12	conservative
A	Jacob			**.11**		
H	James		.01	**.02**	.16	
C	Ken		-1.54	**-.12**	.19	
B	Daria		-.54	**-.13**	.54	
G	Paula		-1.36	**-.25**		
A	Bridie		-.95	**-.29**		
F	Vivian	-2.28	-.69	**-.41**	1.01	liberal
A	Tamara	-2.30	-1.08	**-.49**		liberal
D	Niamh		-1.71	**-1.17**	.60	liberal
A	Claire				1.42	

Note: Unable to compare Claire's data as there were no "at the standard" judgments made.

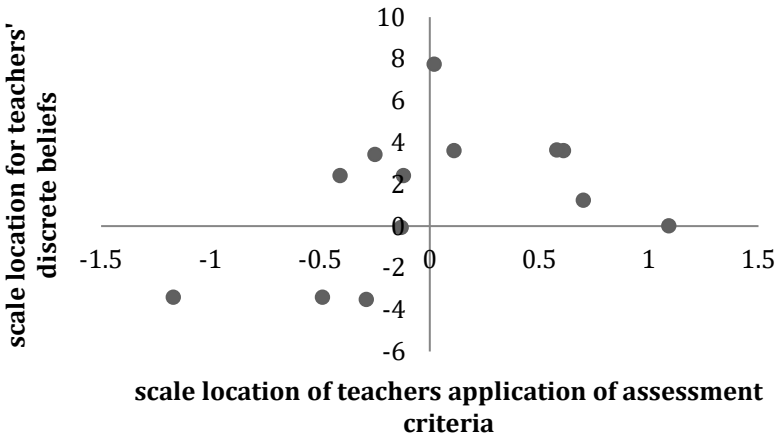
Correlations between teachers' application of assessment criteria when making "at the standard" judgments, from conservative to liberal, and each dimension of teachers' beliefs

The scale locations for teachers' application of assessment criteria when making "at the standard" judgments placed the teachers on a continuum from conservative to liberal. These scale locations were correlated with the scale locations for the six dimensions of teachers' beliefs. Table 3.14 shows the correlation coefficient for each correlation between the six dimensions of

teachers' beliefs and teachers' application of assessment criteria when making "at the standard" judgments. The correlations were explored graphically using scatterplots. Figure 3.1 shows the scatterplot of the correlation between the scale location of teachers' application of assessment criteria and the scale location of teachers' discrete pedagogical beliefs. A correlation of .41 (see table 3.14) was calculated between these two variables. This suggests that teachers who hold discrete pedagogical beliefs tend to be more conservative in their application of assessment criteria when making an "at the standard" judgment against the mathematics standards.

Figure 3.1 Correlation between teachers' discrete pedagogical beliefs and teachers' application of success criteria from liberal to conservative.

Legend: ? ● = teacher?



A correlation between teachers' connectionist beliefs and their application of assessment criteria was r = -.29 (see table 3.14). Figure 3.2 shows the scatterplot of the correlation between the scale location of teachers' application of assessment criteria and the scale location of teachers' connectionist pedagogical beliefs. This shows that the data has very little variation in the measure of teachers' connectionist beliefs as was evident in table 3.4. Investigation of the teachers' responses to the questionnaire show that 11 of the 14 teachers selected "agree" to both question 11 and question 21, which were stating connectionist views. The correlation coefficient shows that these two measures have a weak negative correlation, which suggests that teachers with stronger

connectionist views would tend to be less conservative (or more liberal) with their application of assessment criteria.

Figure 3.2 Correlation between the scale location of teachers' connectionist pedagogical beliefs and the scale location of teachers' application of success criteria from liberal to conservative.

Legend: = teacher

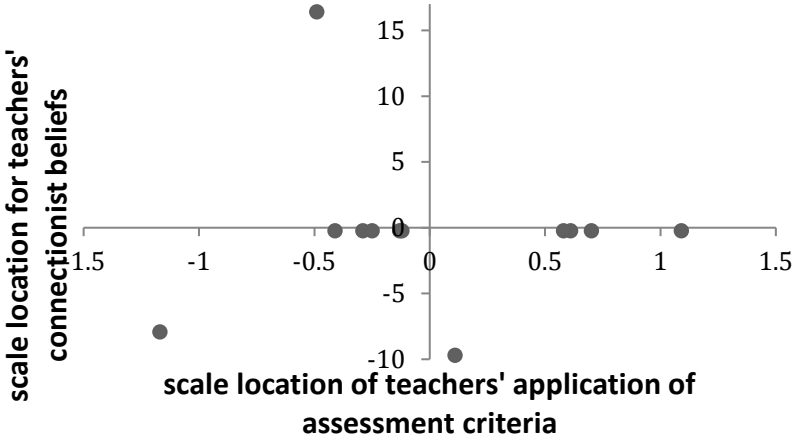


Figure 3.3 shows the scatterplot of the correlation between the scale location of teachers' application of assessment criteria and the scale location of teachers' procedural pedagogical beliefs. A weak correlation of .292 was calculated (see table 3.14) between these two variables. This suggests that teachers who hold procedural pedagogical beliefs tend to be more conservative in their application of assessment criteria when making an "at" judgment against the mathematics standards.

Figure 3.3 Correlation between teachers' procedural pedagogical beliefs and teachers' application of success criteria from liberal to conservative.

Legend: = teacher

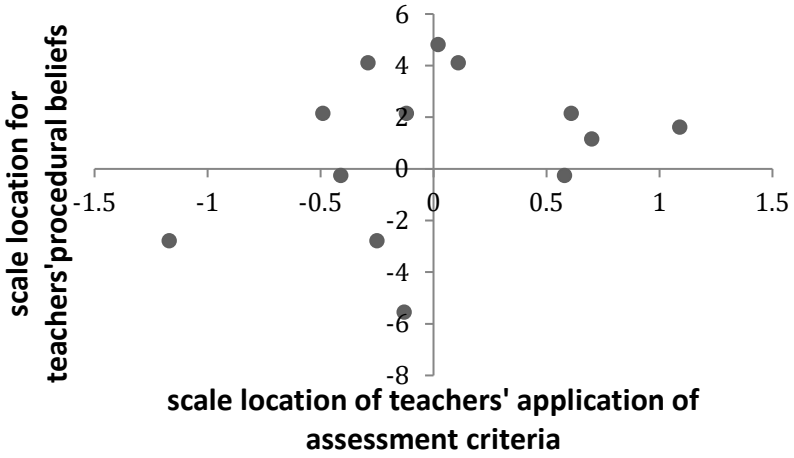


Figure 3.4 shows the scatterplot of the correlation between the scale location of teachers' application of assessment criteria and the scale location of teachers' conceptual pedagogical beliefs. A weak, negative correlation of -.127 was calculated between these two variables. This suggests that teachers with more conceptual pedagogical beliefs tend to be more liberal in their application of assessment criteria when making "at" judgments against the mathematics standards.

Figure 3.4 Correlation between teachers' conceptual pedagogical beliefs and teachers' application of success criteria from liberal to conservative.

Legend: = teacher

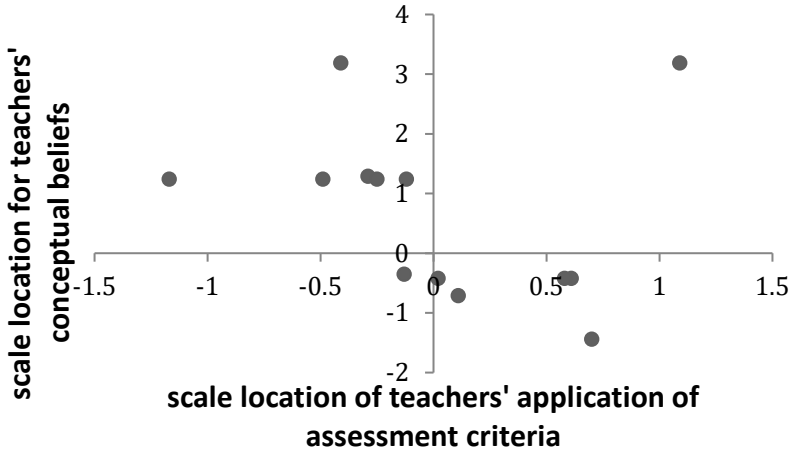


Figure 3.5 shows the scatterplot of the correlation between the scale location of teachers' application of assessment criteria and the scale location of teachers' entity beliefs of students' ability in mathematics. A weak correlation, which was calculated .190, is shown between the scale location of teachers' application of assessment criteria and the scale location for teachers' entity ability beliefs. This implies that teachers who hold entity ability beliefs may tend to be more conservative in their application of success criteria when making an "at" judgment against the mathematics standards.

Figure 3.5 Correlation between teachers' entity ability beliefs and teachers' application of success criteria from liberal to conservative.

Legend: = teacher

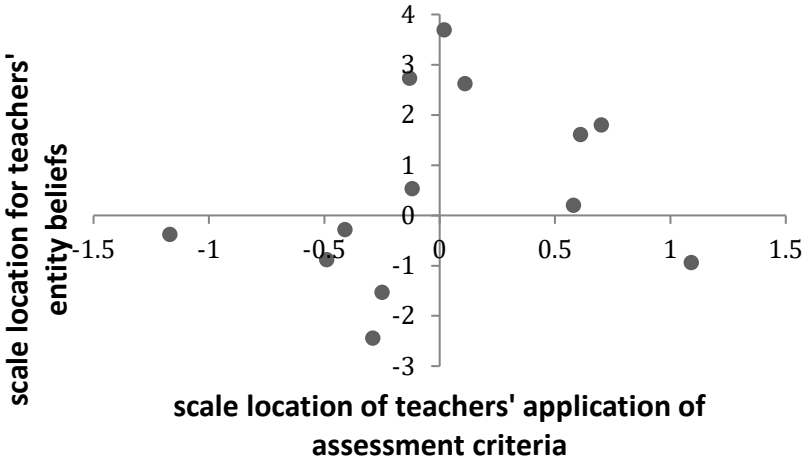
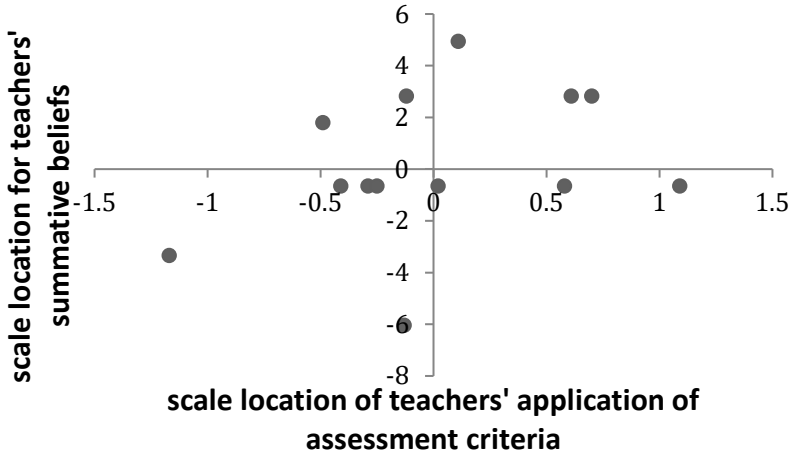


Figure 3.6 shows the scatterplot of the correlation between the scale location of teachers' application of assessment criteria and the scale location of teachers' connectionist pedagogical beliefs. A correlation of .338 was calculated (see table 3.14) between these two variables. This implies that as teachers tend towards more summative assessment beliefs they tend to be conservative in their application of success criteria when making "at" judgments against the mathematics standards.

Figure 3.6 Correlation between teachers' summative assessment beliefs and teachers' application of success criteria from liberal to conservative.

Legend: = teacher



From this analysis, a set of pedagogical, assessment and ability beliefs have been identified as being associated with teachers tending towards a more conservative application of assessment criteria when making judgments against the mathematics standards. These include teachers' beliefs that a more discrete and a more procedural approach is effective when teaching mathematics, an entity belief about the nature of students' ability and a summative belief about the assessment of mathematics. Conversely teachers who hold connectionist and conceptual pedagogical beliefs and more incremental beliefs about student' ability to learn mathematics tend to be more liberal in their application of assessment criteria when making "at the standard" judgments in mathematics.

Table 3.13 compares the responses to the questionnaire for each of the six dimensions of teachers' beliefs measured in this research. The agreement and disagreement sub headings have been collapsed to give a general indication of the agreement or disagreement for comparison between the two groups of teachers. Disagreement contains indications of strong disagreement, disagreement and slight disagreement; the same is true for agreement. From this table, it is clear that the four teachers who have been identified as conservative, on the whole, disagree with discrete pedagogical views, but do so less strongly

than the three teachers identified as liberal. The purpose of this table is to give a realistic perspective to the results of the correlations. Both sets of teachers strongly endorse a connectionist pedagogical approach and disagree with a summative assessment approach; the relative strength of the disagreement will be the difference between the two groups.

Table 3.13 Comparison of the percentage of agreement/disagreement with each of the six dimensions of beliefs between teachers identified as conservative or liberal.

	Teachers' application of assessment criteria	
Teachers' beliefs	Conservative	Liberal
Discrete	62.5% Disagreement	83% Disagreement
Connectionist	100% Agreement	100% Agreement
Procedural	75% Disagreement	83% Disagreement
Conceptual	87.5% Agreement	100% Agreement
Entity-Incremental	73% Incremental	67% Incremental
Summative - Formative	100% Formative	100% Formative

Relationships between increased gains in students' achievement and the six dimensions of teachers' beliefs

Table 3.14 shows correlations between each of the six dimensions of teachers' beliefs and students' progress. The strongest of these is a strong negative correlation between entity ability beliefs and conceptual pedagogical beliefs. As entity and incremental beliefs are dichotomous that means that there is a positive correlation between conceptual beliefs and incremental beliefs. Teachers, who hold conceptual pedagogical beliefs, are likely to believe that it is very important to develop students' understanding of mathematical concepts and how procedures work. Teachers who hold strong incremental ability beliefs are likely to believe that nearly all students who work hard are able to achieve success in mathematics given effective teaching.

Moderate correlations were shown between entity ability beliefs and discrete pedagogical beliefs, between summative view of assessment and procedural

pedagogical beliefs, Teachers with entity ability beliefs view students' mathematical ability as remaining relatively fixed throughout their lives. Teachers with discrete pedagogical views believe that it is important to teach different mathematical ideas and methods separately. This finding is supported by Gill and Hoffman (2009). The correlation between summative assessment views and procedural pedagogical beliefs was hypothesised, as teachers with procedural pedagogical beliefs are likely to focus on learning correct solution methods and procedures, which is consistent with a more summative view of assessment. This view tends to rely on testing at the completion of a unit of work to report on the levels of achievement.

Moderately negative correlations were shown between connectionist pedagogical beliefs and procedural pedagogical beliefs. These two pedagogical views, as described above, would be expected to have a negative correlation. One belief is focused on making the links and connections between mathematical ideas explicit, while the other is focused on learning the correct procedures to solve problems. As teachers tend towards connectionist views they are less likely to hold procedural views. A moderate negative correlation between summative assessment beliefs and connectionist pedagogical beliefs was also found. As both beliefs are described above, it is reasonable to expect that as teachers' connectionist pedagogical beliefs increase, their assessment beliefs would tend to be less likely to be summative. A moderate negative correlation was found between discrete and conceptual pedagogical views. It is reasonable to expect a negative correlation between these two views, described above. One view is focused on teaching mathematical ideas separately and the other belief is concerned with developing students' understanding of mathematical concepts and ideas. As teachers tend towards having discrete views, they are less likely to hold conceptual pedagogical beliefs.

Connectionist pedagogical beliefs were weakly negatively correlated with discrete pedagogical beliefs. The design of the questionnaire predicted a negative correlation between these two pedagogical beliefs. These two beliefs describe different ways of viewing similar pedagogical focus. Both are concerned with the

degree to which the teacher chooses to link or separate ideas in mathematics. The strength of the association is due to the two dimensions both being appropriate for different purposes. Teachers wishing to focus on specific concepts or procedures may be more likely choose a more discrete approach. Conversely, when wanting to focus on developing overarching ideas in mathematics, encouraging students to generalise from the specifics of what has been taught, teachers may opt for a more connectionist approach. Teachers who explicitly identify the connections between ideas in mathematics may well, at times, choose to focus on a more discrete approach to effectively teach a particular concept or method and the converse applies also.

There is no correlation between connectionist and conceptual pedagogical beliefs, which is unexpected. This association was reported by Askew et al. (1997). A connectionist teacher makes links between concepts and procedures explicit and focuses on connecting ideas for students. A conceptual pedagogical approach focuses on the underlying concept and the reasons why mathematics ideas work the way they do. Conceptual pedagogical approach is associated with multiple solution methods and students developing a conceptual understanding of mathematics. Both beliefs are concerned with developing students' understanding by focusing on the big ideas underpinning mathematics.

Of particular concern for this study were the relationships between the six dimensions of teachers' beliefs and students' progress in mathematics. There were three weak correlations worth noting. The largest of these is the weak relationship between entity beliefs about the nature of students' ability and students' progress in mathematics. Teachers' discrete and procedural pedagogical beliefs were also weakly associated with students' progress. Given these three correlations it is unsurprising that there was also a weak correlation between teachers' conservative application of assessment criteria and students' progress, which suggests that students may make more progress in classrooms where the teachers are more conservative in their application of assessment criteria.

Table 3.14 Correlations between increased gains in student achievement, teachers' application of assessment criteria and the six elements of teachers' beliefs.

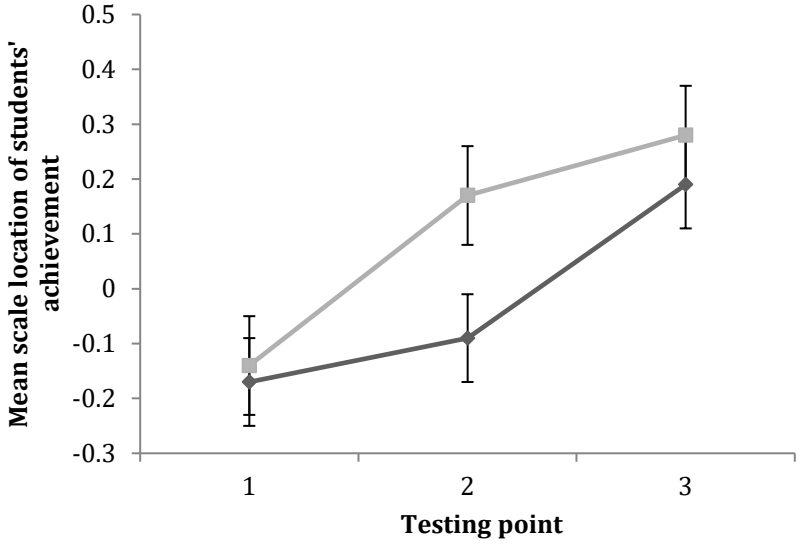
	2.	3.	4.	5.	6.	7.	8.
1. Progress	**.11**	**.13**	**.10**	**.11**	**-.06**	**.19**	**.03**
2. Application of assessment criteria		.41	-.29	.29	-.23	.24	.34
3. Discrete			-.28	.24	-.38	.62	.12
4. Connectionist				-.53	.02	.25	-.51
5. Procedural					-.10	.10	.57
6. Conceptual						-.69	-.10
7. Entity -Incremental							.03
8. Summative							

Analysis of variance of students' achievement by gender, year level and cross-grouping

Figure 3.7 shows the relative mean scale location of students' achievement grouped by gender over the three testing points. The learning trajectory of the two groups differed slightly, however there was no significant difference between the achievement of females and males at testing point 1 and testing point 3.

The data were analysed using a repeated measures ANOVA. Students' scale locations at each of the three testing points were the dependent variable, testing point was a within-subject factor and gender was a between-subject factor. The analysis determined that testing point was a significant factor $F(2, 228) = 14.86$, $p < .001$ $\eta^2 = .115$, evincing an increase in students' achievement over the three testing points. The analysis determined that gender was not a significant factor, $F < 1$, $p = .396$. There was no significant interaction between testing point and gender; $F < 0$, $p = .678$.

Figure 3.7 Mean scale location of students' achievement by gender over the three testing points. Error bars denote standard error of the mean.
Legend: ◆= females, ■ = males

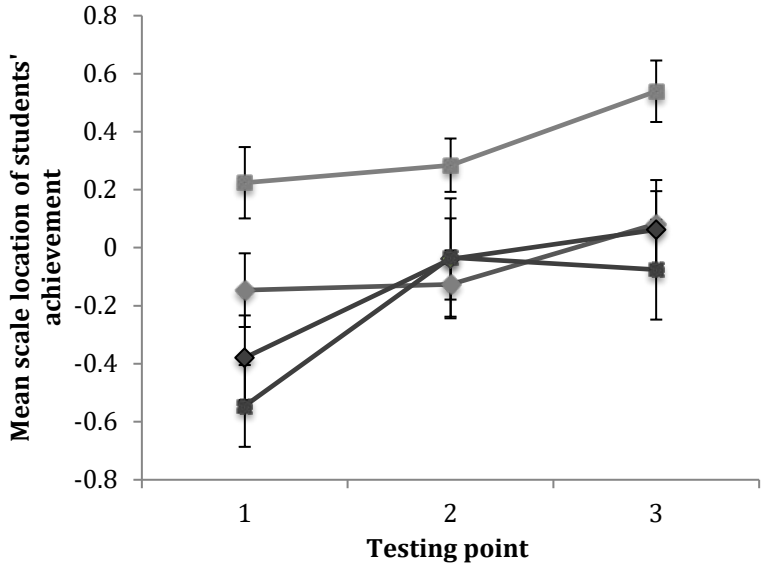


There were differences in the way students were grouped for the teaching of mathematics across the schools participating in the research. School A cross-grouped students for mathematics based on their perceived ability at the start of the year. School B separated students into strict year levels as well as cross grouping for mathematics. Schools C, D, E, F, G and H grouped students within their classes. To identify if grouping was a factor in students' gain in achievement, the data were analysed with both grouping and year level as between-subject factors. This analysis was to identify any possible effect on students' achievement these differences in grouping may have.

Figure 3.8 shows the relative mean scale locations of students over the three testing points based on their year level and groupings for mathematics. The data were analysed using a repeated measures ANOVA. Students' scale locations at each of the three testing points was the dependent variable, testing point was the within-subject factor and grouping and year level were between subject factors. The analysis determined that testing point was a significant factor, $F(2, 226) = 14.99$, $p < .001$ $\eta^2 = .117$, with students' achievement increasing over the three testing points. The analysis determined that grouping was a significant factor, $F(1, 227) = 6.55$, $p = .011$ $\eta^2 = .028$, with a significant difference in students'

achievement over the three testing points based on their grouping for mathematics. Students who were cross-grouped for mathematics having greater achievement than those grouped within their class. The analysis determined that year level was not a significant factor, $F(1, 227) = 1.66$, $p = .199$. There was a significant interaction between testing point and grouping, $F(2, 226) = 4.60$, $p = .011$ $\eta^2 = .039$, which indicates that students increased achievement over the three testing points is related to their grouping in mathematics. The students grouped within their class made significantly more progress (mean scale location at testing point 3 less mean scale location at testing point 1), with an increase in mean scale location of .46 logits compared to an increase in mean scale location of .28 logits for the students who were cross grouped for mathematics. There was a significant interaction between grouping and year level, $F(1, 227) = 4.43$, $p = .036$, $\eta^2 = .019$, which indicates that the increase in students' achievement due to grouping was dependent upon the year level. There was no significant interaction between testing point and year level, $F < 1$, $p = .712$. There was no significant interaction between testing point, grouping and year level, $F < 1$, $p = .720$

Figure 3.8 Mean scale location of students' achievement by grouping and year level over the three testing points. Error bars denote standard error of the mean. Legend: ■ = Year 6, ◆ = Year 5, _____ = cross grouping, _____ = within class grouping.



Discussion

This research found evidence of relationships between teachers' beliefs about elements of effective pedagogy in mathematics, the purpose of assessment and the nature of students' mathematical ability and their application of assessment criteria when making "at the standard" judgments against the mathematics standards. One weak relationship was identified between teachers' entity beliefs of students' mathematical ability and students' increased achievement in mathematics.

Before discussing the relationships found it is necessary to first discuss the findings from the principal components analyses of the dimensions of beliefs measured by the questionnaire.

Four dimensions of beliefs become six dimensions of beliefs

The principal components analysis of the questionnaire data showed very clearly that connectionist pedagogical views were measuring a separate component to the discrete pedagogical views. This was unexpected because the discrete views were written to be the counter viewpoint to the connectionist views and were expected to load negatively onto the same component as the connectionist questions. This suggests that, instead of conceiving of connectionist and discrete pedagogical approaches in an exclusive way, teachers appeared to view these two approaches inclusively. Teachers appeared to believe that effective mathematical pedagogy combines the two approaches to some degree, however connectionist views were much more strongly endorsed than discrete views. It is perhaps more correct to say that teachers appeared to view a predominantly connectionist approach with discrete elements when appropriate as an effective pedagogical approach to the teaching of mathematics. A moderate negative correlation was found between these two beliefs, which suggests that as teachers increasingly favour one of these pedagogical approaches they are less likely to endorse the other.

The results from the analysis of the set of questions designed to measure teachers' procedural and conceptual pedagogical views showed that the

questions stating these two views were measuring separate components. These results suggested that teachers viewed these two pedagogical approaches inclusively. Once again, this result was unexpected. Teachers' endorsement of a conceptual approach was much stronger than a procedural approach, suggesting that teachers viewed effective pedagogy in mathematics as being largely focused on developing students' conceptual understanding whilst ensuring that students also develop a degree of procedural proficiency. There was a weak negative correlation between the two approaches, which suggests that teachers viewed both as effective pedagogical approaches to use as part of their mathematics programme. Favouring one approach does not necessarily preclude the use of the other. Rittle-Johnson and Koedinger (2009) used an inclusive approach in their research, which explored how to effectively incorporate both the development of conceptual understanding and the building of procedural knowledge into a sequence of lessons.

The analysis of the set of questions designed to measure teachers' entity and incremental beliefs about students' ability in mathematics provided evidence of a dichotomous relationship between these two beliefs. Once again data from the four questions loaded onto two components. Interestingly, both of the questions asking about students' lack of achievement in mathematics, one stating an incremental view and one stating an entity view, received very low levels of agreement. This seems to indicate that teachers were responding to the term 'lack of achievement' rather than the incremental or entity views expressed. One possible explanation for this may be teachers' appreciation of the multiple reasons for students' potential difficulties and were therefore reluctance to attribute students' lack of achievement to a single factor. Teachers appeared to be reluctant to endorse an entity view and tended to be mixed in their endorsement of an incremental view. Teachers may find both an incremental and entity view of students' ability inadequate in describing the nature of students' ability.

The analysis of the set of questions designed to measure teachers' beliefs about the purpose of assessment in mathematics identified two components, however

the data from questions loaded onto these components in unexpected ways. The two questions designed to measure formative views loaded onto separate components. There was evidence of summative and formative views being dichotomous as the formative questions loaded negatively when the summative questions were positive and vice versa. Because of this, teachers would be likely to endorse either a formative or summative view and, as mentioned earlier 100% of teachers endorsed a formative view of assessment, so it is unsurprising that very few teachers endorsed a summative view of assessment.

Six dimensions of teachers' pedagogical beliefs were identified from the questionnaire data, a connectionist pedagogical belief, a discrete pedagogical belief, a conceptual pedagogical belief, a procedural pedagogical belief, an entity ability belief and a summative assessment belief. Of interest was teachers' almost universal endorsement of the questions describing connectionist and conceptual pedagogical views and formative assessment views. Of the 14 teachers involved in the main part of the research there was 100% agreement with the two questions stating connectionist views and the questions stating formative assessment views. Teachers also strongly endorsed the questions describing conceptual pedagogical views. These three beliefs are strongly endorsed by the Ministry of Education (Anthony & Walshaw, 2007; Ministry of Education, 2007, 2009a, 2011). Teachers may be more likely to agree with the views that they recognise as being endorsed by the Ministry of Education, because of wanting to conform to the views that they perceive as being valued by their profession.

Teachers' beliefs about effective pedagogical approaches to the teaching of mathematics, their beliefs about the primary purpose of assessment and the nature of students mathematical ability were inferred from teachers' responses to questions about each of these dimensions. Pajares (1992) argues that explorations of teachers' inferred beliefs are complimented by investigation of teachers' enacted beliefs, from analysis of their classroom practice, in order to fully understand the nature of the beliefs held by teachers. It is recommended

that further research into these dimensions of teachers' beliefs be accompanied by analysis of their classroom practice to enhance the validity of the findings.

Relationships between teachers' application of assessment criteria, either conservative or liberal, and each of the six dimensions of teachers' beliefs

Moderate correlations were found between teachers' degree of conservatism in their application of assessment criteria and endorsing a discrete pedagogical view, also between teachers' degree of conservatism and endorsing a summative view of assessment. Weak correlations were also found between teachers' degree of conservatism and endorsing a procedural pedagogical approach, and between teachers' degree of conservatism and an entity view of students' mathematical ability. Bearing in mind that the connectionist pedagogical view and a formative assessment view received universal endorsement, this suggests that teachers who tend towards a more conservative application of assessment criteria were also more moderate in their beliefs about effective pedagogy, the purpose of assessment and the nature of students' mathematical ability. These teachers appear to have a balanced approach to the teaching of mathematics. The findings suggest that they view the four pedagogical approaches explored inclusively, which may indicate that they select the approach that best suits the needs of the students. As teachers tend towards a more conservative application of assessment criteria they also tend towards an entity view of the nature of students' ability in mathematics.

Teachers who tended towards a more liberal application of assessment criteria tended to more strongly endorse pedagogical views that were connectionist and conceptual, they were also less likely to endorse pedagogical views that were procedural and discrete. These teachers tended to more strongly endorse an incremental view of students' ability in mathematics. They were also less likely to endorse a summative approach to assessment. These views are closely aligned with the pedagogical approaches and the assessment approach that are advocated by the Ministry of Education (Ministry of Education, 2007, 2011).

Teachers' degree of conservatism in their application of assessment criteria when making judgments of students' achievement in mathematics appears to be

related to a set of beliefs about effective pedagogy, the purpose of assessment and the nature of students' mathematical ability. The degree of conservatism in teachers' application of assessment criteria is an important factor in describing the variability of teachers' judgments. How consistent teachers are in their application of assessment criteria is the other important factor. Enabling teachers to identify how conservative they are, especially when paired with an indication of the consistency of their application of assessment criteria when making judgments against the mathematics standards may support teachers to reduce variability in the judgments they make against the National Standards. The Ministry of Education has developed a tool, which supports teachers to make increasingly consistent and reliable judgments over time. This is the Progress and Consistency Tool, also known as PACT (Ward & Thomas, 2015). This tool provides teachers with a richly illustrated framework, with annotated examples of students' work at each year level to support teachers make increasingly consistent and reliable summative judgments.

This research identified variability in the form of teachers' degree of conservatism in their application of assessment criteria when making judgments against the mathematics standards (see table 3.12). This finding supports the findings of Ward and Thomas (2015). Research suggests that social moderation is an important element in reducing variability in teachers' judgments (Connolly et al., 2012; Crisp, 2013; Sadler, 2009; Smaill, 2013; Wyatt-Smith & Klenowski, 2013) especially when teachers are given the opportunity to collaborate with colleagues identified as expert in making summative judgments of students' achievement. Developing an expert, novice relationship has been associated with reduced variability in teachers' judgments (Sadler, 2009).

What relationships are there between increased gains in student achievement and the six elements of teachers' beliefs?

The students' achievement gains were correlated with each of the six dimensions of teachers' beliefs. One weak correlation was found between students' achievement gains and teachers' entity beliefs about students' mathematical ability. This correlation needs to be put into perspective, the teachers in this research, on the whole, endorsed a more incremental view of students' ability in

mathematics this correlation is indicating that the teachers with more central views (less strongly endorsing an incremental view of students' ability) had the greatest increase in students' achievement. This finding was unexpected because it is counter to the findings from other research exploring the relationships between teachers' implicit beliefs and students' achievement in mathematics (Ilhan & Cetin, 2013).

Teachers' judgments against the mathematics standards

Analysis of the correlations between teachers' judgments and mean scale locations of students' achievement highlighted some noteworthy patterns. Two of the teachers in the sample made judgments that were not consistent with the independent measure of their students' achievement. For one teacher, the mean scale location for students judged to be "at the standard" was higher than the mean scale location for students "above the standard." The other teacher had mean scale locations from the independent measure virtually identical for judgments "at the standard" and "below the standard." This suggests that these teachers may have based their judgments on different factors then those being measured by the independent assessment. One possible explanation for these is that the assessment may have limited validity. Teachers are required to make judgments of students' solution methods when solving problems in mathematics, students were marked on correct answers rather than their solution methods. Teachers who base their judgments on a range of assessments including informal observations of students' solution methods when solving problems as part of the normal classroom program are likely to make judgments that are moderately correlated to the independent measure of students' achievement.

Another pattern that emerged from this analysis was the difference in the apparent weighting given to students' achievement over the course of the year. This illustrates one of the tensions when making judgments about students' achievement against the mathematics standards that is not such an issue for judgments made against the reading and writing standards. The content covered in mathematics in one year is quite broad. Regardless of how teachers structure their mathematics program over the year it is inevitable that some elements of mathematics will be covered early in the year and not re-visited. Teachers must

weigh this against the desire to use current assessment information when making national standards judgments. It is likely that teachers will either have to incorporate assessment information from across the year, or alternatively have a narrower focus on the elements of the mathematics curriculum that have been taught most recently. The only other alternative is to rely on more formal assessments to provide information covering the breadth of the curriculum.

Related to this, is one possible explanation for teachers giving relatively equal weight to students' achievement from across the year is that schools have generalised the model for delivery of the Numeracy Project Professional Development. This model divided the numeracy domains across the school terms, with addition and subtraction delivered early in the year, usually in term one, multiplication and division delivered in term two, and proportional thinking delivered in term three. This approach is unlikely to adequately reflect the needs of the students in terms of the changes in the relative difficulty and importance of the domains as the stages increase. This also artificially segments the domains rather than connecting ideas across domains. Schools who continue to divide the curriculum in this way require teachers to use assessment artifacts from across the year when making judgments of students' achievement. This is particularly important because students' achievement across the numeracy domains are the predominant factor that judgments are based on (Ministry of Education, 2009a).

Of concern was the evidence of teachers over reliance on standardised tests when making their overall teacher judgments against the mathematics standards. Data from one of the teachers strongly suggest that she may have based her judgments solely on e-asTTle assessments. The high level of correlation between her judgments and students' achievement on the independent measure, which was also constructed from items on e-asTTle, was indicative of using the same type of assessment tool. Data from another teacher suggested she might have based her summative judgments on standardised tests, possibly the Progressive Achievement Test commonly used by New Zealand teachers. The relatively high correlations between her judgments and students' scale locations suggest that the assessment tools used were similar. Relying

solely on these assessment tools will very likely reduce the validity of teachers' judgments. Teachers are required to use a range of assessments, "much of it derived from daily classroom interactions and observations." (Ministry of Education, 2009a, p. 12).

The analysis of variance identified several significant effects when the data were grouped in different ways. Students' increase in achievement over the period of the study was identified. The data were explored to elucidate any effect of grouping, because there were differences in the grouping practices for mathematics in the participating schools. Schools A and B cross-grouped for mathematics, with school B also separating the Year 5 students from the Year 6 students. The remaining schools used in-class grouping, including both Year 5 and year 6 students. The achievement of Year 6 students who were cross-grouped for mathematics was significantly greater than the other students. Surprisingly there were no significant differences between Year 5 and Year 6 students. The interaction between grouping and testing point identified the difference in progress between the cross-grouped and the in-class grouped students, with the in–class grouping students making significantly more progress than the cross-grouped students. The interaction between grouping and year level indicated that the main effect of grouping was dependent on the year level of the students. The Year 6 cross-grouped students' achievement on the first assessment was significantly higher than other students and interestingly the students who received in-class grouping made significantly more progress than the cross-grouped students. It is possible that a ceiling effect limited the progress of the Highly achieving students.

The small sample of schools involved in this research were heavily weighted to the higher decile range, with 87.5% of the schools in the sample in the decile range 8-10 when compared to 32% percentage of schools nationally. The schools were all either suburban or inner city schools located near central Wellington. The schools in the sample were evenly divided between contributing and full primary schools, this compares with 43% contributing and 57% full primary schools nationally. Further research with a more representative sample of

schools, especially lower decile and schools away from major metropolitan areas would be advisable to compare the findings from this research to those from teachers in different socio-economic and geographic settings. It would also be advisable to extend the research to different levels of the curriculum, especially Level 4, based on the evidence suggesting increased variability in teachers' judgments at this level (Ward & Thomas, 2015).

Conclusion

The findings from this research suggest that the beliefs teachers hold about effective pedagogy, students' ability and the purpose of assessment in mathematics are related to their application of assessment criteria when making summative judgments of students' achievement against the National Standards in mathematics. The variability of teachers' summative judgments was analysed in terms of the degree of conservatism of their application of assessment criteria when making "at the standard" judgments.

Teachers who were found to be more conservative in their application of assessment criteria when making "at the standard" judgments tended to be more moderate in their beliefs. As teachers tended towards a more conservative application of assessment criteria their beliefs tended towards a pedagogical approach that was inclusive of both connectionist and conceptual approaches as well as elements of both discrete and procedural approaches. Their view of students' ability tended towards a more entity view whilst also tending to agree with a more summative approach to assessment. Conversely as teachers tended towards a more liberal application of assessment criteria when making "at the standard" judgments in mathematics, their beliefs tended to endorse pedagogical approaches that were more strongly connectionist and conceptual and a view of students' ability that was more incremental.

One weak relationship was found between gains in students' achievement in mathematics and teachers' beliefs. Increased students' achievement in mathematics was weakly related to teachers' entity view of students' ability in mathematics. This relationship prompted further exploration as it was counter to previous research findings (Ilhan & Cetin, 2013). Upon examination of the responses to the four items measuring teachers' implicit theories of students' ability, rather than endorsing an entity view of students' mathematical ability, it is more accurate to describe this as teachers tending to express more neutral views of an incremental view of students' mathematical ability. It may be more accurate to describe this relationship as increased students' achievement in

mathematics being related to teachers' more neutral views of students' ability in mathematics.

Findings from this research highlighted the variability in teachers' summative judgments of students' achievement in mathematics. The variability of teachers "at the standard" judgments was described in terms of teachers' degree of conservatism in their application of assessment criteria. This finding was supported by the findings reported by Ward and Thomas (2013, 2015). Analysis of student achievement data in relation to teachers' summative judgments strongly suggest that some teachers based their judgments on assessment tools like e-Asstle and the Progressive Achievement Test, which is of concern. Analysis also suggests that there was substantial variability in the extent to which teachers base their judgments on recent assessment data

Schools need to be supported to increase the reliability of teachers' summative judgments. This support could be either through professional development, with experts working collaboratively with teachers when making summative judgments, or through tools like PACT that are designed to support teachers when making judgments against the National Standards.

Additional research is recommended to explore relationships between teachers' beliefs and the degree of conservatism in their application of assessment criteria when making judgments against the National Standards in mathematics at different levels of the curriculum, especially Level 4, and in different geographical locations around New Zealand. Additional research on teachers' judgments that includes analysis of teachers' teaching practice is also recommended.

References

Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 20-34. doi:10.1080/0969594x.2012.736364

Anthony, G., & Walshaw, M. (2007). *Effective pedagogy in mathematics/pāngarau best evidence synthesis iteration*. Wellington: Ministry of Education.

Askew, M., Rhodes, V., Brown, M., Wiliam, D., & Johnson, D. (1997). *Effective teachers of numeracy: Report of a study carried out for the Teacher Training Agency*. London: King's College, University of London.

Bahr, D., Monroe, E. E., Balzotti, M., & Eggett, D. (2009). Crossing the barriers between preservice and inservice mathematics teacher education: An evaluation of the Grant School Professional Development Program. *School Science & Mathematics, 109*(4), 223-237. doi:10.1111/j.1949-8594.2009.tb18260.x

Baturo, A. (2004). Teachering enhancing numeracy. *Australian Primary Mathematics Classroom, 9*(4), 54-56. Retrieved from http://search.proquest.com/

Beishuizen, M., & Anghileri, J. (1998). Which mental strategies in the early number curriculim? A comparison of British ideas and Dutch views. *British Educational Research Journal, 24*(5), 519-538. doi:130.195.86.35

Bolden, D. S., & Newton, L. D. (2008). Primary teachers' epistemological beliefs: Some perceived barriers to investigative teaching in primary mathematics. *Educational Studies, 34*(5), 419-432. doi:10.1080/030556908022875

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice, 20*(1), 69-90. doi:10.1080/0969594x.2012.703170

Brown, G. T. L., & Harris, L. R. (2010). *Teacher's enacted curriculum: Understanding teacher beliefs and practices of classroom assessment*. Paper presented at the International Association for Educational Assessment Bangkok, Thailand.

Brown, G. T. L., Harris, L. R., & Harnett, J. (2012). Teachers' beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education, 28*(7), 968-978. doi:10.1016/j.tate.2012.05.003

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education, 27*(1), 210-220. doi:10.1016/j.tate2010.08.003

Brown, M., Askew, M., Millett, A., & Rhodes, V. (2003). The key role of educational research in the development and evaluation of the national numeracy strategy. *British Educational Research Journal, 29*(5), 655-667. doi:10.1080/0141192032000133677

Buehl, M. M., & Fives, H. (2009). Exploring teachers' beliefs about teaching knowledge: where does it come from? Does it change? *The Journal of Experimental Education, 77*(4), 367-407. doi:10.30200/jexe.77.4.367-408

Connolly, S., Klenowski, V., & Wyatt-Smith, C. M. (2012). Moderation and consistency of teacher judgement: teachers' views. *British Educational Research Journal, 38*(4), 593-614. doi:10.1080/01411926.2011.569006

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401-434. doi:10.1080/13803610701728311

Crisp, V. (2013). Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20*(1), 127-144. doi:10.1080/0969594x.2012.741059

Delandshere, G., & Jones, J. H. (1999). Elementary teachers' beliefs about assessment in mathematics: A case of assessment paralysis. *Journal of Curriculum & Supervision, 14*(3), 216-240. Retrieved from http://eric.ed.gov/

DuCloux, K. (2009). *First-year secondary mathematics teachers' assessment conceptions and practices*. Paper presented at the Psychology of Mathematics & Education of North America. Atlanta, Georgia. Article retrieved from http://ebscohost.com/

Dweck, C. S., Chiu, C.-y., & Hong, Y.-y. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. *Psychological Inquiry, 6*(4), 267-285. doi:10.2307/1448940

Even, R. (2005). Using assessment to inform instructional decisions: How hard can it be? *Mathematics Education Research Journal, 17*(3), 45-61. doi:10.1007/bf03217421

Firestone, W. A., Winter, J., & Fitz, J. (2000). Different assessments, common practice? mathematics testing and teaching in the USA and England and Wales. *Assessment in Education: Principles, Policy & Practice, 7*(1), 13. doi:10.1080/713613322

Fives, H., & Buehl, M. M. (2008). What do teachers believe? Developing a framework for examining beliefs about teacher knowledge and ability. *Contemporary Educational Psychology, 33*, 134-176. doi:10.1016/j.cedpsych.2008.01.001

Garcia-Cepero, M. C., & McCoach, D. B. (2009). Educators' implicit theories of intelligence and beliefs about the identification of gifted students. *Universitas Psychologica, 8*(2), 295-310.

Gill, M. G., & Hoffman, B. (2009). Shared planning time: A novel context for studying teachers' discourse and beliefs about learning and instruction. *Teachers College Record, 111*(5), 1242-1273. Retrieved from http://ebscohost.com/

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi:10.3102/003465430298487

Ilhan, M., & Cetin, B. (2013). Mathematics oriented implicit theory of intelligence scale: validity and reliability study. *GESJ: Education Science and Psychology, 25*(3), 116-134.

Jones, B. D., Bryant, L. H., Snyder, J. D., & Malone, D. (2012). Preservice and inservice teachers' implicit theories of intelligence. *Teacher Education Quarterly, 39*(2), 87-101. Retrieved from http://proquest.com/

Jones, B. D., & Egley, R. J. (2007). Learning to take tests or learning for understanding? Teachers' beliefs about test-based accountability. *Educational Forum, 71*(3), 232-248. doi:10.1080/00131720709335008
Jonsson, A. C., Beach, D., Korp, H., & Erlandson, P. (2012). Teachers' implicit theories of intelligence: influences from different disciplines and scientific theories. *European Journal of Teacher Education, 35*(4), 387-400. doi:10.1080/02619768.2012.662636
Klenowski, V. (2013). Towards improving public understanding of judgement practice in standards-referenced assessment: an Australian perspective. *Oxford Review of Education, 39*(1), 36-51. doi:10.1080/03054985.2013.764759
Martinez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: evidence from the ECLS. *Educational Assessment, 14*, 78-102. doi:10.1080/10627190903039429
Ministry of Education. (2007). *The New Zealand curriculum for English-medium teaching and learning in Years 1-13.* Wellington: Learning Media.
Ministry of Education. (2009a). *The New Zealand Curriculum mathematics standards for years 1-8.* Wellington: Learning Media.
Ministry of Education. (2009b). *The New Zealand curriculum reading and writing standards for years 1-8*. Wellington: Learning Media.
Ministry of Education. (2015, July, 27). NZ maths. Retrieved from http://www.nzmaths.co.nz/nzc-and-standards
Ministry of Education. (2011). *Ministry of Education position paper: assessment (schooling sector).* Wellington: Learning Media.
Ministry of Education. (2013). *Reporting student achievement: Guidance for reporting on National Standards.* Wellington: Ministry of Education.
Ministry of Education. (2015a, February, 9). Directory of Schools. Retrieved from http://www.educationcounts.govt.nz/data-services/directories/list-of-nz-schools
Ministry of Education. (2015b, February, 9). Education Counts. Retrieved from http://www.educationcounts.govt.nz/statistics/schooling/national-standards
Nisbet, S., & Warren, E. (2000). Primary school teachers' beliefs relating to mathematics, teaching and assessing mathematics and factors that influence these beliefs. *Mathematics Teacher Education and Development, 2*, 34-47. Retrieved from http://proquest.com/
Numeracy Professional Development Projects. (2008). *Book 1: The Number Framework, revised edition*. Wellington: Ministry of Education,.
Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332. doi:10.3102/00346543062003307
Peček, M., Valenčič Zuljan, M., Čuk, I., & Lesar, I. (2008). Should assessment reflect only pupils' knowledge? *Educational Studies, 34*(2), 73-82. doi:10.1080/03055690701811073
Peterson, P. L., Fennema, E., Carpenter, T. P., & Loef, M. (1989). Teacher's pedagogical content beliefs in mathematics. *Cognition & Instruction, 6*(1), 1. doi:10.1207/s1532690xci0601_1

Rittle-Johnson, B., & Koedinger, K. (2009). Iterating between lessons on concepts and procedures can improve mathematics knowledge. *British Journal of Educational Psychology, 79*, 483-500. doi:10.1348/000709908X398106

Rubie-Davies, C. M., Flint, A., & McDonald, L. G. (2011). Teacher beliefs, teacher characteristics, and school contextual factors: What are the relationships? *British Journal of Educational Psychology*. doi:10.1111/j.2044-8279.2011 .02025.x

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175-194. doi:10.1080/0260293042000264262

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education, 34*(2), 159-179. doi:10.1080/02602930801956059

Smaill, E. (2013). Moderating New Zealand's National Standards: teacher learning and assessment outcomes. *Assessment in Education: Principles, Policy & Practice, 20*(3), 250-265. doi:10.1080/0969594X.2012.696241

Song, Y., & Looi, C.-K. (2012). Linking teacher beliefs, practices and student inquiry-based learning in a CSCL environment: A tale of two teachers. *International Journal of Computer-Supported Collaborative Learning, 7*(1), 129-159. Retrieved from http://proquest.com/

Stipek, D. J., Givven, K. B., Salmon, J. M., & MacGyvers, V. L. (2001). Teachers' beliefs and practices related to mathematics instruction. *Teaching and Teacher Education, 17*, 213-326. doi:10.1016/s0742-051x(00)00052-4

Suurtamm, C., & Koch, M. (2014). Navigating dilemmas in transforming assessment practices: experiences of mathematics teachers in Ontario, Canada. *Educational Assessment, Evaluation and Accountability, 26*, 263-287. doi:10.1007/s11092-014-9195-0

Suurtamm, C., Koch, M., & Arden, A. (2010). Teachers' assessment practices in mathematics: classrooms in the context of reform. *Assessment in Education: Principles, Policy & Practice, 17*(4), 399-417. doi:10.1080/0969594x.2010.497469

Thomas, G., Tagg, A., & Ward, J. (2005). Numeracy assessment: How reliable are teachers' judgments? *Findings from the New Zealand Numeracy Development Projects 2005.* Wellington: Ministry of Education.

Ward, J., & Thomas, G. (2013). *National standards: School sample monitoring and evaluation project, 2010-2012*. Wellington: Ministry of Education.

Ward, J., & Thomas, G. (2015). *National standards: School sample monitoring and evaluation project, 2010-2013*. Wellington: Ministry of Education.

Wood, T., & Sellers, P. (1997). Deepening the analysis: longitudinal assessment of a problem-centred Mathematics program. *Journal for Research in Mathematics Education, 28*(2), 163-186. doi:130.195.86.35

Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 35-52. doi:10.1080/0969594x.2012.725030

Appendix 1 – Ethical Approval

Appendix 2 – Students' assessments

Appendix 3 – Questionnaire

Appendix 4 – Teachers' information letters