The development and validation of a student evaluation instrument to identify highly accomplished mathematics teachers.

Stephen Earl Irving

A thesis submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy
The University of Auckland, 2004

Consent Form

The University of Auckland

Thesis Consent Form

This thesis may be consulted for the purpose of research or private study provided that due acknowledgement is made where appropriate and that the author's permission is obtained before any material from the thesis is published.

I agree that the University of Auckland Library may make a copy of this thesis for supply to the collection of another prescribed library on request from that Library; and I agree that this thesis may be photocopied for supply to any person in accordance with the provisions of Section 56 of the Copyright Act 1994

Signed:… .

Date:… … … …

Abstract

This study describes the attributes of a highly accomplished mathematics teacher as reported by the students in their class, and also determines whether high school students can differentiate between highly accomplished mathematics teachers and others.

The 51-item instrument, Students Evaluating Accomplished Teaching – Mathematics, was developed to map the construct of highly accomplished teaching as articulated by the National Board for Professional Teaching Standards in their Adolescent and Young Adulthood Mathematics Standards. Two focus groups of New Zealand high school mathematics teachers reviewed these Standards, and found that there were more similarities than differences between the Standards and what they would expect of a highly accomplished teacher in New Zealand. Questionnaire items were drafted relating to each of 470 statements in the Standards. These items were trialled in New Zealand high schools, and analysed using factor analysis and item response theory, to select items that completely mapped the Standards. The questionnaire was then administered to 1611 students in the classes of thirty-two National Board Certified Teachers and twenty-six non-Board colleagues in 13 states of the USA.

Multivariate analysis of variance and discriminant function analysis were used to establish that students can record and report the difference between NBCTs and their non-Board certified colleagues, and describe what students believe are the attributes of a good teacher. Highly accomplished teachers build a relationship between their students and the mathematics curriculum, as well as with the language and processes of mathematics, by engaging their minds with challenging material and rich tasks. These results provide further validation of the NBPTS certification process, and indicate that students provide dependable evaluations of their teachers. The student evaluation questionnaire could be used with confidence in both the USA and New Zealand to identify highly accomplished mathematics teachers.

Acknowledgements

This research was only possible through the involvement of a large cast of students, teachers and principals, who willingly gave their time and energy to make it possible. There were no direct, tangible benefits for them yet they were prepared to contribute their thoughts and concerns to further this project. To them I express my sincere thanks for their assistance, often at what seemed like the most difficult time in the annual cycle of school life.

It is the student voice that is at the heart of this research. Over 2500 students (1640 in the USA and 899 in New Zealand) participated in this research, often 'volunteered' by their teacher to act as unwitting reporters of their teacher's strengths and foibles. In this research, they have clearly indicated what they like about their teachers, and what they do not. The good news is that their best ratings help to dispel the myth that high school students can be "bought" by a charismatic teacher who ignores the real job of the teacher – to engage students with the curriculum, and challenge them to strive for the best. Thank you for helping to make this clear. High school students are not very familiar with this role in teacher evaluation, but they seemed to approach their role with due consideration. As always, working with these young people has strengthened my optimism about the future that will eventually be in their hands.

The 116 teachers have been the key people in this research, but their role has been greater than just providing access to their classes. Many teachers feel threatened at the thought of being judged by their students, and then to find out what their class thinks about them, but these teachers welcomed this opportunity and opened the doors that made this possible – the classroom doors, as well as the doors to their minds and hearts. In addition, they provided helpful input into the project. Without this receptiveness as well as the cooperation of their principals, this research would never have been possible.

Professor John Hattie has been an incredible tower of inspiration and strength. He has introduced me to the wonders of Item Response Theory and the National Board for Professional Teaching Standards, and shepherded me through the many ups and downs of postgraduate research. The lengthy gestation has been very trying, and his unfailing patience has been greatly appreciated. All the while, John has provided invaluable advice to keep this research pointing in the right direction. Dr Richard Hamilton has also provided wise counsel as this journey drew to a close. Dr Gavin Brown has helped to keep me on task over the past year while I have been working on Project asTTle. The teams in the Research Centre for Interventions in Teaching and Learning (RCITL), and Project asTTle at the University of Auckland have been very supportive and encouraging, providing a sounding board by listening attentively and responding appropriately.

The final word of thanks goes to my long-suffering family. They have had to live with the longer than expected time it has taken to bring this to fruition, and the time that has not been shared with them. In spite of this, they have continued to share their love, support and encouragement to see me through to completion. No doubt they will be as pleased as I now that this project can finally be "put to bed".

This has been an awesome journey, with many twists and turns - I am hugely indebted to all who have made this journey so fulfilling that it has come alive in my mind.

Table of Contents

Consent Form i

Abstract ii

Acknowledgements iii

List of Tables ix

List of Figures xi

Chapter One: Introduction 1

Chapter Two: Literature Review 6

Section One. Highly accomplished teaching and the National Board for Professional Teaching Standards... 6

The notion of good teaching 6

Professionalism 8

Professional Teaching Standards and The National Board for Professional Teaching Standards (NBPTS) 12

Do Board Certified Teachers make a difference? 25

The NBPTS and its critics 27

Summary 36

Section Two. Student Evaluations of Teaching Performance 37

Teacher evaluation 37

Teacher evaluation models 38

Student conceptions of teaching 43

The SET literature 46

High school studies of SETs 48

Arguments for and against SETs 54

Validity 57

Possible contaminants 65

Issues related to course variables 66

Issues related to student variables (student presage) 68

Issues related to teacher variables (teacher presage) 70

Issues related to instrument variables 78
Issues related to administration and purpose variables 79
Dimensionality 81
Concluding comments on SETs 83

Chapter Three: Study One 84
Membership selection 86
Conduct of the focus groups 87
Data analysis 89
Research questions 90
Results and discussion 92
Acceptability of the Standards 92
Modifications to the Standards 106
Concluding Statement 109

Chapter Four: Study Two 111
Instrument development 111
Trial One 119
Setting 119
Analysis 121
Classical Test Theory and Factor Analysis 122
Item Response Theory item selection and test construction 124
Item Information 130
Results 132
Form A 132
Descriptive statistics 132
Factor analysis 133
Item Response Theory 133
Item selection from Form A 146
Form B 147
Descriptive statistics 147
Factor analysis 147
Item Response Theory 148
Item selection from Form B 158
Form C 159

Descriptive statistics 159
Factor analysis 165
Item Response Theory 166
Item selection from Form C 166
Questionnaire assembly for the November questionnaire 174
Trial Two 176
Form November 176
Descriptive statistics 177
Factor analysis 178
Item Response Theory 188
Form Technology 188
Refinement of Technology Items 188
Descriptive statistics 191
Factor analysis 192
Item Response Theory 196
Item selection from Study Two for SEAT-M 196
Wording and Content analysis of assembled items 199
Discussion 206
Chapter Five: Study Three 209
Procedure 209
Subjects 212
Student participants. 213
Instrument/Materials 214
Results 214
Data processing and analysis 214
Multivariate analysis of variance (MANOVA) 223
Discriminant Function Analysis 224
Discussion 230
Chapter Six: Conclusion and Discussion 235
Implications 241
Future research 243
References 245
Appendices 269

Student Questionnaire Form A 270
Student Questionnaire Form B 274
Student Questionnaire Form C 278
Student Questionnaire Form November..... 282
Student Questionnaire Form Technology 286
Students Evaluating Accomplished Teaching - Mathematics..... 287

List of Tables

Table 1 Overall rating and mean coefficient correlations for teachers and courses 65
Table 2 Applicability of Standards classified by number and percent of participants and analysed units 92
Table 3 Sample of paragraph from Standards and drafted statements 112
Table 4 Example of the synthesis of Standards statements to items for trial 114
Table 5 School Descriptives for Trial One 120
Table 6 Participant Descriptives for Trial One 121
Table 7 Descriptive Statistics for Form A 134
Table 8 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form A 140
Table 9 Classical Test Theory and Item Response Theory Item Statistics for 64 Form A Items 144
Table 10 Factor loadings and IRT parameters for 24 items selected from Form A 146
Table 11 Descriptive Statistics for Form B 149
Table 12 Summary of Items and Factor Loadings for Oblimin Three-Factor solution for Form B 153
Table 13 Classical Test Theory and Item Response Theory Item Statistics for 63 Form B Items 156
Table 14 Factor loadings and IRT parameters for 20 items selected from Form B 158
Table 15 Descriptive Statistics for Form C 160
Table 16 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form C 167
Table 17 Classical Test Theory and Item Response Theory Item Statistics for 67 Form C Items 171
Table 18 Factor loadings and IRT parameters for 21 items selected from Form C 173
Table 19 Wording amendments made to items selected for November Form 175
Table 20 School Descriptives for Trial Two 176
Table 21 Participant Descriptives for Trial Two 177
Table 22 Descriptive Statistics for Form November 179
Table 23 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form November 184
Table 24 Classical Test Theory and Item Response Theory Item Statistics for 66 Form November Items 189
Table 25 School Descriptives for Technology Trial 191
Table 26 Participant Descriptives for Technology Trial 191
Table 27 Descriptive Statistics for Form Technology 193
Table 28 Summary of Items and Factor Loadings for Oblimin Three-Factor solution for Form Technology 194
Table 29 Classical Test Theory and Item Response Theory Item Statistics for 14 Form Technology Items 195
Table 30 Factor loadings and IRT parameters for 50 items selected from November Form 197

Table 31 Final set of items for SEAT-M, their development history, and origins in the Standards.....200
Table 32 Descriptive statistics for National Board Certified Teachers (NBCT) and non-National Board Certified Teachers (non-NBCT) 212
Table 33 Item Statistics for National Board Certified Teachers (NBCT) and non-National Board Certified Teachers (non-NBCT).....215
Table 34 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form SEAT-M 220
Table 35 One-way analyses of variance for effects of National Board Certification status on five SEAT-M factors 224
Table 36 Classification Analysis for NBCT status (51 items, 852 full-data cases).....225
Table 37 Structure Matrix of pooled within-groups correlations between 51 SEAT-M items and the standardised canonical discriminant function.....226
Table 38 Eigenvalues, Canonical correlation, Wilks' Lambda and chi-square for discriminant analysis of SEAT-M five factors 229
Table 39 Box's M statistics for five SEAT-M factors.....229
Table 40 Classification summary for NBCT status (5 factors) 230
Table 41 Structure Matrix of pooled within-groups correlations between 5 SEAT-M factors and the standardised canonical discriminant function. 230

List of Figures

Figure 1 Item Characteristic Curves for an item with ideal characteristics, Item N37 129
Figure 2 Item Characteristic Curves for an item with poor characteristics, Item A60 130
Figure 3 Item Information Curve for low information item, A36 131
Figure 4 Item Information Curve for high information item, A34 131
Figure 5 Item characteristic curves for an item with random responses, Item N07 206

Chapter One: Introduction

Jaime Escalante's Advanced Placement (AP) calculus class stunned the assessment world in 1982. How could all 14 of his students score so well when no one from their East Los Angeles school had ever passed the examination? The examining board was very suspicious, but when twelve of them repeated their results in a second administration of the examination, the world knew that there was something special happening in the school – a special teacher who over a period of a decade had developed a mathematics programme that bred a culture of success. Escalante became the focus of attention as this programme burgeoned under his guidance – indeed a Hollywood movie (entitled "Stand and Deliver") was made to tell the story of his inspirational teaching. He left the school in 1991, and the programme quickly withered without his inspirational teaching.

When asked to account for his success, Escalante responded that four elements were necessary – ganas de triunfar (desire to succeed) on the part of the students; knowledge of the subject; knowledge of how to teach the subject; and respect for the students as people. While Escalante attached the first of these characteristics to the students, it is fair to say that he was underplaying his own role. The students did not arrive at school prepared to learn, particularly when the culture of the school and their peers did not encourage them to engage with learning. He motivated the students to reach for the stars, and his knowledge of the subject and how to teach it, plus the high expectations he had of his students drove them to success beyond most people's expectations.

There is a saying that the three great quests for humanity are the Holy Grail, the fountain of youth, and the answer to what makes a good teacher. Researchers have long sought the answer to the latter, and the story of Jaime Escalante helped to intensify the search for the key to what makes these teachers so successful. Great teachers have their place in the records of antiquity, and have been well studied. Often it is the writing of

their students that has sustained the teacher's place in history, yet today the voice of the students concerning their teachers is often accorded second-class status. This research seeks to give voice to the students by seeking their assessment of their teachers, and seeing the mathematics classroom through their eyes. Can students tell the difference between highly accomplished teachers and their colleagues, and what can the students tell us about these great teachers?

This research was designed to investigate two issues – the use of student evaluations to report on highly accomplished teachers, and, what marks out these accomplished teachers, at least in the eyes of the students. Any discussion of highly accomplished teaching begs the question "what is a good teacher", and there can be as many answers to this as there are commentators. Models of good teaching abound, and the first task was to seek the "gold standard" that could be used to construct a Student Evaluation of Teaching (SET) instrument. Only one of these many models has been used to classify and certificate teachers whose practice is of the highest calibre – the National Board for Professional Teaching Standards (NBPTS) model in the USA.

Two bodies of literature inform this research. The first section reviews the National Board, its critics and the model of highly accomplished teaching that the Board has articulated. This model is premised on three simple elements – the skills, the knowledge and the dispositions that highly accomplished teachers display in their daily work. This section also reviews notions of professionalism, professionalism in teaching, and the manner in which teachers have sought to enhance their work and standing as a profession. The second section considers teacher evaluation, and reports on the voluminous research on student ratings or evaluations of teaching performance. Student evaluations of teacher performance (SETs) have been endlessly scrutinised, praised, criticised and disparaged. In spite of that, the message is clear about SETs – they provide a reliable and valid measure of a teacher's performance in the classroom, for those aspects of teaching that students can appropriately report. Where appropriate, the review refers to the limited amount of research that has been conducted in secondary schools. It is the contention of this research that SETs can be used to identify those characteristics of good teaching that serve as the mark of the very best mathematics teachers. The findings of this review are found in Chapter Two.

The research was conducted as three linked studies. In the first study, the NBPTS model of accomplished mathematics teaching covering the senior high school, the Adolescent and Young Adult (AYA) Mathematics Standards, was examined by two focus groups of highly reputed New Zealand mathematics teachers to determine the applicability of the model in New Zealand. Provided the model received the imprimatur of acceptance from these focus groups, the Standards could be used as a foundation for a SET instrument that could be developed in New Zealand. Modifications to the Standards were expected, but these were minor and the integrity of the Standards remained intact. Qualitative methods were used to analyse the focus group transcripts. Chapter Three reports on the process and findings of the focus groups regarding the Standards.

The second study was focused on the development of a student evaluation instrument that had good psychometric properties, and fully and fairly mapped the construct of interest – highly accomplished mathematics teaching as defined by the AYA Mathematics Standards. The Standards were unpacked into 470 statements, which were then crafted into 191 questionnaire items for trial in New Zealand secondary schools. Data were analysed using Classical Test Theory (CTT) and Item Response Theory (IRT) methodologies. The Students Evaluating Accomplished Teaching – Mathematics (SEAT-M) was derived from these analyses, and has 51 items, one of which is an overall teacher effectiveness item. A five-factor structure was found underlying the instrument. The findings of this study are reported in Chapter Four.

A total of 1611 USA high school students in the classes of thirty-two AYA Mathematics National Board Certified Teachers (NBCTs) and twenty-six of their colleagues completed the SEAT-M questionnaire. Study Three is recorded in Chapter Five, which provides an account of the administration of the instrument to these students, the results and discussion of the results. The discussion addresses the research questions that follow in this chapter. These questions have two main themes – whether students can reliably distinguish between highly accomplished mathematics teachers and their colleagues, and, what characteristics distinguish the accomplished mathematics teacher from their colleagues.

The final chapter, Chapter Six, restates the problem, synthesises the findings and observations from the three studies, and relates these back to the bodies of literature that inform them. Implications of the findings for teaching and teacher education are discussed, along with suggestions for future research.

A number of assumptions underpin the use of student evaluations for this research.

> Student evaluations are a valid, reliable, stable, useful, and cost-effective means of gathering data about what happens in the classroom.
> Students have experienced a wide range of teachers and teaching and have built up an understanding of what they regard as effective teaching.
> Students can communicate this understanding when asked in an appropriate manner.
> Students are willing to provide their viewpoints about teaching and learning in their classrooms if asked.

Based on these assumptions, the two main research questions for this research are:

> Can students reliably distinguish between highly accomplished mathematics teachers and their colleagues?
> What characteristics distinguish the accomplished mathematics teacher from a colleague?

A subsidiary question was also addressed.

> Is the description of the highly accomplished mathematics teacher in the USA also an accurate description of highly accomplished mathematics teaching in New Zealand?

Finally, an answer is provided to the question, "Would Jaime Escalante qualify as a NBCT?"

This thesis investigates one means of identifying high performing teachers, the characteristics that make them stand out as people whose practice is at the top of their

profession, and discusses ways in which this method could be used to complement the National Board's existing palette of assessments. In doing so, it also addresses a major validity issue regarding National Board certification, and whether their assessments identify teachers that are also considered exemplary by their students.

Chapter Two: Literature Review

Two bodies of literature shape the framework that underpins this research – the notion of accomplished teaching articulated by the National Board for Professional Standards (NBPTS) and their development of a model of accomplished teaching, and the use of student evaluations of teaching performance (SETs) as a source of data within the context of teacher evaluation. These two bodies will be addressed in the two sections of this chapter- firstly, professional teaching standards and the National Board, and in the second section, SETs.

Section One. Highly accomplished teaching and the National Board for Professional Teaching Standards.

The National Board for Professional Teaching Standards has articulated a model of accomplished teaching, and has used a set of standards developed on this model to assess and certificate teachers who meet the standard. This model and the standards are embedded in a lengthy history of studies on good teaching, as well as the desire to professionalise teaching. This section of the review will examine good teaching, expertise in teaching, and professionalism and their implications for articulating good teaching. The history of the National Board will be outlined, along with the development of the Standards and assessment tools, together with a review of the arguments of several National Board critics.

The notion of good teaching

In spite of decades of research, there is no universal definition, much less a consensus, about what constitutes good teaching, or even excellence in teaching. The search for the good teacher has occupied the minds of researchers for over a century, and the many proffered answers are useful in several ways as they:

help us understand what good teaching looks like;

provide goals, models and incentives for teachers in their quest to improve teaching;

provide evidence for teachers to explain what they do and why, by describing the tasks and behaviours that can be regarded as important and what we can learn form them;

assist decision makers as they identify and remove teachers who are unable to "make the grade";

address a problem inherent in teacher training that Berliner (1986) highlights, in that the expert teacher who is assisting a trainee teacher behaves automatically in their thinking and behaviour, and has difficulty articulating the basis for their expertise;

inform the debate about good teaching and how to improve it;

influence governmental decisions regarding master teachers, and even remuneration based on performance; and for once and for all

dispel the myth that there is only one kind of good teacher.

Cruickshank (2000) lamented the fact that teachers were constantly pilloried for the perceived failures of education and traced the history in the twentieth century of the search for the "good teacher". In the early to mid twentieth century the focus was on *ideal* teachers, in which the traits and attributes of excellence were defined by "selected significant others" (p. 2). In the early 1960s, the *analytical* teacher attempted to analyse what they were doing, and used the result to modify their teaching - the Flanders classroom interaction studies provide a good example of this model (Flanders, 1973; , 1974). The publication of the Coleman Report in the USA (Coleman et al., 1966) lead to the search for the *effective* teacher. Coleman's report concluded that the student's family background was the main reason for student success in school. The report's findings proposed that children from poor families and homes, lacking the conditions or values to support education, could not learn, regardless of what the school or teachers did. Researchers set out to show otherwise by studying teachers whose students performed better than others on standardised assessments. Stronge (2002, p. 62) characterised an effective teacher as one who recognises complexity, communicates clearly and serves conscientiously. A fourth variation is akin to Scriven's duty-based model of teacher evaluation – the *dutiful* teacher has a set of duties to fulfil including

knowledge of these duties, knowledge of subject matter, knowledge of the school and community; knowledge of students; knowledge of classroom skills for managing the learning environment, and a service orientation to the teaching profession. The accountability movement of the 1970s searched for the *competent* teacher. This paradigm drew on the earlier effective teacher research, as well as task analysis and studies of highly competent practitioners. Teacher testing became a way of assessing competence. The search for the *expert* teacher paralleled studies done in artificial intelligence, software design and other fields (chess, typewriting, medical diagnosis for example), with studies conducted to examine what expertise these teachers had, and how they acted differently from novice and experienced (but not expert) teachers, while the *reflective* teacher provides another, more recent, variation. Reflective teachers critically examine their work as a teacher as both an art and as a science, and seek to learn more about themselves as teachers in a continuous desire to improve.

Cruickshank recognises at least three other variations. *Satisfying* teachers like to please others, and have others visit their classroom. *Diversity responsive* teachers take a special interest in and are sensitive to students who differ in one or more ways, whether culturally, socially, economically, intellectually, physically or emotionally. Finally, *respected* teachers possess and demonstrate a range of virtues like care, honesty, tolerance and fairness. These are the teachers to be found in movies like Goodbye Mr Chips, To Sir With Love and Mr Holland's Opus. However, Cruickshank allows that none of these typographies satisfies all stakeholders. Indeed, an eleventh variation could also be added to the list – the *National Board Certified Teacher* (NBCT), who is deemed to be a good teacher if they can demonstrate that they meet the standards for accomplished teaching described by "discerning colleagues" (King, 1994). While this variation draws heavily on the reflective teacher variant, it has many elements from the other models as well.

Professionalism

The rhetoric of educational improvement has adopted the language of teaching as a "profession". In the USA, Time for Results (National Governors' Association, 1986) and in the UK, Better Schools (Department of Education and Science Welsh Office, 1985) spelled out ways to achieve professional status for teaching, as a means of

improving outcomes for students. However, professionalism (and its related parts of speech), like good teaching, remains a highly contested term (Hargreaves, 1997; Helsby, 1995; Hoyle & John, 1995; Locke, 2001) particularly because of the labour market and political fields on which the contest occurs, or as one commentator put it when discussing the classification of a profession, "less logical and more ideological" (Hoyle, 1974, p. 13). The teacher associations/unions have seen professionalism and professionalisation as the opportunity to restore autonomy and return control of teaching and the curriculum to the teachers (re-professionalisation), while opponents have railed against provider capture, and expressed their desire to keep the general public in control of education. Through 'discourses of derision' (Hoyle, 1974, p. 167), that blame and shame teachers for all the perceived ills of schools and society, these opponents have sought to de-skill and de-professionalise teaching by changing conditions of union membership (The Employment Contracts Act 1991, for example), restricting teacher involvement in decision making, prescribing central curricula (The New Zealand Curriculum, 1993), providing alternative certification routes that enable more unlicensed and uncertified adults to enter teaching (for example, Teaching for America), increased accountability through appraisal (Ministry of Education, 1999b,, 1999c,, 2000a) and inspection (The Education Review Office), and targeting savings in salaries and conditions of service, such as the provision in the collective employment agreement that schools can require teachers to undertake up to five days of professional development outside of school time (The Secretary for Education & The New Zealand Post Primary Teachers' Association, 2002). These strictures have lead one commentator (Helsby, 1995, p. 318) to describe this process as "proletarianisation", although Murphy (1990) would argue that this is simply a rationalisation and bureaucratisation of the professional's role. Secondary teachers have always had the overarching school qualifications system limiting the extent of their academic freedom and autonomy. However, Hargreaves argued that under conditions where teaching is de-professionalised, "practice can at best only be reproduced, not improved" (1997, p.168).

According to Hargreaves (2000), teacher professionalism has passed through at least four ages in its development – pre-professional; the autonomous professional; the collegial professional, and the post-modern or post-professional. In the pre-professional stage, traditional mass public education occurred, with basic teacher centred transmission of knowledge. The class was treated as if it were a single collective

student. Teachers engaged in an apprenticeship, and followed the teaching patterns of their knowledgeable superiors. The model of a good professional was a teacher who knew their stuff, how to get it across and could keep order. The age of the autonomous professional was characterised by the ideological battles over child-centred and subject-centred education, open classrooms and closed classrooms, traditional and progressive methods. However, teaching was still largely an isolated activity, conducted in egg-crate classrooms, and hidden from the gaze of even your colleagues. The third age, the collegial professional, coincided with the explosion of knowledge (of what teachers were expected to teach, and of pedagogical knowledge), the development of information networks, and changes in society that were reflected in schools (that is, a more inclusive and diverse society and school population, and an often alienated adolescent student body). Teachers increasingly turned to each other, developing collaborative learning communities, to provide support, purpose and identity. The current age, post-professional or post-modern professional, is marked by global economic forces, the digital revolution and uncertainty. Rationalisation and cutbacks, competition for students, performance management and merit pay are all indicative of this. In this climate, the certification of advanced skills teachers (often accompanied by increased salary) has the potential to be extremely divisive. To counter this principle of 'divide and rule', Hargreaves argues, teachers need to direct their collaborative energies into improving teaching, learning and caring in school, setting and meeting exacting standards of professional practice, and forging new relationships with their greatest allies - the parents - to protect and advance their professionalism.

Because of the putative advantages associated with professionalism, teachers, nurses and social workers are among a number of occupational groups that have been engaged in long running campaigns to claim professional status. These groups have been described as 'semi-professions' (Etzioni, 1969), as they fail to meet the standard requirements for a profession. Primarily the nature of the difference lies in professional authority, as teachers work in administrative/bureaucratic units (schools/districts) that make decisions that the teacher then carries out. Their workdays are tightly regulated, supervisors are permitted to visit without prior notice, and the supervisors themselves are members of the semi-profession. Salaries are uniformly paid on the basis of years of experience, and no consideration is made for competence, grade level or subject specialisation – merit pay is strongly resisted. In addition, the length of training is

shorter than that required for entry to the professions. Lortie (1969) asserted that teaching lacked a "refined technical culture" (p. 29), with little in the way of a trade jargon that has a universally common meaning. Etzioni also acknowledged that one other significant difference between the professions and the semi-professions is gender – the typical profession is male, while the typical semi-professional is female. Goode (1969, p. 274) argued that history indicated that there are a series of steps that transform a semi-profession into a profession, but admitted that these steps were neither empirical nor convincing – full time activity at the task; establishment of university training; national professional association; redefinition of the core task to give the "dirty work" to subordinates; conflict between the old timers and the new people who seek to upgrade the job; competition between the new occupation and neighbouring ones; political agitation to gain legal protection; and, a code of ethics. In spite of their attempts, Goode predicts that teachers will not achieve professional status, even if they increase their relative income and prestige, not because they work in bureaucracies but because they do not *control* the essential work of the bureaucracy (Goode, 1969, p. 294), even though schools are "loosely coupled" organisations (Weick, 1976), and teachers exercise considerable autonomy once the classroom door shuts. Hoyle (1974) argued that teachers should not use the term "professional" as their day-to-day work is atheoretical.

Although the terms "profession" and "professional" have been modified through their use in sport to distinguish professional from amateur, there has been general agreement (Darling-Hammond & Wise, 1992; Etzioni, 1969; Goode, 1969; Hargreaves, 1997; Hargreaves & Goodson, 1996; Hoyle & John, 1995) on three distinguishing features of a profession:

> a specialised body of knowledge (expertise)
> a strong service ethic (altruism), and
> self-regulatory control of the profession (autonomy) - both at entry to the profession, and for continued membership.

The first two elements imply that professionals focus on bringing their knowledge and expertise to bear in the best interests of the client (or in the case of teachers, the student). This forms a social contract with society, which in turn, delegates to the profession the right to self-regulate. Autonomy in this sense does not include the right for individuals to behave in idiosyncratic ways (that is, the freedom to do whatever the

teacher wishes), but rather refers to the process of responsible self-governance. The profession provides a warranty to society that its members are competent, and will discipline members who fail to meet standards of professional behaviour. This includes removing the right to practise. Essentially, this "suggests greater regulation of *teachers* … in exchange for the deregulation of *teaching*" (Darling-Hammond & Wise, 1992, p. 1359. Original emphasis).

In an English study of teachers and their views on professionalism (Helsby, 1995), the participants felt that teaching certainly met the requirements of a profession in the first two respects, but that they did not have the same autonomy as practitioners in the traditional professions (law, medicine and the church), especially following the introduction of the National Curriculum. The terms professionalism and professionalisation can be neatly distinguished by asking a teacher what it means to be professional – they usually respond in one of two ways (Helsby, 1995). On the one hand they will talk about the quality of the things that they do, and the behaviours and dispositions that guide them. Improving quality of practice and the service provided, or behaving professionally, is what is termed *professionalism.* On the other hand, they will describe how teachers are seen by others – their status, standing and levels of reward – that is, *being a professional* or what is termed *professionalisation*. However, one of the participants in Helsby's study was able to clearly articulate a key distinction that underpins the thinking of the National Board for Professional Teaching Standards:

> Professionalism is how a professional person carries out their particular business, but you can have professionalism without being actually a member of a profession I believe. (Helsby, 1995, p.321)

Professional Teaching Standards and The National Board for Professional Teaching Standards (NBPTS)

Rosenholtz (1984) argued that standards of excellence in teaching are important as they provide a target to aim for, even if the target is imperfect, and that current knowledge lays a foundation for constructing these standards. Berliner (1986) suggested that they [standards of excellence in teaching] show off the best in teaching and provide the gold standard. In spite of the seductive nature of these statements, professional teaching standards have had a somewhat tortured history in being adopted in a number of

countries as a means to "improve learning outcomes for students by improving the quality of teaching and leadership".

In New Zealand, performance management systems have been mandatory in all schools since 1997, and following a trial period with interim standards for classroom teachers, teachers with special responsibilities, and school managers (Ministry of Education, 1995,, 1997a,, 1997b,, 1998a,, 1998b,, 1998c), a revised set of professional standards were released in 1999 for secondary and area schools to use (Ministry of Education, 1999a,, 1999b). From 2000, schools were obliged to incorporate the standards into their performance management systems. These standards describe the "critical knowledge, skills and attitudes needed to perform a particular role effectively" (Ministry of Education, 1999b, p. 4) and were derived from a proposal by the Teacher Registration Board (TRB) that outlined "satisfactory teacher dimensions" for registration and renewal. The TRB described them as the "minimum level of acceptability" required (Teacher Registration Board, 1997, p. 2). While this may be an acceptable level of classroom practice for all teachers, "*minimal acceptability* can not be equated with *quality teaching*" (Upsall, 2000, p. 174. Original emphasis) or standards for an accomplished profession.They have been modified and adapted to reflect a set of "duty of care" competencies required of teachers for registration as fit to teach and to continue teaching. This concept of "duty of care" has parallels in Scriven's duties based approach to teacher evaluation in which he described the "obligation of the employee to discharge the duties of the job to the extent that is reasonably possible with the resources available" (Scriven, 1988b, p.126). The New Zealand Standards are a first step towards the professionalisation of teaching, allied with demands for greater accountability of teachers, but do not go far enough in defining standards for highly accomplished teaching.

In Australia, a Senate report *A class act: Inquiry into the status of the teaching profession* (Australian Senate Employment Education and Training Reference Committee, 1998) argued that there was a need to recognise advanced teaching knowledge and skills, and that good teachers should be rewarded in terms of pay and career advancement. This paralleled developments in Ontario, Canada (Ontario College of Teachers, 1999) and among eleven members of the OECD (Centre for Educational Research and Innovation, 1994), as well as following an earlier moderately successful

attempt in the United States from the mid 1940s to 1960s via the Teacher Education and Professional Standards (TEPS) Commissions.

In the United States, matters came to prominence with the publication of The President's Commission on Excellence in Education report *A Nation At Risk: The Imperative for Educational Reform* (1983). The report identified a serious crisis of public confidence in the teaching profession as they sought answers to why American children were not performing as well on international indicators as expected. The rhetoric in *A Nation at Risk* alarmed the country when, on its first page, it stated that "If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war" (p. 5). The report laid the blame for this state of affairs at the feet of a "rising tide of mediocrity" (p. 5) and an education system that turned out large numbers of students who were ignorant of the past and unprepared for the future. The findings and conclusions of the Commission covered four important aspects of the educational process: curriculum content; expectations; time; and, teaching. Picking up the last of these themes, the Carnegie Forum on Education and the Economy established a Task Force on Teaching as a Profession which issued a report entitled *A Nation Prepared: Teachers for the 21st Century* (1986), which re-iterated concerns that the USA was no longer competitive in the world's marketplaces, and that other countries had overtaken the dominant position enjoyed by the USA. In pursuit of excellence in education, the Task Force addressed the status of teaching and proposed eight key planks for educational reform:

> the creation of a National Board for Professional Teaching Standards to establish high standards for what teachers need to know and be able to do, and to certify teachers who meet that standard
>
> the restructuring of schools to provide a professional environment for teaching, freeing them to decide how best to meet state and local goals for children while holding them accountable for student progress
>
> the restructuring of the teaching force, and the introduction of a new category of Lead Teachers with proven ability to provide active leadership in the redesign of schools and in helping colleagues to uphold high standards of learning and teaching

> that teachers be required to have a bachelors degree in the arts and sciences as a prerequisite for the professional study of teaching
>
> the development of a new professional curriculum in graduate schools of education leading to a Master of Teaching degree, based on systematic knowledge of teaching and including internships and residencies in the schools
>
> the mobilization of the nation's resources to prepare minority youngsters for teaching careers
>
> that incentives for teachers be related to school wide student performance, and that schools be provided with the technology, services and staff essential to teacher productivity
>
> that teachers' salaries and career opportunities be made competitive with those of other professions. (Carnegie Forum on Education and the Economy, 1986, pp. 55-6)

The first of these planks, the establishment of a national board that would "formulate high standards for what teachers need to know and be able to do, and certify teachers who met those standards" can be found in a commissioned position paper by Shulman and Sykes (1986) entitled *A national board for teaching? In search of a bold standard*, in which they discussed one set of options open to teaching as a profession. This paper became the blueprint for the development of the National Board for Professional Teaching Standards (NBPTS). Shulman and Sykes proposed that this national board should be an organisation that was 'by teachers, for teachers'. If teachers were to accept the concept of a national board, then they argued that full and active participation by practicing teachers was crucial to success. This was a central feature of the proposal, and has been carried through in practice by the National Board to this day. At least 50% of people involved in decision-making for the Board and its subsidiaries must be practicing teachers. This requirement was crucial in gaining the support and involvement of the two major teacher unions (American Federation of Teachers, AFT and the National Education Association, NEA), a feature that has brought some criticism from several quarters as noted below.

The Task Force proposed (Carnegie Forum on Education and the Economy, 1986, p. 66-69) that a National Board issue two types of certificates – a Teacher Certificate, and an Advanced Teachers Certificate. The first was to be mandatory and cover high-level

entry to the profession, while the second was to be voluntary and would indicate an advanced standard of competence as well as possession of qualities needed for leadership in the profession. These certificates would be specific to class levels and subjects taught. The two-tiered system was abandoned (entry level certification became the focus of the Interstate New Teacher Assessment and Support Consortium, INTASC, along with the National Council for the Accreditation of Teacher Education, NCATE), and the Board now issues only an advanced teacher certificate (National Board Certified Teacher), although the class level and subject specifications have been retained.

Created in 1987, NBPTS set about establishing "high and rigorous standards for what accomplished teachers should know and be able to do". Through exhaustive consultation with stakeholders in education, the Board articulated a model for excellence in teaching based on five core propositions that form the basis for the development of every standard:

> Teachers are committed to students and their learning.
>
> Teachers know the subjects they teach and how to teach those subjects to students.
>
> Teachers are responsible for managing and monitoring student learning.
>
> Teachers think systematically about their practice and learn from experience.
>
> Teachers are members of learning communities.

From this base, over 30 separate standards committees have developed specific standards for a wide range of primary and secondary teaching. These standards have been developed for specialist subject fields across four developmental levels from early childhood (ages 3-8) to adolescent and young adulthood (ages 14-18+). For elementary school teachers, generalist standards have been developed where the specialty is akin to a general medical practitioner – described by Shulman and Sykes (1986) as a horizontal specialty. Subject specific standards (vertical specialties) are available for middle and high school teachers, in much the same way that an ophthalmologist or anaesthetist specialises in medicine. The majority of each standards committee is made up of teachers who themselves display exemplary practice within the subject/student combination in question. Other members of the committee are acknowledged experts in child/adolescent development, teacher training or the relevant discipline.

From the five core propositions, each standards committee (generally made up of 15 members) drafted the standards to form the basis of the certification process. These standards not only have to reflect the five core propositions, but also have to identify the specific knowledge, skills, and dispositions that support accomplished practice; show how a teacher's professional judgment can be reflected in observable actions; and describe how the standards come to life in different settings. To reflect the complex nature of teaching and learning, a holistic approach rather than a narrow focus on specified behaviours has been adopted. The committee has to capture the essence of accomplished teaching without spelling out what they believe accomplished teachers must do in the classroom. Furthermore, the standards cannot constrain teachers to one or two favoured instructional models – to do so would be to limit the creativity in teachers thinking about their work and students. The Standards were drafted to enable the teacher to draw on a variety of fields related to teaching including "cognition, child development, motivation and behaviour, subject-specific pedagogy, organizational theory, and effective schooling" (Sclan, 1994, p. 2). As the Board stated, "The NBPTS Standards are designed to accommodate the variety of settings in which teachers work, reflect the reality of a range of teaching circumstances, and describe the multiple approaches teachers might take to reach curricular and pedagogical objectives" (National Board for Professional Teaching Standards, 2003b). The draft Standards have undergone several iterations of drafting, circulation for comment and critique (to teachers and educators, and the non-teaching public), then re-drafted until a general consensus was achieved. This process typically lasted from 15-18 months. The Standards so developed were intended to be rigorous, realistic and demanding, but not to the extent of being unattainable.

The rejection of behaviouristic objectives meant that new and innovative methods of performance assessment were required to enable candidates to demonstrate accomplished teaching as articulated in the Standards. Typical methods of teacher assessment (administrator ratings and records derived from isolated classroom observations, inspection, peer and self appraisal, teacher tests, and, student and parent surveys) would not capture the depth and breadth of knowledge, skills and dispositions that the Standards sought to describe. To maintain the authenticity of the assessment, the Board developed a series of portfolio entries, and exercises at an assessment centre.

In addition, they commissioned studies of the processes and outcomes to ensure that the assessments met the highest psychometric standards.

Teacher candidates have to prepare a portfolio that includes three entries that provide multiple sources of evidence about the depth and breadth of their teaching practice and professional maturity against the articulated Standards regarding their skills, knowledge and dispositions. This portfolio is compiled during one school year using one class. Two of the entries require videotaped classroom interactions, while all three entries must include samples of student learning products and other teaching artefacts. A detailed analysis of each entry must show how candidates translate knowledge and theory into practice. This commentary describes, analyses, explains, and reflects on their practice, and provides a rationale for the events that are recorded. In addition to the classroom related activities that are captured, a fourth entry in the portfolio requires candidates to document their involvement in the wider school and professional communities. These entries must indicate the quality of their contribution, as well as comment on the relevance of their accomplishments for student learning. Very detailed instructions are given about all of the materials to be submitted.

In addition they are required to attend one of over 300 assessment centres to complete six 30-minute exercises under test conditions. Designed to complement the portfolio entries, the assessment centre exercises validate the depth and breath of the candidate's content and pedagogical knowledge that was displayed in the portfolio, and cover other aspects as well. The stimulus materials used for these exercises are presented in three ways. Some are sent to the candidate's home in advance of the testing period, some are presented to the candidate on arrival at the assessment centre, while others are presented on screen during the assessment. These tests are computer administered.

The conversion of the Standards into tasks for assessment purpose has undergone change as more has been learned since the first administrations. However, five principles have continued to underpin the development of the tasks:

they should be authentic, and therefore complex;

they should be open-ended, to allow teachers to show their own practice without restriction or limitation to one pedagogical style;

they should provide candidates with the opportunity to analyse and reflect on their teaching;

subject matter knowledge should be evident in all of the entries, and

the tasks should encourage teachers to exemplify good practice. (Pearlman, 2002, slide 14)

In addition, each task was designed to assess a cluster of the Standards, and each standard was assessed by more than one task. By doing this, each teacher's performance was triangulated by the set of tasks. The tasks are only the first part of the assessment process, with the scoring system the second. From its inception, the National Board has paid particular attention to the technical issues related to the validity and reliability of its assessments, and has supported a comprehensive programme of research to assure the technical measurement quality of the assessments and their delivery. Considerable expense in terms of time and money has been expended on validation studies to review processes used for each assessment and scoring session; determine the magnitude and type of rater effects; testing a number of different models for reliability studies; the use of complete double scoring until 1999, then partial double scoring; a complete technical analysis of every set of scores, including descriptives; and the most comprehensive study of adverse impact ever completed on any assessment. On the basis of these studies, modifications have been made to the tasks and scoring processes, especially to reflect a more parsimonious approach to the collection and scoring of evidence. For instance, the assessment centre exercises have been scaled back from six 90-minute to six 30-minute exercises. Throughout, the message for task designers and scorers has been "validity, validity, validity".

The Board claimed that what is unique about the National Board Certification process is that it "assesses not only the knowledge teachers possess, but the actual use of their skills and professional judgment in the classroom as they work to improve student learning" (National Board for Professional Teaching Standards, 2003a). There has been a strong emphasis in the portfolio entries on professional reflection (Argyris & Schon, 1974). Argyris and Schon argued that the conventional wisdom that knowledge is brought to life through 'application' is insufficient to ascertain the difference between espoused theories (what you say you do) and theories-in-action (what you actually do). One of the documented problems inherent in accomplished or expert practice is that experts have difficulty in articulating what they do – for them it is so routine that they

struggle for words to describe it (Berliner, 1986). Reflection-on-action and reflection-in-action are critical in making what Schon (1987) termed 'professional artistry' manifest, and the Board has adopted this approach to emphasise that

> accomplished teachers in every field and at every level are aware of what they are doing as they teach and why they are doing it. This shows itself in the teacher's ability to set high and appropriate goals for student learning, to connect worthwhile learning experiences to those goals, and to articulate the connections between the goals and the experiences. They are conscious of where they want student learning to go and how they want to help students get there. Accomplished teachers also show the ability to analyze classroom interactions, student work products, and their own actions and plans in order to reflect on their practice and continually renew and reconstruct their goals and strategies (National Board for Professional Teaching Standards, 2003c, p. 5).

Two scorers score the candidates on each exercise, and a weighted total score computed across all ten exercises. The candidates can "bank" the scores for exercises that they pass, and re-take the exercises that they fail. The scoring criteria accompany each entry, and these form the basis of the scoring rubric. These rubrics seek to assess and score each entry "in light of the conscious, deliberate, analytical and reflective criteria the NBPTS Standards endorses" (National Board for Professional Teaching Standards, 2003a). However, the Board has been careful to point out that no one particular teaching style is mandated or sanctioned, and NBCTs do in fact demonstrate a variety of pedagogical approaches. Experienced teachers, who do not have to be NBCTs, score the portfolios and assessment centre exercises. These scorers are given thorough training to reduce variability and bias in scoring. The scorers are responsible for scoring a single exercise, and do not examine the entire *corpus* of a candidate's work.

Matters of validity have been paramount in the work of the Board, and for its credibility – the validity of the Standards, the validity of the assessment tools and scoring rubrics (especially the congruence between the tools/scores and the domain to be assessed), and the validity of the cut-scores set to distinguish those worthy of certification from those who are not. Standards for test construction and validation are contained in *Standards for educational and psychological testing* (American Educational Research Association, American Psychological Association, & National Council on

Measurement in Education, 2002), but the most relevant section and standard (Section 14 and Standard 14.14) states that the "focus of performance standards is on the levels of knowledge and performance necessary for safe and appropriate practice" (p. 156) and "defining the minimum level of knowledge and skill" (p. 157). The Board had set itself the goal of defining a high level of practice, so the content standards and their development, as well as the assessment standards and their development and outcomes, had to reflect a higher level of proof than a minimal level of competence for public protection.

In an investigation into the process validity for establishing the Standards for the Adolescent and Young Adulthood Mathematics certification standards, Hattie (1996), listed thirteen criteria for checking the development process. These criteria are:

> that the integrity of certification requires that the certifying board be administratively independent of any professional organisation.
>
> that the certifying board be solely responsible for constructing the standards.
>
> that the certifying board be composed primarily of those who are already highly accomplished teachers.
>
> that the universe of competencies be clearly defined.
>
> that the process of defining the complex content domain be developed on a sound scientific basis.
>
> that formal instructions be provided to the Standards Committee delineating their roles and responsibilities in setting the standards and demarcating the boundaries of the universe of content.
>
> that the process of developing the standards be formally documented.
>
> that after the standards are formally approved, Committee members have confidence in the process
>
> that the process involve defining critical aspects of practice that are distinguishing characteristics of highly accomplished teachers.
>
> that a process be followed that ensures that "high standards are set that recognise the variety of contexts in which teachers practice and that do not prescribe a single model".
>
> that the work of disciplinary groups, the states, NCATE, other standards committees, and others inform the standard-setting process.
>
> that the Standards Committee serve as the sounding board for the development

of the assessment measures and assist in designing fair and trustworthy processes.

that a wide sampling of agreement be sought for the standards from the major professional groups regarding the appropriateness and level of the standards.

Hattie concluded that each of these criteria had been met and that the process for establishing standards could be defended.

Every National Board assessment and the related content standards were subject to validation studies involving panels of highly experienced teachers in the relevant certification area from at least twelve states. The panellists responded independently to a series of questions regarding the relevance, importance, necessity and representativeness of the exercises following two days of training about the Standards, the assessments and the scoring of the assessments. These panels have found that the exercises and scoring rubrics were relevant and appropriate for the content being assessed (Crocker, 1997).

Further validation exercises were conducted on the scoring rubrics and their application by another panel of experienced teachers who had no previous association with the Board's work. These panelists worked in pairs, independently of the assessment panels, and ranked a sample of candidates' entries and exercises. These evaluations were compared with the scores awarded by the original assessors. In addition, they reviewed the evaluative comments that the assessors/scorers made when originally assigning scores to examine the extent to which the assessors based their scores on the content standards. These checks on the scoring rubrics indicated that with rare exceptions, the scoring system was consistently and appropriately applied (Jaeger, 1998).

The final psychometric validation study was to estimate the precision of the measurement involved in certifying accomplished teachers, and the setting of cut-scores. Once again, traditional reliability statistical approaches for the newly developed performance assessments were not appropriate, and new approaches had to be devised. Jaeger (1998) applied a stratified coefficient alpha approach to estimating reliability, and found that the estimated value increased from .82 without stratification to .84 with stratification, and that the estimated standard error of measurement was reduced

from 22.0 to 20.4 points. In addition, he was concerned to examine the error in the classification of candidates – that is, those who were of inferior standard but obtained certification (false positives), and those who were of superior standard but did not obtain certification (false negatives). Using Livingston and Lewis' methodology (1995), Jaeger found that the probability of a false-negative was .20 and the probability of a false positive .09. In effect, for the 258 examinees studied, there was a thirteen percent chance of misclassification, with 34 examinees being misclassified (18.5 denied certification when their true score indicated superior standard, and 15.5 obtaining certification whose performance was below the certification standard).

Standard setting is the process of establishing cut-scores to distinguish between two categories of applicants – those who meet the content standards for accomplished teaching, and those who do not. The Judgmental Policy Capturing procedure was initially employed (Jaeger, 1982,, 1995) as the method for computing performance standards. After familiarisation with the exercises for the assessment under consideration, the standard setting panels were provided with score profiles of the candidates, and independently classified each profile using a four-point scale – meets few of the standards, meets some of the standards, satisfies the standards, and, exceeds the standards for certification. Ordinary least squares multiple regression was used to fit each of the panellists' judgments to a model, and a weighting for each panellist calculated. This was followed by a session in which panellists received feedback about their rankings relative to all other rankings, and discussed the rationale for their judgments. A second iteration was completed, and similar feedback provided. The standard for performance was then computed from the average of each panellist's rankings. More recently, the Board has used a less complex standard setting procedure, the direct judgment method (Edwards, 1977; Edwards & Newman, 1982). This involves ranking all of the exercises, and then assigning the lowest ranked exercise a weight of 100 as the benchmark. All other exercises are then assigned a weight in terms of how much more important they are relative to the lowest ranked exercise. This is repeated twice, first using the second lowest exercise as the benchmark, and then the third lowest exercise. The entire process is repeated after each of the panellists reports on their judgments.

In either procedure, once the weights are calculated, the cut-scores can be determined. Again, judgment by panels of expert teachers is required. The Board also had to decide whether a compensatory system (where superior performance on one exercise compensates for poor performance on another exercise) or a conjunctive system (which specifies a minimum standard on the most important exercises in the assessment package that is the requirement for passing, without any amount of compensation) would be used to determine the final outcome for candidates. In fact, a modified combination of the two was used. The following example for the EA/Generalist portfolio entries used the dominant profile judgment method (see, for example Plake, Hambleton, & Jaegar, 1997; Traub, Haertel, & Shavelson, 1996) designed for use with profiles of polytomous scores on exercises in a performance-based assessment. The standard required for certification was:

a minimum score of 3 on the Teaching and Learning exercise
no scores of 1 on any of the exercises
no more than two scores of 2 on any of the other exercises
sum of scores across the six exercises must be 18 or higher

A similar approach has now been adopted across all certificates, and the performance standard is a total of 250 points.

In a psychometric study for a doctoral dissertation, Neustel (2001) investigated the information function of the final averaged exercise scores for the six portfolio and four assessment centre exercises of 8455 candidates from 1997-2000 in EA/English Language Arts, EC/Generalist, MC/Generalist. Information functions for both tests (portfolio and centre exercises) within each certificate area across years revealed that the assessments were roughly parallel, and that this was strongest in the portfolio assessment. Further comparisons of the average information functions indicated that the portfolios discriminated most between candidates on the construct, accomplished teaching. A further analysis, using logistic regression, confirmed this result and affirmed the Board's weighting scheme.

The first certificates were awarded in 1994, when 177 candidates for the Early Adolescent/English Language Arts and Early Adolescent/Generalist certificates were successful, and the first AYA/Mathematics certificates were awarded to 47 candidates

in 1997. To the end of 2003, a total of 23935 certificates have been awarded across 26 fields, including 1108 certificates for AYA/Mathematics.

Do Board Certified Teachers make a difference?

A major validity question that remained unanswered for many years concerned the relationship between Board certification and improved student achievement. In other words, do the students of NBCTs perform better on external validity measures of student achievement than the students of their non-Board certified peers? The ice began to melt with a comprehensive study, described by Vandervoort, Amrein-Beardsley and Berliner (2004, p. 10) as a 'unique and creative study', which compared the teaching practices of National Board Certified Teachers (NBCTs) with other experienced teachers who had failed to pass the NBPTS assessment. This study (Bond, Smith, Baker, & Hattie, 2000b) compared samples of student work from classrooms of the two groups of teachers, and is important because it compared the work of two experienced groups of teachers who thought that they were highly accomplished. The results indicated that NBCTs significantly outperform their peers who are not Board Certified on 11 of 13 key dimensions of teaching expertise, and outperform them on all 13 measures. The effect sizes ranged from just over .25 to 1.13 standard deviations. Further more, a discriminant function found that 85% of these teachers could be correctly classified into the two groups. However, this study did not use any standardised measures of achievement, opting instead to devise its own measures based on research-based features of expert classroom performance. Consequently, this study has been criticised for this perceived failing (Finn & Wilcox, 1999; Wilcox, 1999).

One of the first studies to explore the relationship between Board certification and improved student achievement (Stone, 2002) used Sanders' Tennessee Value Added Assessment System (TVAAS) data and set the criteria for "exceptional teaching as teaching that brings about an improvement in student achievement equal to 115% of one year's academic growth in the local school system". The study analysed performance of grade 3-8 students in the classes of sixteen of the forty Tennessee NBCTs for which state data from year 2000 was available. Stone's analysis lead him to conclude that none of these sixteen teachers could be considered effective as they failed to meet the standard in one of the required subjects (reading, language or mathematics),

or failed to meet it for three consecutive years. Indeed, he noted that none of them would meet Chattanooga's criteria for a salary bonus. However, the sample size was very small, and even smaller once it was noted that the substantive conclusions were based on only six of these teachers.

The relationship between Board certification and student achievement at the elementary level has also been studied in North Carolina (Goldhaber & Anthony, 2004). North Carolina has been at the forefront in adopting and promoting the National Board, and is the state with the greatest number of NBCTs. The researchers accessed the data held by the state on the performance of grade 3 to 5 students in reading and mathematics over a period of three years from 1996-7 to 1998-9. It was possible to link the teacher and student data and track both over time. The merged files represented a matching of over 770,000 out of 880,000 student observations with their teachers, with over 609,000 matched reading records, and over 611,000 matched mathematics records. The results indicate that the performance and growth in performance (in both reading and mathematics) of students taught by NBCTs were significantly better than the performance of students taught by unsuccessful candidates as well as the students of non-applicant teachers. One of the more interesting results in this study was when they compared teachers who were already NBCTs with teachers who would become Board certified in the future. In this case, the data indicated that the future NBCTs were more effective prior to the certification process than after they received their certificate. The authors suggested that the time requirements of the assessment process made the NBCTs less effective in the year they received their certificate. In addition, they explored the human capital benefits and the costs of National Board certification, and concluded that the greatest benefit of Board certification could be obtained by assigning NBCTs to the younger grades. Furthermore, they estimated that the cost per pupil of raising reading achievement by one standard deviation is about $7300.

This study together with the most recent investigation of the effectiveness of NBCTs (Vandervoort et al., 2004) provides the National Board with an answer to their critics. Vandervoort and colleagues were able to compare the achievement data from the students of 35 NBCTs and their non-certified colleagues in Arizona. Four years of data from the Stanford Achievement Tests in reading, mathematics and language for grades 3-6 indicated that the performance of students of NBCTs was superior to that of their

non-NBCT peers in almost three-quarters of the 48 comparisons made. In all of those cases where the students of non-NBCTs out-performed the students of NBCTs, the differences were not significant. On average, the effect size was .12 in favour of NBCTs, with greater effect sizes in reading and mathematics than in language. This effect translated into grade-equivalents amounts to a one month advantage for those students in the classes of NBCTs. Put another way, this amounts to the equivalent of an additional 25 days of instruction in a typical 180-day school calendar year. The authors concluded that

> given the weakness in the studies that show no relationship between Board certification and students achievement (Stone, 2002 and Stephens, 2003) and the strengths of the Bond, Smith, Baker and Hattie (2000) study (showing deeper student classroom work) as well as the Goldhaber and Anthony (2004) study and our own, the preponderance of the evidence suggest that students of NBCTs achieve more. (p. 36)

The NBPTS and its critics

Even in the foundation report *A nation prepared: Teachers for the 21st century*, the proposed concept of a National Board was not fully supported. Mary Futrell, who as President of the National Education Association was one of the two trade union representatives on the Carnegie Forum, attached a dissenting view in which *inter alia* she expressed concern about some aspects of the National Board proposal and the potential for abuse of the Lead Teacher proposal. Specifically, she was concerned that National Board certification could lead to differentiation among teachers and that this would send the message that "some teachers are more equal than others." Furthermore, she felt that there was the distinct possibility that the standards might be established and overseen by a body that was too far removed from the classroom. This latter concern was addressed when it was decided that a majority of the Board had to consist of practising teachers, and Futrell went on to become a valued member of the Board. As a counter-balance, the other teacher union representative, Al Shanker of the American Federation of Teachers, added a note to encourage support for the proposals, noting that while the report may not be the perfect document, it represented the views of a wide community, and for that it deserved support.

The Carnegie Forum had widespread representation from the education establishment (the leaders of the two major teacher unions; two state education superintendents, the Dean of an education school; the *New York Times*' lead education columnist, and John W. Gardner who had been Secretary of Health, Education, and Welfare from 1965-1968 and was well known for his belief in the responsibility of government as an agent of social change). Subsequently, the Board's by-laws stipulated that the Board must have a majority of active teachers, which gives the power to the teacher unions - 42 of the 63 directors are active teachers, with 14 selected for their teaching accomplishments, another 14 for their leadership in their teaching subject, and the final 14 from the leadership of the two major teachers unions (half from each union). Effectively, two thirds of the directors are union members or leaders. For this reason, several critics have accused the Board of provider capture. Interestingly, when making these claims, none of those critics have levelled similar accusations concerning the boards governing standards in law, medicine, pharmacy or accountancy, which have an even more pronounced representation from within their respective profession.

In opening her comprehensive critique, Wilcox (1999) acknowledged the importance of enhancing and maintaining teacher quality, but does not feel that the NBPTS has achieved what it set out to do. She explored the certification process, the validity of its scoring system, surveyed the extant literature on its effectiveness, and concluded that the Board's standards and assessments were too flawed to support the claims that were made on its behalf. In particular, she addressed six areas of concern:

1 The quality of the standards.

2 The validity of the scoring system.

3 The rationale for federal funding.

4 The objectivity of research on the Board's effectiveness.

5 The Board's ties to the teacher unions.

6 Its connections with state policymakers and teacher-training institutions.

1 The quality of the standards. Wilcox raised three questions – are the standards based on research or blind faith; are the standards focused on content or pedagogy; and, are the pedagogical assumptions based on rigorous research? She noted that according to Chuck Cascio the NBPTS Vice President for Certification, Standards and Teacher Development, the standards are not based on empirical research but on the "experiential research" of the teachers who serve on the committee. That is, they are the professional

consensus or "gut feeling" of those on the standards committee. On the second question, Wilcox claimed that the emphasis in the Standards is on the way a teacher teaches - they have a particular type of teaching in mind. From personal correspondence with Lawrence Braden and Ralph Raimi who carefully examined the AYA Maths Standards, Wilcox concluded that the Boards Standards are written "at the expense of a rigorous assessment of teachers' substantive expertise in their subject field" (Wilcox, 1999, p. 13). Braden and Raimi, she noted, claim that not enough emphasis has been placed on proven methods for teaching maths, a lack of specificity in geometry and algebra, and too much attention has been given to how teachers should interact with students (including different cultural approaches). That is, the National Board has mandated a particular teaching style - one that is inclusive and learner centred. Using medicine as a point of comparison for the final question, Wilcox asserted that very little research in education fully uses the scientific method including replication, random assignment and control groups (for a contrary view on the rigour and relevance of educational research, see Berliner (1987)).

2 The validity of the scoring system. In five sections, Wilcox raised issues concerning content and cheating; subjective assessments; the time and money needed to complete the portfolios and the cost of the assessment centres; and, the banking policy whereby unsuccessful candidates are able to "bank" or retain credit for any sections of the assessment for which they obtain pass scores. In the shortest of the five sections, Wilcox stated that "NBPTS operates on the assumption that the mark of the true professionalism is peer review" which she described as "the process by which educators judge one another … [and] that no other group, such as principals or local administrators, can gauge successful teaching". Her argument is that good teachers can be identified by tracking student academic progress over time (p. 19).

3 Government entanglements. To 1999, the National Board had received approximately $70 million from the federal government to conduct its research programme and to support needy candidates. No other professional organisation has received such funding. Furthermore, the General Accounting Office (GAO) has never audited the National Board and the use of these funds. Indeed, according to Wilcox, the Board has not "kept its promise to seek rigorous, outside reviews of its performance", as the sole review to date (Bond, Smith, Baker, & Hattie, 2000a) is contaminated by the fact that the Board paid for the research (which is exactly what the federal funding was granted for) and the researchers had professional ties to the Board. The Board's goal was to be

self-sustaining by 2001, but this objective has not been achieved. At the bi-annual conference in Washington DC in 2003, the delegates who attended Hill Day were to lobby their local and state congressional representatives for funding in fiscal year 2005 (FY05) for candidate support programmes. As Phillips and Kanstoroom put it "one is hard pressed to name any organization that has voluntarily left the federal gravy train once it was aboard." (1999, p. 71)

4 Research on Board effectiveness. Wilcox regarded the sole review to date (Bond et al., 2000a) as "seriously flawed" (p. 21) for two reasons – because the Board paid for it, and the researchers were professionally connected to the Board. A forthcoming study planned by the National Partnership for Effectiveness and Accountability in Teaching (NPEAT) has similar difficulties with conflict of interest - the principal researcher, Ann Harman was formerly the Director of Research for NBPTS, and the two university researchers (Lloyd Bond and John Hattie) had already conducted a review of the Board's effectiveness (Bond et al., 2000a). The measures used to determine teacher effectiveness steer away from the single measure that will satisfy the author and her funding sponsor – student outcomes measured by standardised test results. The two studies quoted in the previous section (Goldhaber & Anthony, 2004; Vandervoort et al., 2004) have provided convincing evidence that NBCTs do make a difference using this criterion.

5 Union involvement. While acknowledging that it is a very smart political move on the part of the Board to have a large proportion of teachers on its governing body, Wilcox regarded this as a threat to improving teacher quality, and the integrity of the Board. Teacher unions are interested first and foremost in the welfare of their members, and this will be uppermost in the minds of the union representatives on the Board, and not the learning of the students.

6 Connections with state policymakers and teacher-training institutions. The adoption of NBPTS-compatible standards by many states with respect to their teacher training and licensure practices has been an example of "state officials failing in their civic duty" (p24) as these standards have not been proven to improve student learning.

Wilcox concluded her paper by acknowledging that such a non-governmental organisation with laudable goals sounds wonderful, but "the Standards and assessments that the Board uses remain unproven and of questionable value" (p. 24). A similar critique can be found in Leef (2003) who asked whether North Carolina (the state with

the greatest number of NBCTs) was getting its money's worth from National Board certification. His conclusion was that the state was not getting value for money, and called for the cessation of the candidates' subsidy and of the salary increase that North Carolina NBCTs receive until there is proof of the beneficial impact of NBCTs on student learning.

Wilcox's colleague at the Fordham Foundation, Chester E Finn Jr (a former US Assistant Secretary of Education) has argued and promoted reports that are critical of the National Board (Finn & Wilcox, 1999,, 2000). The thrust of his argument has been that there was no evidence to show that the students of National Board Certified Teachers performed better academically than students of other teachers. He would like to see research like that conducted by William Sanders in the Tennessee Value-Added Assessment System (see, for example, Sanders, 1998; Sanders & Horn, 1998; Sanders & Rivers, 1996). Without that evidence, he argued, the National Board lacks credibility and validity in its work. A similar call for hard evidence before federal and state governments commit any more funds to the National Board has been taken by two economists, Ballou and Podgursky (Ballou & Podgursky, 1998a,, 1998b; Podgursky, 2001a,, 2001b). Their argument has been that Board certification is a poor substitute for merit pay, which should be determined in a more cost effective way by local supervisors, colleagues, as well as parent-consumers (Ballou & Podgursky, 1998b; Solmon & Podgursky, 1999). They have favoured an outcomes-based approach over the Board's preference for certification based on peer-review. Two recent developments have taken up this challenge. First, an alternative board has been established (the American Board for Certification of Teacher Excellence, ABCTE) which will certificate teachers at two levels – at entry to teaching and for veteran teachers – on the basis of a value-added portfolio. For entry to teaching, this consists of a pen and paper examination that covers subject content knowledge, and their knowledge of 'professional classroom skills'. Veteran teachers will have to demonstrate and prove their classroom effectiveness through student results on standardised achievement tests, and submit portfolios containing data based on an objective, external examination of their students' work. The ABCTE and its assessment procedures are still in development, and it has yet to define how the value-added data will be collected, presented and assessed. The first certificates are due to be issued in 2004. Secondly, as part of its continual research and development programme, NBPTS has now contracted

Sanders and his team to study the academic performance of the students of NBCTs and other teachers.

While Wilcox, Finn, Ballou and Podgursky have supported only teaching methods that are strongly correlated with student achievement as the basis for good or exemplary teaching, Scriven (1988b) dismissed this notion. He argued that correlates of good teaching are not adequate measures of good teaching, as it possible for a teacher to possess or demonstrate that correlate, yet not be a good teacher. For example, the use of advance organisers has correlated positively with increased student learning (Ausubel, 1978,, 1980; H. C. Johnson, 1980; Luiten, 1980), but there are teachers who consistently use advance organisers but are not good teachers. Instead, Scriven favoured a duties-based approach where 'responsibility' and 'demonstrability' are the keys to accountability in teacher evaluation. Although he was not addressing the issue of exemplary teaching, Scriven has provided an answer to the question 'How do you define good teaching?' He asserted that good teaching was whatever scores well on the duties list. His model of good teaching consisted of the following: knowledge of duties; knowledge of school and community; knowledge of subject matter; instructional design; gathering information about student learning; providing information about student learning; classroom skills; personal characteristics; service to profession. "The test of whether a given factor is part of the definition of "good teaching" must be distinguished from the question of whether teachers at a particular moment in history think it is a preferred practice" (Scriven, 1988b, p. 135). For this reason, he has not favoured the behaviours articulated in the Board's Standards through professional consensus, yet his list of duties bears a remarkable resemblance to the Standards enunciated by the various committees of the National Board.

One of the National Board critics has spoken with some inside experience, at least in the development of a Board assessment instrument. Anthony Petrosky spent four years leading a team to develop an assessment scheme for the Early Adolescent English Language Arts certificate. The portfolio exercises that Petrosky and his team devised were intended to produce "thick" case studies that demonstrate the complexity of the candidate's teaching. These exercises were well received by the National Board, but it was in the judging of these that the two parties fell out. Petrosky wanted to engage prospective judges in a system that took them one year to prepare for, through month

long summer institutes and on-going weekly workshops. The system he envisaged was "an approach grounded in an educative process … that emulated the kinds of training done by other professions and high stakes assessments (such as the education of judges for Olympic judging)" (Petrosky, 1994, p. 36). The NBPTS staff decided that this process was too expensive and time consuming (especially for the large number of candidates they were anticipating), and abandoned it in favour of a system that could train judges in three or four days, that placed emphasis on assigning scores to exercises, and provided "canned" feedback to candidates for each exercise, or, as Petrosky put it "reduce reality to a narrow over generalized description" (p. 37). In his view, the National Board took the traditional psychometric approach to reliability and validity in considering the defensibility of the assessments by seeking to objectively identify and scale behaviours, and in doing so, overlooked the stance that Petrosky advanced which is that the judging would be a matter of interpretation made by clinical and expert judges, who were well trained for this task. His position is premised on the view that teaching is a complex task that requires complex and innovative approaches to performance assessment, as this field of assessment is in its infancy. He felt bitter that his schema provided that innovation and degree of complexity, but was rejected.

Several of these critics have raised questions about the way that the Board has developed the content standards, and the assessment and scoring of these standards. Using the NBPTS as a case study of the growing use of standards-based performance assessment, Moss and Schutz (1999) presented the Board's "assessment development and evaluation practices as typical of those considered sufficiently sound to support the use of an assessment for high-stakes decisions" (p. 681). Without engaging in rhetoric, they carefully considered a range of psychometric issues associated with Board assessments at four critical points: development of the content standards; developing assessment tasks; developing rubrics and selecting benchmarks; and, setting the performance standard. In essence, their review acknowledged that the Board's assessment development "represents an impressive accomplishment. It served as evidence that complex portfolio and performance-based assessments can be mounted on a large scale and can meet professional standards of technical quality" (Moss & Schutz, 1999, p. 685). However, there were limitations in these assessments that exist as much in conventional practice as they do in Board assessments. The authors feared that by exposing these limitations, they would risk providing the Board's "more polemical

opponents" with the ammunition they needed. For example, they identified the need for a "strong" program of validity research, and not merely "confirmationist bias" which accumulates results consistent with the proposed interpretation. A 'strong' programme would involve making the "theoretical ideas as explicit as possible, and then devising deliberate challenges" (Cronbach, 1989, p. 152).

The development of the content standards is a distinctly social process informed by "experience, research evidence, values, personal and professional theories about good teaching, stories about their own teaching and that of others, and argument" (Moss & Schutz, 1999, p. 682), or as each of the standards themselves say "a professional consensus on the critical aspects of practice that distinguish exemplary teachers in this field" (see for example, National Board for Professional Teaching Standards, 1996, p. 1). However, the very nature of this consensus has been that the standards were broad and general, and while the response of teachers to the standards has been overwhelmingly positive, as Hattie and Clinton (in press) noted in their review of the EA/Generalist standards, the lack of detail in the form of supporting statements or vignettes can cause difficulties when teachers try to self-assess against the standards.

The development of the assessment tasks followed standard procedures. "Each assessment task can be matched to one or more standards, and … each standard has one or more assessment tasks that represent it" (Moss & Schutz, 1999, p. 683). However, the assessment tasks have only sampled from the Standards, and what is not known is how a candidate may perform (indeed, the likelihood that successful candidates would fail, or unsuccessful candidates pass) if different tasks, sources of evidence (for example, classroom observations or interviews) or sampling of the Standards were to occur.

The development of the scoring rubrics and selection of the benchmarks has been one of the more problematic stages in the certification process. The rubrics and benchmarks anchored the process and converted the generalities of the Standards into concrete examples that then reveal the accomplishments of the candidates. The authors are satisfied with issues of reliability in scoring using these rubrics and using the benchmarks given the training that markers receive, but there are questions of validity that need addressing. For example, while considerable care has been taken in writing

the rubrics and selecting the benchmarks, there has been no routine review by outside professionals to ensure that the scores sufficiently represent the content standards, and therefore to know just what exemplary teaching actually looks like.

With regard to the setting of the performance standards, the aspect that has stood out is that no single marker examines complete cases of a teacher's work. The panels involved in standard setting have based their decisions on profiles of scores across exercises, and have been unable to consider the actual performance on which the scores are based. It would be better, Moss and Schutz argued, if a panel read over all the evidence available about a candidate (or selection of candidates) to see whether that teacher's performance was truly indicative of accomplished teaching. Indeed, at no stage of the process does the entire body of evidence from a single candidate undergo examination by any group, and the public has to rely on the measurement community for the system's credibility. However given that reservation, Moss and Schutz asserted that the "passing standards are not inconsistent with the Board's uniform standard" (p. 685).

In setting out the form that the portfolio and exercises will take, the National Board has dictated the nature of the discourse that candidates engage in to successfully demonstrate their practice. One commentator has described this discourse as "self-disclosure, a form of confessing" (King, 1994, p. 102). The mechanism of these confessions may well become the new orthodoxy for describing behaviours, dispositions and expertise in teaching, and lead to institutionalising the Board's Standards, assessments and certification program as the legitimate and official view of teaching. Indeed, in a nation traditionally marked by state and local control of education, the National Board may well be the front-runner in the development of national standards for teaching, and thereby impose a de facto national pedagogy, structure, curriculum, teaching resources, and ideology. A consequence of this has been that the very diversity that the Board espouses becomes standardised into a set of universal truths and principles that define accomplished teaching. Once this occurs, difference and diversity are eradicated, and certain groups of people may be excluded. There has been some evidence to support this in the reports by Burroughs (Burroughs, 2001; Burroughs, Schwartz, & Hendricks-Lee, 2000) which provided case studies of the writing abilities of Board candidates, and concluded that writing apprehension, the difficulty in expressing tacit knowledge, reluctance to accept the sampling logic

required to assemble the portfolio and the role of evidence all worked against the African American candidate in the study. Moore (1999) investigated the relationship between learning style and success on NBPTS assessment, and found that there was no significant relationship between success and learning style as measured by Learning Type Measure (LTM), but that other factors were significant - those who assess themselves as better writers had more success; minorities do less well; and, suburban district teachers had an advantage. These three factors all work to minimise the likelihood of certain groups of teachers gaining certification, and may lead to stratification of the teaching force (King, 1994), with a status conscious elite that defends its privileged position in teaching .

Summary

The National Board model outlined above provides a clearly articulated set of performance standards and well-tested model of accomplished teaching. The demanding and rigorous standards are assessed using multiple sources to certificate exemplary teachers whose demonstrated practice meets the requirements of the standards. Validation studies show that teachers who meet the Board Standards differ from their non-certified colleagues on all indicators, although the Board has only recently been able to assemble evidence to demonstrate that the academic achievement on standardised tests of students in NBCT classes is measurably better. The National Board is confident that research in this area will vindicate its goal of strengthening teaching and improving learning, and there is emerging evidence to demonstrate that NBCTs do in fact make a difference in the classroom, however that is measured. As such, the Board model provides a gold standard for the purpose of developing items for a student evaluation instrument that can be used to identify accomplished teaching and to distinguish those characteristics that discriminate between accomplished teachers (taken as teachers who are Board Certified, NBCTs) and their non-certified colleagues.

Section Two. Student Evaluations of Teaching Performance

Student evaluations of teaching performance have been used for almost a century, and remain a controversial tool in the teacher evaluation toolkit (Penny, 2003b; Theall & Franklin, 2001). This section reviews SETs in the context of teacher evaluation in general, gives a brief history of their use and development, examines the many variables that have been subjected to study in the SET literature, and reviews the arguments that have been marshalled for and against their use. Of particular interest will be those few studies that report the use of SETs in high schools.

Teacher evaluation

Teacher evaluation, paralleling that of teacher professionalism, is a highly contested field, with enormous implications at the personal, inter-personal, legal, ethical, institutional, practical and political levels. Everyone it seems has a vested interest in such evaluation, and in the quality of the teaching force.

Teacher quality has been seen as the key to unlocking greater student achievement in schools. This is borne out by numerous studies that indicate that what goes on once the teacher closes the classroom door is most critical in determining student outcomes – in other words, the largest differences in student achievement occur between teachers (see for example, Rowe, 2003; Wenglinsky, 2000; Wilkinson & Hamilton, 2003; Wilkinson et al., 2000). Consequently, the movement to increase student achievement has focussed on improving teacher quality, and teacher evaluation has been considered a key element in achieving this goal. Quality teaching and the evaluation of teaching/teachers have become a focus of school improvement, teacher development and institutional decision-making.

In practice, teacher evaluation has been used for the two competing paradigms of formative diagnosis/feedback and summative monitoring of teaching effectiveness. The tension between these dual missions has frequently created difficulties as attempts have been made to create models of evaluation and instruments to serve both purposes – the

same dilemma the gardener faces when deciding whether 'to weed' or 'to cultivate' (Sawa, 1995). Frequently the school principal has been the person caught in this dilemma of purpose of use, and while they express the desire to have better teacher evaluation systems to aid them in both roles, principals exhibit a distinct reluctance to "weed" (Tucker, 1997). Furthermore, teacher evaluation systems and instruments have been designed by decision makers (political and educational) to "rid classrooms of incompetents, improve the performance of the average teachers, and differentiate and reward the expert pedagogue" (Berry & Ginsberg, 1989, p. 125). In their conclusion, Berry and Ginsberg noted that there was little evidence that many of the state-mandated teacher evaluation systems fulfilled any of the stated purposes. Teacher evaluation has now extended beyond the two original purposes to include personnel decisions concerning hiring new faculty; annual reviews of faculty; promotion decisions; school accreditation reviews; teaching awards and honours; and assigning teachers to courses, among others.

Sawa (1995) proposed six main purposes for teacher evaluation: (1) to improve instruction by fostering self-development and peer assistance; (2) staff development activities can be rated and identified; (3) the selection process can be validated; (4) to provide a major communication link between teachers and the school system; (5) for personnel decisions such as retention, transfer, tenure, promotion, demotion and dismissal; (6) to protect students from incompetent teachers by providing structured assistance to marginal teachers. Each of these purposes is individually and collectively important, but this thesis serves a seventh purpose – to identify high quality teachers (Prybylo, 1998, p. 560), and to determine what teaching characteristics make them stand out.

Teacher evaluation models

Shinkfield and Stufflebeam (1995) briefly described fifteen models of teacher evaluation, of which only the first four are based on classroom observation. These are (1) traditional impressionistic, with judgments made based on the experience and beliefs of the observer; (2) clinical supervision, proposed by Madeline Hunter (1985; , 1993); (3) research-based checklists, favoured by many states in their evaluation instruments; (4) high inference judgments by skilled and trained observers; (5)

interviewing which has been useful for teacher development, and is a common device in teacher selection, but divorced from what the teacher actually does in the classroom; (6) paper and pencil tests, such as those used by teacher training institutions for entry to the profession, or those proposed by the Fordham Foundation (1999), (7) management by objectives, which has agreed goals and measures for meeting those goals; (8) job analysis which describes what teachers do from observations and other data; (9) duties-based evaluation, which describes what teachers can legally be expected to do; (10) theory-based evaluation derived from a theory that links student achievement with certain teaching practices; (11) consumer ratings, which can include parent ratings as well as the more common student ratings; (12) peer ratings; (13) self-evaluation; (14) meta-evaluation of existing models; and the one that has gained increasing popularity in recent years (15) outcomes and value-added student learning such as the Tennessee Value-Added Assessment System (TVAAS).

In the past, the most common method used to ensure quality in teaching, was school and teacher inspection. Inspection was a common practice in Europe, the USA and New Zealand in the nineteenth century, and through the first half of the twentieth century (D. Peterson, Micceri, & Smith, 1985). These evaluation programmes were grounded in Taylor's theory of scientific management (F. W. Taylor, 1911), with its emphasis on standardisation, supervision, efficiency and productivity. In New Zealand until the early 1970s, school inspectors graded all teachers. The teacher was involved in some dialogue with the inspector and received formative feedback in that way, but the purpose was largely summative, and it was of very high stakes given that it was used for appointment and promotion decisions (Goddard, 2003). Since the advent of Tomorrow's Schools in 1989, school principals have largely taken over this role through the appraisal process, by visiting classes and looking at teacher performance in the classroom. There has been concern that principals are expected to fulfil the twin roles of academic leader and summative judge (Costa, 1988), with the consequence that the one conflicts with the other. Furthermore, almost eighty years of research has consistently reported low reliability of assessments of teachers by principals (K. D. Peterson, Wahlquist, & Bone, 2000).

Test data has also been suggested as a means of evaluating teacher competence. Two types of test results have been used – the achievement of students on a variety of tests,

and results on tests designed to measure a teacher's knowledge, skills and other factors believed to be pertinent to successful teaching – but these have proved to be problematic. An early example of testing teachers for their knowledge and skills can be found in the late nineteenth century, when the British government proposed a merit-based pay system for teachers so that teachers whose students scored above national norms received an increase in pay. Inspectors administered the tests to ensure objectivity in outcomes. The idea of a merit-based pay system was soon abandoned, as it proved difficult to determine the value-added by any individual teacher – other variables were soon acknowledged as making a contribution such as the impact of other teachers, prior student learning and aptitude, student background variables, school variables, and the difficulties of making adjustments for regression, an effect one critic called the "Robin Hood" effect, because it steals from the rich and gives to the poor (Medley, Coker, & Soar, 1984). With advances made in measurement techniques, state legislators believe that they can now overcome these difficulties, and a variety of attempts have been made to use student achievement results as measures for teacher evaluation. Perhaps the boldest of these has been the Tennessee Value-Added Assessment System (TVAAS) devised by Sanders and his colleagues (Sanders, 1998,, 2000; Sanders & Horn, 1994,, 1998) which purported to partial out all other variables and leave the value-added contribution of an individual teacher on a student or group of students. In the UK, value-added assessment has been developed by a team at the University of Durham (Fitz-Gibbon, 1985,, 1996,, 1997; Fitz-Gibbon & Tymms, 2002), which monitors teacher and school performance from results in the O-level and A-level external examinations. Some schools in New Zealand are adopting this process as a way of monitoring teaching and learning. Fitz-Gibbon and her associates have acknowledged that despite the advances in achievement measurement, there are still problems (particularly student mobility and tracking) with this form of monitoring which threaten the validity of the exercise (Fitz-Gibbon, 1997).

Considerable store has been placed, in some quarters (Finn, Kanstoroom, & Petrilli, 1999; Finn & Wilcox, 2000; Wilcox, 1999), on student achievement data as the most appropriate indicator of teacher effectiveness. Attaching consequences or accountability to student achievement data is "high stakes" for the two major participants – students and teachers – and these consequences are not always beneficial. Amrein and Berliner (2002) studied data from states that had introduced high stakes testing programmes and

showed that if the intended goal was to increase student learning, then the policy is not working, even if official state reports indicate to the contrary. Schools and districts have successfully subverted state testing programmes by careful coaching, and the exclusion of certain students from the test. After a careful analysis of ACT, SAT, NAEP and AP tests in these 18 states, and using archival time series to examine the effects, Amrein and Berliner concluded that testing programs may be increasing student scores on the high stakes state tests, but they are not increasing student learning as independently measured by these four testing programmes. In response, Rosenshine (2003) countered and noted that in the Amrein and Berliner study there was no comparison group. His study used states that had not attached consequences to their statewide tests, and concluded there was a "meaningful carryover" (p. 4) from attaching high stakes to state wide testing. In a re-analysis of Rosenshine's critique, Amrein-Beardsley and Berliner's findings (2003) contradicted this, as they noted that those states with state wide testing exempted more students than states without high stakes testing. At best, the difference between high stakes testing states and the comparison group was indeterminate. The most recent study on this topic, Braun (2004), agreed with Rosenshine and ascribed faults in Amrein-Beardsley and Berliner's re-analysis as "undermined by a too-zealous use of changes in exclusion rates as a basis for eliminating states from consideration" (p. 33). While teachers in Florida, one of the states to introduce high stakes testing, were not averse to accountability, the teachers believed that the Florida Comprehensive Assessment Test was not taking schools in the right direction (Jones & Egley, 2004). They noted the negative effects of the testing programme by narrowing the curriculum, on student and teacher motivation, and on teaching and learning. Positive comments on these themes were relatively infrequent.

As a means of assessment for entry and continuation in the teaching profession, teacher tests (for instance, the Praxis series of professional assessments for beginning teachers currently used in the US, the Content Mastery Examination for Educator (CMEE), National Teachers Examination (NTE), the Mississippi Teacher Assessment Instruments (MTAI), and the Massachusetts Educator Certification Test (MECT)) have been used. The latter test attracted considerable criticism when it was revealed that teachers performed poorly on the test, with the consequence that politicians were able to make the headlines at the expense of teachers (Ludlow, 2001). These tests in Massachusetts came under the spotlight as a result, and it was found that the tests would

not meet the industry standards laid out in the *Standards for educational and psychological testing* (American Educational Research Association et al., 2002) or earlier versions of the Standards (Ludlow, 2001). Twenty years earlier, a class-action suit in Alabama had exposed flaws in the development of the Alabama Initial Teacher Certification Testing Program. The tests discriminated against African-American candidates, and were deemed to be culturally biased largely because of the presence of a number of items with negative point-biserial correlation coefficients, and incorrectly keyed responses. While the Praxis tests are still in use for entry to the profession, the evaluation of practicing teachers now focuses on performance assessment rather than tests of content or pedagogical knowledge.

The NBPTS has placed considerable store in the use of performance assessments, as they are believed to hold the greatest potential for a comprehensive form of teacher evaluation. The face validity of the NBPTS performance assessment is high, as it assesses what the teacher does in the classroom, without the intrusion of a third party. The teacher has to document their work with the class, and submit a portfolio of exercises for assessment. Extensive training of the scorers is required for this form of evaluation, to ensure high reliability among them. As a consequence, the costs of this form of assessment and evaluation are high, which is one of the reasons that (the cheaper method of) student evaluations have become so common.

Observational methods of teacher evaluation have moved from impressionistic inspectors visit to more focused observations. Teacher peers, supervisors, students, external experts, parents and even ex-students some years after graduation have all been canvassed, using a variety of rating scale instruments (see for example, Howard, Conway, & Maxwell, 1985; Marsh, Overall, & Kesler, 1979; Murray, 1983). Medley, Coker and Soar (1984) were interested in determining whether student ratings could be used in preference to any of these other sources, and when they reviewed the literature for other sources they found:

> a few other studies [in] each of which the relationship between ratings and measured pupil achievement gains had also been estimated. Each study reported the same result, no relationship. Each study reached the same conclusion independently; that ratings made by reasonably sophisticated observers of

classroom performance had no validity as predictors of teacher effectiveness. (Medley et al., 1984, p. 46)

Student conceptions of teaching

All evaluators bring to their task a set of conceptions of what good teaching looks like, and students are no exceptions. Professionals may have a theoretical framework, but students' conceptions are grounded in experience. It is therefore useful to know the conceptions that students have, as there is a complex interaction between these conceptions and ratings. There have been numerous studies of teacher and student teachers' conceptions of both teaching and learning (Ayala & Martin, 1997; Boulton-Lewis, Smith, McCrindle, Burnett, & Campbell, 2001; Eklund-Myrskog, 1998; Jensen, 1998; B. L. Johnson, 1997; Kember & Wong, 2000; Pratt, Kelly, & Wong, 1999; Purdie, Pillay, & Boulton-Lewis, 2000; P. G. Taylor, 1996). Student conceptions of learning have also been well researched, as well as referenced against their approaches to learning in the classroom (see for example, Biggs, 1987) but their conceptions of school teaching are less well known. Batten (1994) conducted a survey in four Australian secondary schools. Students were asked to name three teachers that they had in the previous twelve months who were thought to be the best, and say why these teachers had been nominated. The results show that almost every teacher had their fan club (even if it had only one member). Analysis of the comments on what made the teacher the best revealed remarkable similarities to the kinds of things that are found in textbooks on teaching. The high frequency responses described a teacher who: helps you with your work; explains well so that you can understand; is friendly, doesn't yell at you and is easy to get on with; makes lessons interesting and enjoyable; cares about you, listens to you and understands you; has a sense of humour and will have a laugh with you; and controls the class. The study also grouped teachers according to their curriculum area and found that there were different student perceptions about mathematics/science teachers compared with humanities teachers, for example. Batten also provided a rationale for listening to and taking note of the student voice, when she wrote (p.4):

> The student perspective on learning and teaching is too often neglected or underrated in educational research. ... It is understandable that educators should

> be a little apprehensive about inviting student comment ... particularly in the middle years of secondary school, the student voice is more often heard in complaint than in praise. ... Underneath the negative and nonchalant veneer sometimes lies real discernment and perception; if probed, students will often give considered and insightful opinions on teaching and learning, based on their many years in the classroom. ... In the classroom setting, students are better placed than either principals or other teachers to identify and comment on examples of good teaching, because only students are constantly and directly exposed to the professional practice of a range of teaching styles.

Sizemore (1979; , 1981) asked black and white ninth and twelfth graders to select their three best teachers and three worst teachers, and state what were the most important behavioural differences between their nominated best and worst teachers. Although there were many significant differences between the teacher behaviours identified by each of the four groups (ninth grade blacks, ninth grade whites, twelfth grade blacks and twelfth grade whites), four of the top five behaviours that distinguished "best" teachers from the "worst" were identical: (1) the willingness to explain material adequately; (2) the ability to make the material interesting; (3) a willingness to help students with work; and (4) a caring attitude towards students.

In a Finnish study, Pekhonen (1992) examined the views of seventh graders with regards to the teaching and learning of mathematics. Replies from 514 students suggested that they have a task-oriented view of mathematics that emphasises process over product, that they stress working procedures and learning through practice, and favour student-centred activities and small group learning. Teachers are expected to help and provide direction for students. Some similar themes emerged when Brown and McIntyre (1993), convinced that students could provide the answer to what constitutes "good teaching", analysed student comments on the ten attributes of their best teachers. There was considerable agreement amongst the students particularly in the way the teachers constructed the learning environment, rather than the specific learning experiences. These categories were: (1) creation of a relaxed and enjoyable atmosphere; (2) retention of control in the classroom; (3) presentation of work in a way which interests and motivates students; (4) providing conditions so students understand the work; (5) making clear what students are to do and achieve; (6) judging what can be

expected of the student; (7) helping students with difficulties; (8) encouragement of students to raise their expectations of themselves; (9) development of personal, mature relationships with students; and (10) teacher's personal talents (subject-related or other). They concluded that "good teachers" were able to achieve and maintain a Normal Desirable State (NDS) of student activity, and that this state varies from one class/teacher to the next, and even from one lesson to the next. This state was achieved through automating their procedures and routinising classroom activity, which Berliner (1986) describes as one of the hallmarks of expertise.

In a study that involved undergraduate students at a Hong Kong university, Kember and Wong (2000) found that their conceptions fell along two continua – passive versus active learning, and transmission versus non-traditional teaching. They then compared the qualities described by these conceptions with the dimensions found in SET instruments. Students who had a passive belief about learning were more likely to attach higher importance to organisation, clarity of structure, workload, level of difficulty and specification of objectives. Students with an active orientation to learning attached greater importance to stimulation of interest, promotion of interaction in the class and displays of enthusiasm. They noted that one of the most notable absentees from SET questionnaires is any kind of statement about a variety of teaching approaches, and that the assessment of innovative or student centred teaching is problematic with standard student feedback questionnaires (p. 91). Kolitch and Dean (1999) reached similar conclusions when they analysed standard evaluation instruments with reference to the transmission paradigm and the engaged-critical paradigm.

The SET literature

Student evaluation of teacher performance is just one of many methods of teacher evaluation. However, of all the methods used for teacher evaluation, Student Evaluation of Teacher Performance (SET) is the most researched, debated and contested (Abrami & d'Apollonia, 1999; Greenwald & Gillmore, 1997; Marsh & Roche, 1999; McKeachie, 1997; Page, 1974). No other method of teacher evaluation has generated so much hostility, resentment, and distrust from teachers. SETs have been subject to considerable scrutiny, yet controversy continues to this day. This section of the thesis will briefly present a discussion for and against the use of SET, and argue that SETs are valid, reliable and useful when used in an appropriate way. Greater emphasis will be directed to those matters that might relate to this research, while other variables will be treated lightly.

Since the time of their first classroom encounter, students have expressed an opinion about their teachers. This is just as true of teachers in the ancient world as it is today. The works of Plato and Xenophon both expressed admiration for their mentor, Socrates, in a similar way to the writings of the disciples of Jesus in the four gospels. Indeed, we know that the great teachers of antiquity were those who managed to gather a group of students – Socrates in the marketplace, Aristotle in the Lyceum or Jesus at the Mount – and they were effective because they managed to attract and retain their disciples. To survive, professors in medieval universities had to attract students prepared to pay the fee, and if no students were attracted, then they had no source of income. The criterion for effectiveness was simple – the ability to attract and retain fee-paying students (Blank, 1985; Bonner, 1977; Corey, 2002).

While the comments and faithfulness of their students may not be generally regarded as student evaluations of the teacher in the sense that we use the term today, they contain two of the essential elements that have become part of the student evaluation literature in the twentieth century – the student has something to say about the teacher, and that message was heard in one forum or another. According to Darling-Hammond, Wise and Pease (1983) the use of student ratings assumes four things: (1) no other person is better

placed than the student to say whether the teacher has been able to motivate them to learn; (2) the intention of teaching is to change student behaviour; (3) student ratings provide feedback for the teacher; and, (4) good teaching may be motivated by recognition from students. These assumptions, they argued, provide a defence for the use of student ratings.

There are references in the late nineteenth century (Kratz, 1896) and the first half of the twentieth century to studies using student rating instruments to determine the characteristics of teachers from the student perspective (for example, Remmers & Brandenburg, 1927). At this time, interest in teachers was focused on the traits that they possessed that determined whether they were good or bad teachers. This interest was paralleled by the first student "anti-calendar" at Purdue University in 1924, in which students collected their own evaluations of faculty and published them so that future students could be alerted to the strengths and weaknesses of the instructors and courses they proposed taking. For many years, university authorities dismissed these publications as irrelevant and useless, but a study at Cornell University (Rayder, 1968) found that poor teaching was widespread and student unrest and dissatisfaction with teaching was justified.

Typically, data is collected using a rating form, with a set of statements about teacher proficiencies, and the students are asked to rate the teacher using a Likert-type scale on each of these. The scale commonly provides an ordered measurement continuum with gradations from *Strongly disagree* to *Strongly agree*. However, not all SET instruments have followed this rating scale pattern. For example, Scriven (1988a) developed an instrument that provided cues (prompts) and allowed the student to mark the cue if they felt that the feature mentioned was "particularly important" – an all or nothing approach. The cues were in two sections and indicated ways in which teachers fell short (Scriven argued that these negative cues have had the effect of eliminating the ceiling problem that frequently occurs with rating scales) and ways in which teachers could excel. The students could mark as many cues as they deemed appropriate. Salient scoring involved counting the number of times each cue was marked, and this indicated whether this cue was a strength or weakness for the teacher.

In addition to teacher evaluation, student feedback has been used for course evaluations, by providing data about the contribution of course related characteristics to the teaching/learning nexus (for example, meeting academic needs, volume of work required, selection of texts and course materials, assessment of the course, and the provision of facilities for teaching the course). Course evaluations are not a part of this thesis.

Haskell (1997) noted that the systematic collection of student evaluation data by the institution was recorded at the University of Washington in the early 1920's. Their use has grown to the extent that almost all higher education institutions now routinely collect this kind of data for one purpose or another with Seldin (1993) reporting an increase from 29 percent of colleges using SET in 1973 to 86 percent in 1993, while a more recent survey (G. D. Johnson et al., 2003) indicated that over 95% of all colleges and universities in the USA gathered student ratings data. The use of SETs in high schools was much less common, such that in 1996 it was estimated that as few as five percent of school districts systematically use student views in teacher evaluation (Loup, Garland, Ellett, & Rugutt, 1996; K. D. Peterson, 2000). There are even fewer SET studies in primary/elementary schools.

High school studies of SETs

Over half a century ago, Beecher (1949, p. 44) claimed that "it is reasonable to believe that high school pupils are as good subjective judges as any other group of persons". In keeping with this assertion, it is possible to find several instruments designed for use with school age students from the late 60s and early 70s, and they were not restricted to high school age students. One of the first in the field was the Purdue Teacher Evaluation Scale (PTES) (Bentley & Starry, 1970) which was designed for use with students from junior high school and beyond. It has six dimensions (1) ability to motivate students, (2) subject matter orientation of teacher, (3) student-teacher communication, (4) teaching methods and procedures, (5) ability to control students, (6) fairness of teachers, plus an overall evaluation item. The Illinois Ratings of Teacher Effectiveness (IRTE) (Blanchard, 1967) is an instrument for recording senior high school student perceptions of teacher performance in ten trait areas: teacher appearance, ability to explain, friendliness, grading fairness, discipline, outside classroom

assignments, enjoyment of teaching, voice, mannerisms, and command of subject matter. Students from Grades 3 to 12 in Anchorage Alaska, and their parents, have routinely completed a written evaluation of teachers, and the chief of the evaluation centre commented, “student input is about as good as or better than other teachers and even principals." (Bushweller, 1998, p.26). In Austin Texas, the Student Evaluation of Teacher II (SET II) was designed to capture the ratings of students below the fourth grade or, with disadvantaged students, through to the sixth grade (Haak, Kleiber, & Peck, 1972).

In an interesting variation on the usual pattern of researcher-designed instruments, a panel of high school students sitting on the Student Advisory Board to the Secretary of Education of Pennsylvania developed an instrument to provide feedback to teachers on their teaching. The Student Observation of Teachers and Teaching Techniques (StOTT) questionnaire (J. R. Masters, 1979) contained thirty-two positive statements concerning teacher classroom behaviour, using a five point scale for student responses. The first twenty-nine statements are included in five sub-scales: student-teacher relations; grades and testing; materials; teacher personality; and, teaching methods and techniques. The last three items in the questionnaire ask students to report on distracting mannerisms, the reason they took the course taught by the teacher, and whether they would recommend the course to another student. Several studies (J. R. Masters, 1977,, 1979; J. R. Masters & Weaver, 1977) have indicated that the StOTT has good estimates of reliability (0.80), and the scales make sense conceptually. In one study involving thirty-six teachers and 925 students, comparisons with other teacher ratings instruments indicated that student ratings compared favourably, and that students make judgments about teachers that are relatively stable over time. However, they found that students of differing abilities rated the same teachers differently, and recommended that the StOTT is useful for teachers to use in their own classrooms, but they should not be used to evaluate or compare high school teachers.

One of the most notable features of these high school student evaluation programmes is their incorporation into a much broader teacher evaluation scheme, which may also include parents. This was rare in tertiary institutions, but the high school studies often reported SET findings similar to those found in tertiary studies. For example, the Davis County School District in Utah offered teachers the option of incorporating student

survey data in their evaluation portfolios. Peterson, Wahlquist and Bone (2000) conducted a study, designed to test these survey instruments, empirically test the items, determine norms to assist in the interpretation of results, and assess the extent to which the participants (teachers and students) were satisfied with the process. They concluded that "student surveys are not merely popularity contests; students distinguish between merely liking a teacher and recognizing one who enables their learning" (p. 148). Of special interest was the positive response of the teachers to this system, given the level of dissatisfaction with evaluation based on the reports of their school principal. However, the study cautioned against the belief that high ratings equate to good teaching – rather they should be used along with other positive indicators as markers of quality classroom teaching. The Davis County policy of optional student surveys is indicative of the "other sources of evidence" approach that many evaluation systems adopt regarding student and parent surveys. Danielson and McGreal justified this by commenting "because they [parent and student surveys] are based on perceptions, evaluators should not consider parent and student surveys as *entirely* reliable sources of evidence." (2000, p. 51. Original emphasis). While this may be true for parent surveys, there is ample evidence to suggest that student surveys are very reliable sources of evidence, and certainly as reliable if not more so than any other source used for teacher evaluation.

One indication of this was the main finding of a study premised on the principle of 360 degree feedback (Wilkerson, Manatt, Rogers, & Maughan, 2000). Wilkerson and colleagues found that student ratings of teachers were the best predictor of student achievement on district-developed, criterion-referenced tests and showed the strongest positive relationship to student achievement when compared with those of principals and teachers. This supported two earlier studies on the usefulness of school student ratings. Tuckman and Oliver (1968) focused on the effectiveness of student rating feedback as a function of source. The teachers involved were vocational teachers in high schools and technical institutes, and they were subject to one of four feedback conditions: (1) from students only; (2) from supervisors only; (3) from students and supervisors; and, (4) no feedback. Their results showed that feedback from students only had the most beneficial effect, and that the addition of feedback from supervisors added nothing to this. Supervisor ratings alone had the opposite effect, such that teachers would have been better off receiving no feedback rather than feedback from

only their supervisors. These teachers were less receptive to feedback from their supervisors than they were to feedback from their students, which may reflect the student-centred nature of their work. Fox and colleagues (1983) surveyed 1657 Grade 6 students regarding 53 teachers, and compared the student ratings with adult observer assessments of teacher behaviour and with student characteristics measured by the Piers-Harris Children's Self-Concept Scale. They concluded that the student ratings were reliable and useful, even though these students were relatively young.

Mertler (1999) sought to determine whether high school teachers were receptive to SETs and whether they found them useful. Approximately 600 students in five Florida high schools evaluated the performance of 14 teachers on two occasions. On each occasion, the results of the surveys were tabulated and feedback provided to the teachers within a week. The overall reaction of the teachers was positive, who reported that SETs were feasible in high schools, and while they had some reservations about the honesty of students when completing the rating forms, they had endeavoured to modify their behaviour in the light of the feedback.

Two groups of academically talented high school students (N = 851) in a six-week university summer school programme rated the teachers in the programme using items describing low inference classroom behaviours (Worrell & Kuterbach, 2001). Outstanding teachers from local high schools were recruited for the programme, and university faculty and graduate students also participated. The seven highest rated and seven lowest rated classes were identified using global effectiveness scores, and these two groups compared using multivariate analysis of variance. The overall comparison was significant ($p<.001$), and the multivariate effect size was .69. Factor and regression analyses indicated that academically talented high school students could produce reliable scores when using a low inference behavioural rating instrument. In addition, the researchers found that these behaviours formed factors that were more predictive of overall effectiveness than student ability, predicted grade and workload.

Several studies with high school students have been conducted in New Zealand. A study conducted in three New Zealand secondary schools (Tod, 2000) revealed that many of the issues that have surrounded SETs at the tertiary level overseas also arise in high schools. Tod interviewed senior managers/principals, middle managers, teachers

and Year 12 and 13 students to assess the effectiveness of including student evaluations as part of a 360-degree feedback process. When asked whether SETs should be included, senior managers answered with an unequivocal Yes. However, Heads of Department believed that SETs were important, and that they needed to act on the information they received, but viewed student evaluations as important for informal feedback - for improvement of teaching, and of more use to individual teachers in improving teaching. The students overwhelming believed that teachers should seek feedback, but felt that only the senior school students in Years 12 and 13 should be asked. In their eyes, several of their teachers did not want any form of constructive criticism. The students also admitted to not always taking the evaluations seriously for a number of reasons including lack of anonymity; that end of year evaluations favoured by teachers would not benefit them [the students] personally; and that the forms used were felt to be too general with the focus on course and workload and not enough about the teacher. The students commented that they wanted to be in the classes of those teachers who were prepared to include items about themselves and their teaching practices. On the other hand, the students felt that the very teachers who needed to make the biggest changes were the ones who sought such information the least. When asked, students could recall examples of when teachers had changed their teaching as a result of feedback - and most of these examples matched those given by the teachers. Tod concluded that: (1) teachers recognise that feedback from students is a part of good teaching; (2) where the school's performance management system requires evidence of SET, these requirements are met, but that for many meeting the requirement seemed to be the only purpose; (3) a few teachers still find asking students for feedback threatening, (4) students, particularly senior students, want to have a say; (5) to offer meaningful and effective feedback, students have to believe that teachers will genuinely want to receive this feedback and that it will result in change; (6) the majority of teachers change their teaching as a result of feedback, and set personal PD goals as a result; (7) better processes are needed to ensure better anonymity and more valid data; and, (8) students prefer to offer informal feedback to teachers with whom they have a good working relationship.

Irving (1996) conducted a study involving 16 high school teachers in Auckland, and used student evaluations collected from 344 students on two occasions to measure the impact of feedback with consultation on teacher performance. At the same time,

teachers completed a self-evaluation. The teachers agreed that student evaluation data provided useful feedback for the improvement of teaching. Overall, there was little measurable impact on Time 2 ratings, but there was one important difference. Teachers who rated themselves as good teachers, but whose ratings they perceived as poor, made a significant improvement in their Time 2 ratings. In all other groups, the effect was negligible. It was argued that the cognitive dissonance (Festinger, 1957) between the student ratings and the teachers' own self-evaluation was a major catalyst that promoted positive change.

In a cross-cultural comparison (Chapman & Kelly, 1981), 880 Iranian and 599 American high school students completed the Classroom Behavior Survey. Although the dimensions on which they rated the course and content were similar, the study found that the students in each country used different dimensions to evaluate their teachers. Johannessen, Gronhaug, Risholm, and Mikalsen (1997) collected almost 500 Norwegian high school student ratings, and found that affective and emotional aspects played a major role in the way the students rated their teachers.

The evaluation of teacher trainees by the students in their practicum classrooms is not very common. Based on the thirty-eight item Pupil Observation Survey (POSR) (Veldman & Peck, 1967), Veldman (1970) selected two items from each of the subscales, slightly reworded them, and re-named it the Student Evaluation of Teaching (SET) instrument for use by teacher trainees and their supervisors to discuss implications for teaching practice as seen through classroom eyes. More recently, in a small case study of teacher trainees, ratings by high school students were compared with ratings by university professors, classroom teachers and the trainees' self-ratings (Stroh, 1991). The high school student evaluations were "consistently very close to those of the more experienced and 'knowledgeable' evaluators" (Stroh, 1991, p. 88). In an interesting aside, Stroh noted (p. 90) that the student teachers were "totally unafraid of the evaluation process" and that the high school students took considerable time to complete the evaluation forms as the process was so new to them.

Arguments for and against SETs

A major appeal which favours the use of SET simply states that students are inescapably there in the classroom and are therefore ideally placed to report on the teacher (Fraser, 1986). Students do not produce an "academic Heisenburg effect" (Page, 1974, p. 1), where their very presence changes the nature of what happens in the classroom. In addition, as one early commentator recorded, students "are no fools" when it comes to sizing up their teachers (Cole, 1940). In comparing different forms of teacher evaluation to the way in which the law courts weigh up evidence, Kulik and McKeachie (1975) noted that SETs relied on eyewitness, not hearsay, evidence. Worrell and Kuterbach (2001, p. 245) observed that "every year, teachers provide reliable and useful ratings of students using a variety of behavioral checklists. It is perhaps not surprising that students can also provide accurate ratings of teacher behavior as students spend as much time observing their teachers as their teacher spends observing them."

Consumerism provides another argument for SET. The advent of user pays education means that students have a desire to speak and be listened to regarding the education that they receive. In much the same way that the store customer is the consumer of the goods available to them, the student (along with their family and society in general) is the consumer of the work that teachers do. This may not be a legitimate reason for the use of student evaluations, as the evaluator (the student) is driven by personal self-interest and may not be well placed to adjudge all aspects of the *goods/services* on offer. When making choices, does the customer know whether the product (curriculum and pedagogy) is out of date, whether there are better choices at the next store (school), and do they know whether they are being ripped off at this checkout (school or class)? However, as Seldin (1993, p. A40) concluded, "The opinion of those who eat the dinner should be considered if we want to know how it tastes." McMillan and Cheney (1996) also addressed the use of the "student as customer" metaphor in education, and argued that while SETs are appealing, there can be negative consequences. There are at least four limitations regarding the use of this metaphor – it suggests undue distance between the students and the educational process; it highlights and promotes an entertainment model of classroom learning; it emphasises education as a product rather than a process; and, it reinforces individualism at the expense of community. Dowling (2000) argued that SETs are merely consumer satisfaction surveys that undermine teaching and

learning, by focussing on keeping the customer happy. For these reasons, the student as consumer is not a preferred rationale for SETs.

Advocates of SET have also used the "teacher as learner" movement to support their case. As a learner, the teacher wants to improve their teaching by learning more about their subject and more about their performance in teaching that subject. Therefore, the teacher will seek feedback on their performance to help them to improve. Meta-analytic studies have shown that the most powerful influence on learning is immediate and purposeful feedback (Bangert-Drowns, 1991; Hattie, 1999; Menges, 1990). More specifically, Dyer (2001), Manatt (2000) and Santeusanio (1998) have argued for the power of "360-degree feedback" which involves feedback from all possible stakeholders, including the students. SETs provide targeted feedback to teachers from the people who are best placed to observe on a daily basis the teacher's teaching behaviours and dispositions, and not those behaviours put on display for the "inspection visit" of a colleague, principal, or external evaluator.

When formative feedback is the purpose, then the feedback is likely to be only moderately successful in improving teaching through simple feedback loops (Braunstein, Klein, & Pachla, 1973; Miller, 1971), moderately successful with augmented feedback loops (teachers receive their results with some notes on how to interpret and use them) (P. A. Cohen & Herr, 1979, 1982; Friedlander, 1978; Toney, 1973), and most successful when consultation about best practices accompanies the SET results (Marsh et al., 1979; McKeachie et al., 1980; R. C. Wilson, 1986). However, SETs have little value if they do not address three key questions: (1) is the information new to the teacher, (2) is the teacher motivated to improve, and (3) does the teacher know how to improve? (McKeachie et al., 1980; Seldin, 1993).

In spite of a wealth of data in the research literature on SET, teachers still have reservations about their validity, reliability and utility, and prefer other forms of teacher evaluation, even when these alternative methods have been shown to be less valid and reliable. Over a quarter of a century ago, one author noted

> better teacher evaluation will only occur when educators attend to the issues of reliability and validity that have so far only been addressed in any substantial way in the research on student ratings of instruction. (Aubrecht, 1984, p. 89)

while more recently, other commentators have suggested that:

> it is not the quantity or quality of the research *per se* that has slowed the transfer of the empirical evidence, but rather other factors that have limited its application … in part, the responsibility for ignoring this wealth of scientific knowledge lies with the instructors themselves who feel secure in the knowledge that they know all there is to know about teaching, simply because they are actively engaged in the teaching process (Perry & Smart, 1997, p. 3).

The picture does not appear to have changed in any substantive way (Abrami, 2001a, 2001b; Kulik, 2001; Lewis, 2001a, 2001b).

Opposition comes from a number of quarters. In 1972, Hildebrand listed 23 common objections regarding the use of SETs, and answered them perceptively, even if he did so without reference to the growing body of literature that was developing to support their use. Hildebrand recognised that any change is a threat to the status quo, and that the occupants of an institution feel comfortable in the system they have lived with and have come to know. The literature on attributions (Schunk, 1996), efficacy (Bandura, 1977, 1986) and expectancy (R. Rosenthal, 1973, 1997) all indicates that there is a psychological foundation to teachers' fears – teachers feel threatened when others judge the quality of their work, especially when those judges are regarded as less qualified to act in this role than the teacher. Anxiety and resistance are common.

The objections that Hildebrand identified have persisted to the extent that a more recent journal article was entitled "Current concerns are past concerns" (Abrami & d'Apollonia, 1999). Penny (2003a) recently identified what she describes as four shortcomings in SET research: (1) insufficient attention to the validity of the construct "teaching effectiveness" and improving rating forms; (2) the failure to examine how administrators use ratings data to make decisions (which fuels on-going teacher discontent with the use of SETs); (3) too little attention to the practice and needs of teachers in using SETs for feedback; and, (4) the lack of consideration of the interaction between the students' own conceptions of learning and the teaching process. As a consequence, it is not surprising that whenever SETs are proposed at an institution, it is necessary to address concerns by publishing a list of the "myths" about SETs, and rebut them with research evidence (see for example, Bullock, 2004; Centre for Professional Development, 2004; Massey University, 1993). One survey of several hundred

administrators and teachers found that there was a 'surprising' lack of knowledge about literature on student ratings or even the basic statistical knowledge needed to read and interpret them properly. Indeed, this lack of knowledge was highly correlated with negative attitudes towards evaluation in general, students ratings in particular, and the value of student feedback (Franklin & Theall, 1989).

Several authors have conducted much of the research on student ratings (Abrami, 1977; Abrami, d'Apollonia, & Cohen, 1990; Aleamoni, 1999; P. A. Cohen, 1982; Feldman, 1997; Franklin & Theall, 1990; Marsh, 1984, 1994; Marsh & Roche, 1992; McKeachie, 1997; McKeachie, Lin, & Mann, 1971; Theall, Abrami, & Mets, 2001; Theall & Franklin, 2001), and have found that almost all concerns can be dismissed. More recently, Kulik (2001) reviewed the literature and addressed the controversy that surrounds SET, focusing on the agreement between student ratings and four of the most credible indicators of teacher effectiveness – student learning; student comments; alumni ratings; and, ratings of teaching by outside observers. In each case, his conclusion was that student ratings provide convincing support for teacher effectiveness.

Validity

The most common criticism of SETs is that students cannot validly and reliably evaluate a teacher's performance, particularly as they have little or no knowledge about the curriculum and pedagogy. However, no other group of people have had such constant and frequent exposure to teaching. Referring to the fifteen thousand hours (Rutter, Maughan, Mortimore, Ouston, & Smith, 1979) that a student typically spends at school, Fraser (1986, p. 1) noted that students had:

> a large stake in what happens to them at school, and students' reactions to and perceptions of their school experiences are significant.

Conceptions of validity have changed significantly in the past four decades in which SET have received considerable attention, especially since Messick re-defined the framework for validity (Messick, 1989). The traditional approaches of content, criterion and construct validity did not address the fundamental issues related to the meaning of the scores, their interpretation and the consequence of these interpretations. Messick

proposed that validity was concerned with the extent to which the empirical evidence and the theory supported the adequacy and appropriateness of interpretations using scores from an assessment. As a consequence, validity is now defined in the *Standards for educational and psychological testing* (2002, p. 1) as:

> the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.

It is clear from this definition that validity is a matter of degree rather than an absolute state (Linn & Gronlund, 2000), so in considering issues of validity, we wish to maximise the validity of our interpretations, and minimise any sources of invalidity that may lead to erroneous interpretations of the assessment. However, much of the key literature on student ratings was written in the 1980s, using the more traditional conception of validity, with the emphasis on content, criterion and construct validation of assessment instruments and this framed the way in which validity was addressed in the literature. Validity was an attribute of the instrument, not of the way it was used and interpreted.

Abrami, d'Apollonia and Cohen (1990) argued that "student rating validity" should be addressed by two key interpretations - one, if they reflect accurately student's opinions about the quality of instruction, regardless of whether ratings reflect what students learn; or second, if they accurately reflect instructional effectiveness. In addition, they underscore the importance of studying student ratings of instruction in the local institutional context to determine reliability, validity and utility more accurately.

In the first instance then, student evaluations must accurately reflect what students think about the quality of the instruction they receive. Several studies have investigated the correlations among several methods of obtaining student feedback. Correlations ranging from .81 to .94 were obtained in a study that used three sources of student feedback to measure the degree of congruence amongst them (Ory, Braskamp, & Pieper, 1980). There were no discernible differences between the feedback received from ratings on standard student questionnaires, open-ended questions and group interviews. A similar study (Tiberius, Sackin, & Cappe, 1987) compared student questionnaire data with the results of a discussion with the same class. Although the teachers preferred the ‘depth’

of feedback they received from the discussion, the authors found no difference in terms of what the feedback actually covered. Further evidence of the congruence between student ratings and other forms of gathering data from students can also be found (Greimel-Fuhrmann & Geyer, 2003; Kember & Wong, 2000; Pehkonen, 1992; Schmuck & Schmuck, 1989). Students themselves have indicated a preference for a variety of interview methods (Abbott, Wulff, Nyquist, Ropp, & Hess, 1990), but they question whether there would be any specific advantage in using the more time- and people-intensive methods such as interviews, focus groups, or small group instructional diagnosis (SGID) over the cost effectiveness of a standard SET instrument.

To reflect instructional effectiveness, student ratings need to correlate highly with some other measure(s) of teaching effectiveness. Those other measures do not have to be perfect measures of effective teaching, nor do they have to be complete. Indeed, there is no known single measure of teaching effectiveness that could be considered perfect or complete (Howard et al., 1985; Marsh, 1984; Marsh & Overall, 1980). As Cook and Campbell (1979) pointed out in their seminal discussion on quasi-experimentation, there may even be advantages in not achieving a single perfect and complete definition:

> we cannot in reality achieve widely accepted definitions of most constructs. This is because propositions about constructs are more reliable if they have been successfully tested not only across many overlapping operational representations of a single definition of a construct but also across representations of many overlapping definitions of the same construct. (pp.62-63)

In other words, the numerous studies of SETs that use different operationalisations of effective teaching strengthen rather than weaken the argument that SETs reliably measure teaching effectiveness.

The concept (and, hence, an agreed criterion) of teaching effectiveness remains elusive, even after many decades of study. If researchers and teachers are grappling with the problem of defining a conception of teacher effectiveness, then surely this is beyond the reach of students. A legitimate question to ask is "How can students be expected to evaluate teaching and their teacher, if they don't have a clear conception of what good or effective teaching is?" The answer appears to be that they do, at least as clear as that of teachers and researchers. Surveys of high school students using a variety of

techniques have found that they have a remarkably clear and consistent conception of effective teaching (Batten, 1989,, 1994; Clark, 1987; Greimel-Fuhrmann & Geyer, 2003; Johannessen et al., 1997; Kember & Wong, 2000; McCabe, 1995; Pehkonen, 1992; Schmuck & Schmuck, 1989; Wragg & Wood, 1984; Younger & Warrington, 1999). In a study with college students (Greenwood, Bridges, Ware, & McLean, 1973) it was shown that college students also had a consistent conception of good teaching to use as the yardstick for evaluating teaching. Indeed, the study was able to show that if you believe that students do not know what good or effective teaching is, then you have to believe that their teachers are even more inconsistent. As already noted, it would be more realistic to expect a consistent assessment from your students than from any other group of people involved in education (Blackburn & Clark, 1975; Doyle & Crichton, 1978).

While commentators have generally agreed that increased student learning should be the criterion for effective teaching, they have been unable to agree on how that criterion should be measured. Some (Finn & Wilcox, 2000; Stone, 2002; Wilcox, 1999) insisted that measured learning gains, such as those from the Tennessee Value-Added Assessment System (TVAAS) should be the only acceptable measure of teacher effectiveness. On the other hand, Scriven (Scriven, 1993; , 1994,, 1995) argued that student learning gains can result from unethical or bad teaching. The actual performance of a teacher should be the only legitimate criterion for effective teaching, and evaluations should not use learning gains as the criterion for that performance. Student learning gains, he claimed, are at best one indicator of teaching effectiveness but not a measure of that effectiveness. The construct validation approach to validity that predominates in the SET literature has endeavoured to demonstrate that SETs are logically related to a variety of indicators of effective teaching, and then supported this by using those indicators that have high correlations with the SETs. In the absence of an agreed criterion for effective teaching, and in the overwhelming presence of SETs, there has been a very real concern that student ratings are becoming the de facto criteria for effective teaching (Doyle & Whitely, 1974), with what Cohen (1997, p. 293) has called an “illusion of scientific exactitude” and a “form of legerdemain, serving some political agenda of questionable benevolence.”

Validation studies

Multi-section studies have been the most common way in which validation studies have been conducted. An ideal multi-section study would involve: (1) a course with large numbers of students in a number of different classes/sections, taught in "parallel"; (2) random assignment of students to each section; (3) pre-test measures that correlate substantially with final course performance; (4) each section is taught by a different teacher; (5) course characteristics (outline, textbooks, objectives; assessment, and the final examination) are the same for each section; (6) the final examination is constructed by someone who was not responsible for teaching the course (Marsh, 1984, p. 720). In addition, an external marker should mark the examination, and the sections would be pre-tested with the findings used as a covariate. By adhering to these requirements, the only difference between the sections is presumed to be the teacher and their teaching. However, there have been methodological problems with these studies, as most multi-section studies do not meet all or even most of these criteria. Firstly, the random assignment of students to each section has not been very common – students are placed in sections based on timetable constraints and other subject choices. Typically, the number of sections has been relatively small which generates large sampling errors. Finally, most multi-section courses were at the introductory level and results may not be generalisable to more advanced courses.

In spite of these methodological difficulties, multi-sections studies have been very popular, and have generally shown a strong positive correlation between the criterion of effective teaching (student achievement) and student ratings (see for example, Centra, 1977; P. A. Cohen, 1986,, 1987; Marsh, Fleiner, & Thomas, 1975; Sullivan & Skanes, 1974; Watkins, Marsh, & Young, 1987), although some individual studies may not be sufficiently consistent to convince teachers that they should use SETs to obtain reliable, valid and useful feedback. Teachers have been able to find at least one study that suits their preferences. For instance, one study in particular completely contradicted the evidence of so many other studies – the Rodin and Rodin study (1972). This highly publicised study still has its devotees, in spite of its significant methodological weaknesses. Rodin and Rodin reported that they had found a very strong negative correlation (-.75) between student ratings and student achievement. This result was startling, as it indicated that the better the students rated their teacher, the lower their achievement on the course. The calculus course they surveyed consisted of lectures (3

hours per week taken by the professor), “recitation sessions” or tutorials (one hour per week) and one hour devoted exclusively to administering test problems. Teaching assistants took the recitation and test sessions. In this study, the students were not rating the professor, but the teaching assistants, who had a relatively minor teaching role. The criterion used was not an end of course examination, but the total number of correct solutions obtained on the series of problems administered each week. Mastery of the problems was sought and students could re-take these problems as many times are they wished without penalty. Over the period of the course, students were able to change teaching assistants, so the results could not be reliably attributed to any one individual teaching assistant. No other reputable study has been able to replicate Rodin and Rodin’s negative findings. In one attempt to replicate Rodin and Rodin’s findings, the results were almost exactly the reverse, with correlations of .91 and .60 between student achievement on the final examination and teacher ratings for the two courses studied (Frey, 1973). Frey replicated his results two years later (Frey, Leonard, & Beatty, 1975), and an accompanying commentary (Scott, 1975, p. 445) used these findings to make the assertion that student ratings “constitute one of the most credible indicators of professional performance available”. Although largely discredited in the literature, some commentators (for example, Sproule, 2002) still cling to the Rodin results when they seek to critique the use of student evaluations.

Meta-analyses are a method for synthesising empirical studies, and calculating effect sizes to paint a composite picture or generalised effect for an intervention. These syntheses lead to general conclusions, showing the conditions under which the relationship is positive or negative, strong or weak, and providing some idea about the representativeness of the literature. In 1981, Cohen conducted a meta-analysis of all known SET studies involving multi-section courses, and found that student achievement was consistently correlated with student ratings factors of skill (.50), overall course (.47), student progress (.47), structure (.47), and overall instructor (.43). Several years later, Cohen updated (P. A. Cohen, 1986) then reanalysed (P. A. Cohen, 1987) his 1981 meta-analysis, and came to similar conclusions. There is a moderate positive correlation between student learning and overall instructor ratings, and on the dimensions of skill, rapport, structure, interaction, and evaluation. Other meta-analyses (d'Apollonia & Abrami, 1996; Dowell & Neal, 1982; McCallum, 1984) have produced similar positive correlations, though they have not always interpreted the results in the

same way. Cohen concluded that student evaluations were valid measures of teaching effectiveness, while Dowell and Neal suggested that the "validity of student ratings is modest at best and quite variable" (p. 59). This does not justify the abolition of student ratings in favour of other teacher evaluation processes – after all, as Dowell and Neal ask "where are the studies of the validity of the faculty peer evaluation process?" (p. 61).

Another approach used to determine the validity of student ratings is the multi trait, multi method (MTMM) methodology. This is a form of triangulation using different raters to assess the same traits or factors. McCall and Krause (1959) in an early attempt at value-added assessment measured the growth in the nine R's (reading, writing, arithmetic, research, reasoning, reporting, relationship with people, responsible work skills, and recreation) of sixth-grade students and compared this growth with the ratings of superintendents, supervisors, principals, colleagues, teachers themselves, and the students. They concluded that "superintendents, supervisors, principals and colleagues tended to rate good teachers low and poor teachers high ... the only persons in the school system who were found to be professionally competent to judge the worth of teachers were their sixth-grade pupils and the teachers themselves when giving a confidential self-rating" (McCall & Krause, 1959, p. 73). Aubrecht, Hanna and Hoyt (1986) compared student evaluations and teacher self-evaluations in eleven public and private high schools. Over 500 teachers cooperated, and they each provided self- and student-ratings for two of their classes. Aubrecht et al conducted parallel factor analyses for the student data and the teacher data, and these yielded almost identical structures. Measures of convergent validity were all significant, while those for discriminant validity were "inconsistent but predominantly favourable" (Aubrecht et al., 1986, p. 230). Another MTMM study (Drews, Burroughs, & Nokovich, 1987) tracked teacher and student ratings daily for 15 days. They noted that the student ratings paralleled the teacher self-ratings on a daily basis, which overcomes one of the theoretical objections to student ratings, namely that there is usually only one data point. Analysis showed that the student ratings and the teacher's self-ratings were significantly correlated in three areas: material covered, instructor performance and overall impressions of the success of the course. Drews and colleagues concluded that the finding that students agree with their instructors about day-to-day variability in their performance argues for

the “credibility of students as judges of teaching effectiveness” (Drews et al., 1987, p. 25).

Marsh conducted MTMM studies in which teachers and students completed the same evaluation instrument (Marsh, 1982b; Marsh et al., 1979). The same factors emerged when the student and teacher data were analysed separately. The agreement between students and teachers were significant on every dimension, the differences were small and non-significant for most items, and any significant differences were non-systematic. Howard and his colleagues (Howard et al., 1985) investigated 43 college instructors who were rated by students, colleagues, trained classroom observers, former students and themselves on four teaching dimensions (skill, rapport, structure and difficulty) plus one non-instructional variable, athletic ability. They found that students and former students had the highest validity coefficients for teacher effectiveness, and ascribed this to the amount of exposure that they had to the instructor’s teaching. Curiously, given complete exposure to their own teaching, the teacher self-ratings had the lowest validity coefficients. The measures of athletic ability served to confirm the discriminant validity of the ratings from the five groups. In a similar vein, Murray used trained observers in the classrooms of 54 teachers classified as high, medium or low on the basis of their student ratings (Murray, 1983). He found that the observers recorded similar differences between the teachers as were reflected in their SET ratings. In a subsequent investigation, Murray, Rushton and Paunonen (1990) studied colleague ratings in comparison with student ratings for 29 personality traits. There was clear evidence that for the 46 psychology teachers involved, student ratings of personality traits were strongly related to peer ratings.

Marsh conducted a reliability study (Marsh, 1981; , 1984) that considered the ratings of more than 1300 courses over 13 years to examine the teacher by course correlation. Table 1 shows the overall teacher rating and mean coefficient correlations for the same teacher teaching the same course; same teacher different course; different teacher same course; and, different teacher different course.

Table 1 Overall rating and mean coefficient correlations for teachers and courses

		Same teacher	Different teacher
Same course	Overall teacher	.72	-.05
	Mean coefficient	.71	.14
Different course	Overall teacher	.61	-.06
	Mean coefficient	.52	.06

Over the years under consideration, students on the same course gave very consistent ratings to the same teacher, and quite different ratings to other teachers on that same course. On different courses, the same teacher again received more consistent ratings than different teachers received. Furthermore, these indicated that students do not complete the course or teacher evaluations capriciously, but clearly distinguish good teaching according to what they do experience. In a similar study, the teacher ratings accounted for 5 to 10 times as much variance as the course variable which suggests that student ratings are influenced more by the teacher than any course factors (Marsh & Hocevar, 1984).

Validation studies based on multi-section studies, MTMM and meta-analyses all point in the same direction – that SETs fairly reflect what students think about their teacher and the teacher's performance, and that SETs are valid measures of teacher performance against selected criteria, especially student achievement. That is, students achieve well in the classes of teachers to whom they give good ratings.

Possible contaminants

SETs are frequently assailed as invalid because of a variety of influences. Variables identified as possible influences on student ratings are course variables, instructor variables, student variables, administration variables and instrument of measurement variables (Papalewis, 1990). Indeed, the literature often discusses these variables as 'contaminants' or biases in student ratings, an indication of the polarisation that occurs when SETs are under discussion. Marsh (1984) noted that many researchers infer that because they have found a correlation between a background variable and SETs, then the ratings are biased. To constitute a bias, he argued (1984, p. 709), it is not sufficient

to show that "some variable is correlated with students ratings and that a causal interpretation is warranted; it must also be shown that the variable is *not* correlated with effective teaching" (p. 734 – original emphasis), and this latter proposition is extremely problematic as it involves trying to prove the null hypothesis.

Issues related to course variables

Compulsory/elective

Several studies have found that students in compulsory courses give lower ratings to the course and teacher, whereas students in elective courses are more favourably disposed to both the course and the teacher (Gillmore & Brandenburg, 1974; Pohlmann, 1975). This relationship may reflect student motivation, as it is assumed that students are more motivated in courses they have freely chosen, than they are in courses that are required

Course difficulty and workload

Marsh (1994) stated that harder courses are rated more favourably on the basis that harder courses that require more work are worth striving for, whereas easier courses are not challenging enough and therefore receive lower ratings. Within a discipline, courses with a more difficult workload receive higher ratings (Cashin & Downey, 1992), with correlations of .44 between a measure of "working hard on this course" and an overall composite rating. This can also be interpreted in terms of the student's perceived value of the course. On the other hand, a number of studies indicated that there was only a weak or negligible correlation between course workload and student ratings (Burdsal & Bardo, 1986; Centra, 1977; Chang, 2000b; P. A. Cohen, 1981; Millea & Grimes, 2002). A New Zealand investigation found that all items except the difficulty/workload items of the Students' Evaluations of Educational Quality (SEEQ) and Endeavour rating instruments could distinguish among tertiary teachers who were rated as good, average or poor (Watkins et al., 1987).

Level of class

Several authors (Aleamoni & Hexner, 1980; Feldman, 1978; Marsh, 1987) have reported that higher-level courses receive better ratings than lower level undergraduate

courses, though the overall effect is small. It is interesting to note that teachers prefer to receive ratings from higher-level or graduate classes, than from their large lower level classes, and that students similarly believe that only upper level students should be asked to evaluate their teachers (Tod, 2000).

Time of day/year

There is a limited amount of research in this area, but it indicates that the time of the day that the course is offered or the time of the year has little influence on student ratings. (Feldman, 1978; Guthrie, 1954).

Class size

Studies on class size as an intervening variable on SETs are mixed. A study involving 1157 Spanish students (Mateo & Fernandez, 1996), found that class size had an effect on student ratings, but the effect sizes were small, while an English study (Watkins, 1990) with over 20,000 students found that the effects of class size were significant. Several studies have shown a curvilinear relationship. Pohlmann (1975) set out to replicate an earlier study by Gage (1961) and confirmed a curvilinear relationship, but unlike Gage found that the minimum rating occurred with class sizes between 106 and 120, and not 30 to 39. Marsh (1984, p.736) similarly found that small (N <15) and very large (N >200) classes evaluated more favourably, but not medium to large classes.

Full time/part time

This has not been the subject of extensive study but Watkins (1990) indicated that SET ratings are stable across varying proportions of full time students in the class.

Discipline/subject

Cashin and Downey (1995) were able to determine a hierarchy of subjects/disciplines according to whether they are rated highest or lowest. They found that teachers in the arts and humanities received the highest ratings, then, in descending order, biological and social sciences, business and computing, mathematics, engineering and the physical

sciences. Several other authors supported these findings (Centra, 1973b; Centra & Linn, 1973; Feldman, 1978).

Issues related to student variables (student presage)

Gender

The results of studies conducted regarding student and teacher gender have been inconsistent. Generally, they have indicated that there was no difference in the ratings made by male and female students. Most studies indicated no significant differences between the ratings given by male and female students to any one teacher or the ratings received by female or male teachers (Basow & Silberg, 1987; Feldman, 1992,, 1993). On the other hand, there have been indications that students give same gender teachers slightly higher ratings, but the differences overall were not significant (Kierstead, D'Agostino, & Dill, 1988).

Level of education

There are those who have argued that only the most senior students should evaluate their teachers – if at tertiary level, seniors or graduate students, or if at school level, then only the most senior students in the school (Tod, 2000) – on the basis that students generally lack maturity and experience, and rate their teachers capriciously. One study (Sailor, Worthen, & Shin, 1997) found that in lower-division and upper-division undergraduate courses, correlations between grades and ratings were positive, whereas in graduate courses, the grade-rating correlation was negative. This was attributed to a more critical approach to evaluating teaching in line with their more highly developed critical faculties as students progressed through the education system. On the other hand, Whitworth (2002) found that ratings of business teachers differed significantly according to level (graduate/undergraduate), with graduate students tending to give higher ratings than undergraduates.

Age

This variable may be more related to the level of students than specifically to their age, and has received little empirical attention. Worthington (2002) investigated the impact

of student age (inter alia) on student evaluation of teaching in Australian finance classes, and concluded that age had its greatest effect on items related to curriculum design, subject aims, and overall teacher ratings. However, this result could not be easily generalised as the age variable was classified in two ways – students between 21 and 30, and students over the age of 30. Klann and Hoff (1976) could not differentiate between the teacher ratings of the two age groups that they classified in their study – students over 20, and those under 20. Both of these age filters are too coarse to be of general use.

Motivation and Prior interest

Feldman (1978) recorded a small positive correlation between ratings and average intrinsic interest, and Cashin (1988) reported a correlation of .39 between a measure of "strong desire to take the course" and the overall course rating item. In one study, student motivation correlated .43, .71, and .53, respectively, with overall instructor rating, overall course rating, and goal attainment, while a general-interest item correlated .21, .49, and .31, respectively, with the course-interest item (Prave & Baril, 1993). However, this effect may not be uniform, as Marsh (1980) showed that prior interest was stronger than 15 other variables on the dimension to which it is most logically related, Learning/Value. As the Prave and Baril results indicated, prior interest is more of a function of the course rather than the teacher, so caution should be exercised when using and analysing student ratings for summative purposes, especially from large service courses such as introductory mathematics and statistics.

Expected grade

There has been a consistent finding in research related to this variable – expected grades are positively correlated with student ratings (Eiszler, 2002; Goldberg & Callahan, 1991; A. Moore, Masterson, Christophel, & Shea, 1996; Munz & Munz, 1997; Scherr & Scherr, 1990; Tatro, 1995). Marsh (1984) carefully examined three possible explanations for this consistent result – a "grading leniency hypothesis" (the teacher gives higher than deserved grades and in return receives higher than deserved ratings), a "validity hypothesis" (better expected grades are a reflection of better teaching and this in fact supports the validity of the ratings), and a "student characteristics hypothesis"

(certain pre-existing student characteristics have an impact on learning and actual grades such that the influence of expected grades is limited if not spurious). He concluded that there is evidence to support the validity hypothesis, and the student characteristic hypothesis, and the evidence "does not rule out the possibility that a grading leniency effect operates simultaneously" (Marsh, 1984, p.741), although he argued that this evidence is weak in effect, from experimental studies and may not be generalisable to authentic settings.

Attitude to SET

There are no studies that have specifically examined the relationship between students' attitudes towards completing SETs and the ratings that they give to teachers. It has been shown that students prefer to complete mid-term evaluation rather than end-of-course evaluations, as they can see if the feedback they have provided has had any effect (Abbott et al., 1990). However, they are sceptical as to whether teachers and administrators pay more than scant regard to the ratings (Marlin, 1987).

Ethnicity

There is anecdotal evidence that Asian students do not like SETs as they are reluctant to criticise or judge their teachers (Ory & Ryan, 2001). However, there is no empirical evidence to support this. In her doctoral dissertation, Holmes (1996) found no difference in the ratings obtained from Caucasian or non-Caucasian students in a college and a university setting.

Issues related to teacher variables (teacher presage)

Gender

As noted above, the relationships between student ratings and the gender of the teacher are complex. Empirical research can be found to support the argument that men receive more favourable ratings than women (Basow & Silberg, 1987), and vice versa (Feldman, 1992,, 1993). Kierstead D'Agostino and Dill (1988) suggested that teachers who behave in a gender stereotypical way receive better evaluations, and that students expect more of women instructors than men in regard to "both educational and

interpersonal aspects of teaching. If a higher level of performance is required in order for women to get SRIs [student ratings of instruction] that are comparable with those of men, then the same level of performance should elicit lower SRIs for women than for men. This is what we found in our experiments" (1988, p. 344). On the other hand, Theall and Franklin (2001) noted that women were often assigned to large lower level classes, and that their slightly lower ratings are a function of class level rather than gender. In a study of what gender preference students had for their teacher, Leeds, Stull, and Westbrook (1998) found that students preferred a male teacher who was a native-born first language speaker. Feldman (1992; , 1993) conducted a comprehensive synthesis of 39 gender studies, and found that the relation between teacher gender and student ratings is small (r = .02 in favour of women) and inconsistent. Exploring the most common control variable in teacher gender studies – the gender of the students evaluating the teacher - Feldman used a rank order analysis to search for patterns (many of the studies had insufficient information for any other analysis), and concluded that there was no consistent pattern of differences in the way in which male and female students rated male and female teachers, although there did tend to be a same-gender preference.

Status – tenured/professor

Some early studies have indicated that the higher the rank, the higher the ratings, although the relationship is weak (Gage, 1961; Guthrie, 1954). Full professors obtained higher ratings than assistant and associate professors, and teacher assistants received the lowest ratings. Aleamoni with various colleagues (Aleamoni & Graham, 1974; Aleamoni & Thomas, 1980; Aleamoni & Yimer, 1973) has found no significant relationship between rank and ratings. In a study that stands at odds with these, Schuckman (1990) found that Teacher Assistants received better ratings than tenured staff.

Research productivity

Many studies (Aleamoni & Yimer, 1973; Centra, 1983; Linsky & Straus, 1975; Marsh, 1987) have reported at best a very weak positive relationship between research productivity and the ratings a teacher received. Linsky and Straus' study (1975) noted

no correlation on the global rating item, but the correlation with teacher's knowledge (the factor to which it is most logically related) was .27.

Popularity and reputation

One of the classical myths concerning SETs is that teacher reputation and popularity have an undue positive influence on ratings. This myth is fed by the notion that a teacher who teaches an easy course and awards easy grades will increase their popularity and receive more favourable ratings. Felder (1992) noted that these courses and teachers in fact receive lower ratings. This myth also depends on the conjecture that students can be "seduced" into awarding better grades to teachers who are more popular, but, as stated previously, there is ample evidence that students are discriminating judges. Several studies have specifically examined the relationship between reputation and ratings. In one study, students who had heard positive prior information about the teacher awarded higher ratings to both the teacher and the course than students who had heard negative information (Griffin, 2001). This supported an earlier study in which students, who had selected a specific course section on the basis of teacher reputation, gave higher ratings than classmates who were simply assigned to the same section (Leventhal, Abrami, & Perry, 1976). However, these studies did not examine the relationship of popularity with multidimensional factors of SETs, and there is no evidence that popularity loads on any factors other than those to which it would be logically related.

Expressiveness – the Dr Fox effect

The Dr Fox studies are very pertinent to this research. A frequently expressed concern with student ratings is that in which a charismatic, entertaining teacher can "seduce" good ratings from their students, even if their teaching has little or no value in terms of possible student learning – hence the term "educational seduction". If this concern is true, then this is the triumph of style over substance. It is further exacerbated by the contention that school students may be more capricious and more susceptible to influences of this nature and therefore their ratings may be less valid than those obtained from more mature audiences. The original Dr Fox study (Naftulin, Ware, & Donnelly, 1973) employed a professional actor to perform for a group of graduate

medical students at a conference, using an expressive manner. The lecture content was designed to have little educational value. After the lecture the students were surveyed and the ratings were favourable across the evaluation criteria. Naftulin and his colleagues argued that the seductive nature of the delivery unduly influenced the ratings of the students. Critics pointed out a number of methodological difficulties with this study, and Williams and Ware devised several other studies to address these shortcomings (Ware & Williams, 1975,, 1977; R. G. Williams & Ware, 1976). A number of their critics (Abrami, Leventhal, & Perry, 1982; Perry, 1985; Perry, Abrami, & Leventhal, 1979) had difficulty replicating these results and concluded that educational seduction is not supported by empirical evidence, and that educational seduction may not be a reliable effect. Marsh and Ware (1982) reanalysed the Dr Fox data, and concluded that a high degree of expressiveness positively affected the ratings on the instructor Enthusiasm scale (which is most logically related to expressiveness), while the low content score negatively affected the Instructor Knowledge and Organisation/Clarity scales (again the logically related scales on the instrument). There is little support for this hypothesis now, but the controversy was recently revived when Williams and Ceci (1997) published a study that showed that a change in vocal expressiveness produced a large change in student ratings, but no effect on exam scores. However, this was a single subject study (one of the authors, Ceci), and on previous ratings, expressiveness was his lowest rating item, so there was the greatest potential for increased ratings on this particular item.

Personality

Surprisingly, this has been only infrequently studied. When personality was inferred from student or colleague reports, Feldman (1986) found that there were some significant correlations with student ratings, but when personality was inferred from teacher self- reports, very few characteristics correlated significantly. In a more recent similar investigation of the influence of personality on ratings, Murray, Rushton and Paunonen (1990) concluded that dependent on the type of course, peer ratings of teacher personality traits were strongly correlated with the ratings that students gave to those teachers. These traits correlated highest with student ratings for “leadership, extraversion, liberalism, supportingness, intellectual curiosity and changeableness” (p.

259). However, they noted that these characteristics did not generalise across types of course, with considerable variation between undergraduate and graduate courses.

Humour

Allied with teacher expressiveness is the use of humour in the classroom, and the purported effect it has on student ratings. The importance and value of a sense of humour should not be underestimated, as in studies about what students think makes for an effective teacher, one of the consistent features mentioned is a sense of humour (Batten, 1994; Cravens, 1996; Ogden, 1994; Phelan, Davidson, & Cao, 1992; Schmuck & Schmuck, 1989; B. N. Young, Whitley, & Helton, 1998). However, studies have not shown this to impact on student ratings on other than the specific scale (instructor enthusiasm or individual rapport, for example) where a sense of humour is appropriate (Benz & Blatt, 1995; Krehbiel & McClure, 1997; Marsh & Bailey, 1993; Rodabaugh & Kravitz, 1994). In reviewing one of his earlier studies, in which he analysed the written comments that students are often asked to make in conjunction with rating their teachers, Aleamoni (1987) found that students would compliment their teachers for their warmth and humour, but roundly criticise them if their lessons and courses were not well organized, or if their methods of encouraging and supporting students were poor. He concluded that "students are not easily fooled. In rating their instructors, students discriminate among various aspects of teaching ability: if a teacher tells great jokes and has the students in the palm of his or her hand in the classroom, he or she will receive high ratings for humor and classroom manner, but these ratings do not influence students assessment of other teaching skills" (p. 27). This is a further indication that students are discriminating judges of effective teaching, and acknowledges the strengths and weaknesses of teachers appropriately.

Grading leniency

One variable that has received considerable attention has been grading leniency. A general finding in multi-section studies has been the moderate degree of positive association between student ratings and student achievement as measured by their grades (Chang, 2000a; Eiszler, 2002; Goldberg & Callahan, 1991; A. Moore et al., 1996; Munz & Munz, 1997; Olivares, 2001). This attention has been focused on grades

as they are often used as the proxy for student achievement, and consequently teachers may "buy" good ratings through lenient grading practices. On the contrary, a study which measured the impact of grading leniency on high school student ratings of their teachers (Brooks, 1990) found that teachers who graded strictly received better ratings than those who graded leniently. Franklin, Theall and Ludlow (1991) found no relationship between frequency of evaluation and two indicators of grade inflation: average end of term grades and student ratings of workload.

One of the most discussed grading leniency studies used a revised ratings instrument that added several measures to those previously used at the University of Washington. In this study, Greenwald and Gillmore (1997) found that there was a positive grades-ratings correlation, with an average standardised path coefficient of .45 over three studies. They then reviewed five possible theories that may explain the grades-ratings correlations, and concluded that of the two direct-cause relationships (where grades influence ratings), the leniency theory best explained the negative relationships between grades and workload. Marsh and Roche (2000) rebutted the positive effects of grading leniency and low workloads on student ratings that Greenwald and Gillmore concluded, describing them as popular myths. Marsh and Roche re-analysed the published Greenwald and Gillmore data, and noted that the single measure of student achievement used by Greenwald analysis was a student self-report measure, for want of anything better. Marsh and Roche argued that a stronger test would involve multiple measures of student achievement. Once achievement, prior student characteristics, and course characteristics had been controlled for, the grade-leniency relationship was substantially reduced. Their findings in this study indicated that there was not a systematic relationship between grades and ratings, and that the curvilinear relationship (with grades well below average generating the steeper part of the curve) is better explained by attribution theory that posits that a poor student will attribute poor grades to external causes (including poor teaching), while good grades are attributed to internal causes. Marsh and Roche found that there was a slight increase in SETs over time, but these were not significant. Workload increased, not decreased, and there was a small quadratic effect in grades where they initially rose and then fell.

Body language

There are several studies that have examined student ratings and teachers' body language. Ambady and Rosenthal (1993) showed silent video clips ("thin slices of expressive behavior") to a group of "stranger" observers and asked them to rate the teachers. They found a correlation of .76 between these "stranger" ratings and students ratings, and that principal's ratings of the teacher could also be predicted from the "strangers" ratings. The study used video clips of only 13 teachers, with consequent large standard error. Babad, Avni-Babad and Rosenthal (2003) also used "thin slices", this time with high school teachers. They found that the relationship between non-verbal behaviours and SETs vary across teaching situations. The relationship is positive while the teacher is disciplining the class and interacting with the class, but negative when frontal teaching and especially so when the teacher treated high and low students differentially. They concluded that student anger at what they saw as unfair treatment was visited upon those teachers who acted in this way.

Physical attractiveness

Laboratory studies dominate research in this area. Typically, photographs of teachers are shown to students, together with a description of certain teaching characteristics. Ambady and Rosenthal (1993) in their studies of the non-verbal characteristics of college teachers and high school teachers found that the correlation between ratings of teachers' physical attractiveness and the criterion variable (one or some combination of student ratings, peer ratings and supervisor/principal ratings) was not strong. Buck and Tiene (1989) used college seniors preparing for a teaching career, and found that physical attractiveness had no effect on rating of competence. There was, however, a significant interaction between gender, attractiveness and authoritarianism. In general, authoritarianism was associated with negative evaluations, but an attractive authoritarian female teacher received significantly higher ratings than the three other authoritarian groups (attractive and unattractive males, and unattractive females). In a more recent study that used photographs taken from faculty websites (Hamermesh & Parker, 2003), six students (three male and three female) rated instructors on their physical beauty. Those instructors who were viewed as better looking received higher ratings, to the extent that a shift in the beauty rating from one standard deviation below the mean to one standard deviation above resulted in an increase of .46 in the mean

rating, almost a one standard deviation increase. The standard error in each of these studies is likely to be large, as the number of raters in each study was relatively small.

Teacher strategies and styles

Strategies that teachers use influence the way in which the teacher is perceived and rated by their students. In a study involving 448 high school social studies students, Smith (1984) assigned them to one of 16 groups defined by four bipolar conditions (uncertainty versus no uncertainty; bluffing versus no bluffing; discontinuity versus no discontinuity; and, notes handouts versus no notes handouts). After each lesson, students completed a comprehension test of the material taught, and completed a lesson evaluation. Bluffing and discontinuity negatively affected the SETs, indicating that students did not appreciate these behaviours in teaching style. The use of active learning techniques (cooperative learning, small group learning, and peer reviewed essay writing) was trialed in teaching probability to advanced level college students (J. S. Rosenthal, 1995). The students favourably received the emphasis on participation in a subject that is traditionally presented by lecture. Forty-nine ninth-grade algebra teachers were the subject of an investigation into the relation between their attitudes to mathematics, teaching and students, and the ratings they received from external observers and students (McConnell, 1978). Student ratings favoured those teachers who were clear in their presentation, more varied in their presentation, more enthusiastic, more indirect, asked more higher order cognitive questions, and were less critical of students. When the effects of active teaching strategies (such as group assignments and discussion sessions) on teaching evaluations were measured in economics classes (Leeds et al., 1998), they found that there was very little direct impact on SETs. However, active teaching strategies also had no impact on actual or perceived learning, and the only direct teaching strategy that approached statistical significance (the use of class discussion) had a negative impact on perceived learning. The study authors concluded that economics may not lend itself to these more direct instruction techniques, and that a well-prepared and structured lecture may be more effective than poorly organised class discussions.

Attitude to SETs

Surveys have been conducted in many institutions to determine staff attitude to SETs, with only a few assessing the extent to which those attitudes are reflected in SETs. Seventy members of the faculty of a teachers college in Taiwan completed an attitudes to SET questionnaire (Attitudes to Student Rating of Instruction, ASRI), and were rated by their students (Chang, 2002). Those instructors who had positive attitudes to SETs received student ratings that were significantly better than their colleagues who had negative attitudes to SETs. There is evidence to suggest that morale is not negatively affected by SETs (Jacobs, 1987), that tenured staff have less favourable views on the summative uses for SETs when compared with non-tenured staff (Avi-Itzhak & Kremer, 1985), and that when teachers become better informed about the SET literature they become more positive about using SETs (Franklin & Theall, 1989).

Issues related to instrument variables

Instrumentation

The actual instruments used for student evaluations of teacher performance can and do have an impact on the validity of the inferences that can be drawn from the data – students need to be asked the right questions without bias, ambiguity or inappropriate content acting as confounding variables. The principles of questionnaire and test construction have been well defined over the years with several authors paying attention to them (Aiken, 1996, 1997; Berk, 1979; Oppenheim, 1992). Content of the items and instrument are critical to validity. At the time that Tagomori and Bishop (1995) conducted their study, they could find no prior studies that specifically examined the content of SETs. Their study involved a content analysis of the items in a sample of 200 evaluation instruments used in 414 schools of education. They found that approximately 58% of the instruments had flaws such as (1) ambiguous or unclear items (54.6% of items and over 90% of instruments), (2) lack of correspondence between the item and observable behaviours that teachers have control over and which students could fairly be expected to evaluate (24.5% of items and over 90% of instruments), or (3) response patterns/anchors that were skewed, unclear or ambiguous (58.0% of instruments). In all, an average of 82.3% of instruments contained one or more of these flaws. Many of these instruments were “home grown” varieties, often

cobbled together from other sources, and this indicates the need for very careful analysis of the items prior to use.

Issues related to administration and purpose variables

Anonymity, and the presence of the teacher in the room

There are a number of opponents of student ratings (Crumbley, 1995; Emery, Kramer, & Tian, 2003) who have argued that the only valid rating is when the student has to attach their name to the form. Their rationale is that even in a court of law, you are entitled to know who the witnesses are. Furthermore, they argue, students know the identity of the teacher who wrote their [the student's] evaluation/report. While these arguments have certain appeal, they ignore the unequal power relationship that exists between students and teachers, especially with regard to the success or otherwise of the student on the course. Students feel that if the teacher could track any criticism (even constructive) to a particular student, then that student's grades could suffer. Argulewicz and O'Keeffe (1978) with high school students, Feldman (1979) and Blunt (1991) have all found that ratings are higher when the students had to put their name on the ratings. Dommeyer and colleagues (Dommeyer, Baum, Chapman, & Hanna, 2002; Dommeyer, Baum, & Hanna, 2002) observed a similar phenomenon when they investigated two methods of administration – paper and online. Online response rates were lower, a consequence of the lack of anonymity according to the students. The presence of the teacher in the room also had the effect of raising ratings (Feldman, 1979), which is one of the reasons why administration by a third party is generally recommended.

Summative/formative

While there is almost universal agreement that student ratings can and should be used for instructional improvement, caution is often advised when using student evaluations for summative purposes (for example, tenure and promotion). In many cases, the use of student ratings may be used along with other sources of information. Ironically, the information obtained from these other sources (self, peer, supervisor and other forms of evaluation) is less psychometrically sound than SETs. The summative/formative debate continues, and is the source of more heat than light, often supplemented by anecdotal "evidence". Svinicki (1998) lamented this situation, and noted how personal stories

were more powerful than "facts" in this debate. In an experiment to determine whether the purpose of the evaluation influenced the way in which students rated their teachers, Young, Delli and Johnson (1999) used Marsh's SEEQ instrument which has been well validated (Marsh, 1982a,, 1983,, 1984), and randomly assigned students in a number of courses to one of three conditions – control (the directions on the ratings sheet were neutral), formative (the directions indicated that the results would be used for instructional improvement), and summative (the directions indicated that the results would be used for determining the teacher's annual salary increment). There was no statistically significant difference between the conditions on the dimensions on SEEQ, nor on either of the two canonical discriminant functions when analysed for classification to actual group membership. These findings support Frankhouser's (1984) conclusions, but run counter to those of Aleamoni and Hexner (1980) and Braskamp and colleagues (1984) who suggested that SETs tended to be higher if the stated purpose was promotion or tenure.

End- or mid- course evaluations

Student ratings obtained mid-course are highly correlated with those obtained at the end of the course (Marsh & Overall, 1980), and seem to be immune to timing (before, during, or after) relative to the final examination (Feldman, 1979). Frey (1976) also measured the effect of timing relative to examinations for a multi-section calculus course, and found that ratings taken at the time of examinations were very similar to those taken at the start of the next quarter. Indeed, ratings taken at the start of the next quarter show a slightly stronger relationship with exam performance. This could suggest a grade inflation effect as the students knew their examination results, but when the between-section and within-section correlations between ratings and class performance on the final exam relative to the predicted performance based on a linear regression for the mathematics SAT score were considered, he found that students were not rewarding or punishing teachers on the basis of their calculus results.

Publication of results

According to a survey conducted by a committee at the University of South Alabama (G. D. Johnson et al., 2003), the results of student ratings usually remain confidential to

the teacher and the relevant sections of the organisation (the teacher, department chair, deans, and other involved in ensuring teacher quality in the university). They reported that publication of the results is not common, and noted that teachers use student ratings for improvement and buy into their collection provided they are not published. Freedom of Information laws and privacy provisions are at odds here, and while publication was more common in the 1970s, the data suggest that publication has declined through lack of interest.

The net effect of all of these background variables however is relatively small. Brown (1976) estimated that they accounted for 14% of variance, most of which is accounted for by expected grade. Marsh's 1980 study of 16 background variables indicated that they accounted for approximately 13% of variance, and this ranged from 20% on the overall course item to 2% on the factors of Organisation and Rapport. Burton (1975) had an estimate of 8-15%, most of which was accounted for by student enthusiasm. McKeachie concluded that "most of the factors which might be expected to invalidate ratings have relatively small effects and those factors which affect ratings also affect learning" (1979, p. 390).

Dimensionality

The question of whether teaching is a unidimensional or multidimensional activity has been debated quite vigorously. Through logical, theoretical and empirical analysis, Marsh (1984) posited that teaching effectiveness is multifaceted, that there is no single criterion of effective teaching, and that "different dimensions or factors of students' evaluations will correlate more highly with different indicators of effective teaching" (p. 709). Therefore, he argued, the evaluation of teaching must also be multidimensional, and no single measure can convincingly capture the effectiveness of a teacher's performance. Furthermore, it would not be possible to fully understand the effect of so-called biases without consideration of the different factors or dimensions of teaching. Often, a bias has an impact on the factor to which it is logically related, but little or no influence on other dimensions. If the purpose of the evaluation is teacher improvement, then a single averaged score based on the ratings conveys insufficient information to help the teacher address aspects of teaching that need improvement. A description of strengths and weaknesses would be more useful (this teacher is

enthusiastic but not well organised), and this can be obtained from the items and factors that measure the construct of effective teaching. Several studies (Aleamoni & Thomas, 1980; McBean, 1991; McBean & Lennox, 1987) have shown that a single global item has a low correlation with specific items, but has a higher relationship with the many different variables related to the assessment, and that a single item should not be used as the basis for making decisions. Psychometrically, the reliability of items used individually is problematic, and the relationship between "true score" and "observed score" is lower for individual items than for summated scales – multiple observations increase reliability.

In summative situations, a full picture can be obtained from a multidimensional approach, and it is not necessary to adopt a single score approach. Cashin and Downey (1992) suggested that it would be better to use a shorter form of the instrument for summative purposes rather than rely on a single global item. In a carefully considered factor analytic study, Burdsal and Bardo (1986) challenged the notion that single items could be used as indicators. The instrument they used had three such global items (overall course evaluation, general ranking of the teacher, and a recommendation to other students about the teacher), but all of them had their highest loading on different primary factors, suggesting that single items convey "overly simplistic representations of students perceptions of the instructor" (p. 75). However, that is not a view shared by all, as other researchers have argued that decisions for summative purposes often require a single measure that enables a choice to be made between competing interests. Abrami, D'Apollonia and Rosenfield (1997) analysed the inter-item correlation matrices derived from 1184 items taken from the 17 ratings forms used in 43 multi-section validity studies. Using principal components extraction, their factor analysis indicated a single general factor that accounted for 68.2% of the variance. On rotation, the variance was re-distributed across three correlated factors, and one uncorrelated factor. The global rating items (such as, overall, this teacher is a good teacher or overall, this course is a good course) all loaded on this first factor. Although they agreed that teaching effectiveness was multidimensional, they suggested five reasons why it was appropriate to use global items for summative purposes: (1) most rating forms share a similar set of global ratings items; (2) global items have relatively high validity coefficients; (3) different settings have a greater impact on specific dimensions than they do on global items; (4) global items mostly load on the first few factors, even

on well designed multidimensional instruments such as SEEQ; and, (5) the factor analysis in their study confirms the preceding four points.

Concluding comments on SETs

SETs have been extensively researched with respect to a wide range of context, presage, process, purpose and outcome variables. In each case, critics and supporters of SETs have assembled competing evidence. However, the net effect of these studies suggests that SETs are a valid, reliable, and useful means of indicating teacher effectiveness, when used in an appropriate way.

This review has also demonstrated that where studies have been conducted with high school students, they parallel the comprehensive evidence compiled about the use of SETs in tertiary education. With this in mind, this study will regard the opinions of high school students, expressed through a student evaluation instrument developed specifically for the purpose, as valid, reliable, dependable and useful reports about their mathematics teacher.

Chapter Three: Study One

This study began with the search for the "holy grail", a set of standards that comprehensively describe what highly accomplished teachers know, do and care about, and provide a system for certificating teachers who meet or surpass those standards. New Zealand does not have a set of standards that describe accomplished teaching in high school mathematics. Indeed, there are no such standards for any of the levels of schooling or for any of the curriculum subjects. Several countries (Australia, England) are currently grappling with ways to identify and recognise the best teachers that they have. The National Board for Professional Teaching Standards (NBPTS) in the United States of America has done just that.

The AYA/Mathematics Standards are described in 11 Standards organised around four large themes, with each Standard in two parts – a summary of the Standard, followed by an elaboration that provides the texture against which the assessments for certification are structured. The first theme is Commitment, consisting of Standard I, Commitment to Students and Their Learning. The second theme is Knowledge of Students, Mathematics and Teaching with three Standards – Standard II, Knowledge of Students; Standard III, Knowledge of Mathematics; and, Standard IV, Knowledge of Teaching Practice. There are four Standards in the third theme, The Teaching of Mathematics – Standard V, The Art of Teaching; Standard VI, Learning Environment; Standard VII, Reasoning and Thinking Mathematically; and, Standard VIII, Assessment. The final theme, Professional Development and Outreach, consists of Standard IX, Reflection and Growth; Standard X, Families and Communities; and, Standard XI, Contributing to the Professional Community.

In the absence of any description of accomplished teaching in New Zealand, a qualitative study was designed to determine the extent to which the NBPTS AYA/Mathematics Standards could be used as a proxy for highly accomplished mathematics teaching in New Zealand secondary education. More specifically, can the

Standards developed in the USA by the NBPTS be applied in New Zealand, and what modifications (if any) are necessary to make them more accurately reflect the New Zealand teaching environment? Research on teaching and learning most often claims that conceptions of teaching, learning and knowing are rooted in cultural and social structures (Biggs, 1996; Nasser & Abouchedid, 2000; Pratt et al., 1999), and as a consequence notions of exemplary practice in one environment/culture may not be transportable to another. Furthermore, using instruments developed in one culture for a particular purpose without demonstrating the relevance of the construct to the new culture is to engage in what Triandis calls "pseudo-etic research" (Triandis, 1972, p. 39). This study will show that there are few differences between the notions of highly accomplished mathematics teaching in the USA and in New Zealand, and this will enable the instrument developed for the research to be readily used in either context.

To determine the applicability of the AYA/Mathematics Standards, a series of focus groups were conducted. A focus group can be defined as "a group of people, with certain characteristics who provide data of a qualitative nature in a focused discussion" (Krueger, 1988, p. 27). A major benefit of using a focus group is that the explicit use of group dynamics and interaction produces "data and insights that would be less accessible without the interaction found in a group" (Morgan, 1988, p. 12). Focus groups provide an interactive process that is an alternative to the more commonly used interview or questionnaire. However, this much-vaunted interaction between participants may be more illusory than the methodology purports to represent. Kitzinger (1994, p. 103) completed a review of 40 published studies that used focus group data, and found that she could not find a single study that concentrated on a conversation between participants, while the overwhelming majority included quotations from only one participant at a time.

Focus groups have relatively high face validity as a means of gathering a large and rich amount of qualitative data in the respondent's own words, at a reasonably low cost, while providing speedy results for the researcher (Krueger, 1988; Stewart & Shamdasani, 1991). However, as Krueger (1988) noted, focus groups require careful management for success, as it is possible for the group to highjack the agenda. Responses are not independent, but the validity of focus group data is enhanced when other data is used to triangulate the results and to confirm the findings as well as to

obtain depth and breadth. The relatively small number of respondents involved can limit the generalisation of the results to the population at large - a difficulty that careful selection may overcome.

Membership selection

The selection and conduct of the focus groups was adapted from the procedures outlined by several authors. (Krueger, 1988; McLennan, 1992; Morgan, 1988; Stewart & Shamdasani, 1991). Three focus groups were proposed (two for mainstream mathematics teaching in the English language, and one for teachers of Pāngarau and Māori mathematics teachers in the mainstream), to be comprised of experienced and accomplished teachers from the Greater Auckland region who were chosen by the reputational method.

Nominations for the first two focus groups were obtained from two sources – the district mathematics advisor, and the senior lecturer in secondary mathematics at the local college of education. These people have ready access to all schools in the region and are in a position to observe teachers within their own school environments. These two sources were asked to independently identify accomplished mathematics teachers who were experienced at teaching classes at AYA/Mathematics level (that is, in New Zealand, at Years 11 to 13). A total of 21 names were obtained by this method, with 16 receiving nominations from both sources. These 16 teachers were invited by letter to participate in the focus groups. A total of 14 agreed to participate, but one of these was not able to attend either of the sessions and did not wish to be subsequently interviewed. A second person was also unable to attend the sessions, but was willing to be personally interviewed. A one-to-one interview was conducted, but this latter discussion focussed on pedagogical issues related to the evolving Numeracy Project in New Zealand, and not whether the standards could be applied in New Zealand as a fair and reasonable measure of exemplary teaching. Therefore, this discussion has been excluded from further analysis.

The third planned focus group was a hui with experienced and accomplished Māori teachers of mathematics, teaching either the Pāngarau curriculum in te reo Māori or the Mathematics curriculum in English. They were also to be chosen by the reputational

method. After consultation with the district mathematics advisers and college of education staff, and correspondence with prominent Māori educators and mathematics educators, it was not possible to conduct this hui. The teaching of Pāngarau in high schools is rare above Year 10 (the NBPTS AYA Standards refer to teaching at the equivalent of Years 11-13), and teachers approached felt that their current experience of teaching at the desired level was minimal and therefore their contribution would not be appropriate. One Māori mathematics teacher who teaches in a mainstream high school was available, and agreed to review the Standards and comment on them. He was later interviewed by telephone, and his comments reinforced the dominant themes from the two focus groups included in this study. However, his comments could not be used to generalise to Māori teachers in the mainstream, nor to teachers of Pāngarau.

Conduct of the focus groups

Two focus groups were conducted with a total of twelve participants. The first had eight participants, and four participants attended the second. Approximately two weeks before the focus group meeting, each participant received a copy of the NBPTS Standards, presented in a three column format. The units for analysis were numbered in the first column, and the second contained the text of the Standards, and the third column was left blank for the participants to make notes. The participants also received a set of protocols for the meeting, which outlined the way in which the focus group meeting would be conducted. The meetings were held after school at a convenient location for the participants, and afternoon tea was supplied for them. No other inducement was supplied.

In preparing for the focus group meeting, the participants were asked to read the Standards with four categories in mind, and to note these on their copy of the Standards:

a Standard OK i.e., no change needed to this standard. This standard substantially describes accomplished teaching in a New Zealand context

b Amendment desirable note what needs amending, and why

c Delete i.e., this standard is irrelevant to New Zealand

d Missing there is something missing that needs to be included.

For each part of the NBPTS AYA/Mathematics Standards document, the unit of discussion was each individual paragraph of the document, represented by a hard return in the text file. The document was systematically numbered to include the headings, sub-headings, paragraphs and footnotes as units of text. These units were numbered from 1 to 236.

Units 1 to 101 contained the generic material common to all of the NBPTS Standards documents (the preface, an outline of the NBPTS and its core philosophy, the certification process, Standards and assessment development, an introduction that specified the format of the Standards), and an overview of the eleven Standards for AYA/Mathematics. The exposition of the specific AYA/Mathematics Standards, which are assessed by the National Board, commenced at Unit 102 and continued to Unit 231. The final five units (232-236) were not relevant to the discussion as they listed the members of the Standards committee and acknowledgements. Units 102 to 231 provide a description of the teacher behaviours that could be converted into items for the student evaluation instrument, and therefore were the focus of the discussions. However, participants were asked to read all of the units in the document to provide background information for the discussion.

At the beginning of each focus group, participants were invited to make a brief opening comment about the Standards and their applicability to the New Zealand context. This was followed by a close reading of each unit, making particular note of those units that the participants felt deserved comment or amplification. Units of text that were accepted "as is", were passed over without comment and the discussion moved forward. More detailed comments were made of the units where participants felt that modifications were necessary either by addition or deletion. The focus group sessions finished with the opportunity for participants to make a closing comment to draw the threads together. In the conduct of each focus group, the researcher's role was to adopt an "unobtrusive chameleon-like quality" (Karger, 1987, p. 54) and to allow the discussion to flow with minimal interruption and to keep the momentum moving forward.

The meetings were audio taped, and field notes were also made during the discussion. The tapes were professionally transcribed, and the transcriptions reviewed against the tapes to ensure their accuracy. In addition, the copies annotated by the teachers were collected, and their notes reviewed for congruence with the recorded/transcribed text and the field notes.

Data analysis

The text was analysed in 3 ways - first, to determine whether each unit of analysis should be included or excluded from a New Zealand version of the Standards suitable for item development in New Zealand; second, to determine any changes that may be required to each unit of analysis; and third, to search for any other subtext that revealed the views of the focus group participants towards the Standards, and their use in the proposed study.

Kaplan (1943, p. 230) defined content analysis as the "statistical semantics of … discourse", or as a "research technique for the objective, systematic and quantitative description of the manifest content of communication". This systematic description of the communication enables researchers to answer questions of interest. This view is supported in the literature (Carney, 1969; Cartwright, 1953; Paisley, 1969), although Carney (1972) noted that content analysis is not necessarily limited to frequency counts, with "pattern-fitting" and other methods of data recording also practised. Three components are desired when using content analysis – *objectivity*; *system*; and, *generality*. Objectivity refers to a set of "explicitly formulated rules which will enable two or more persons to obtain the same results from the same document" (Holsti, Loomba, & North, 1968, p. 598). This ensures that any pre-conceptions or biases the researcher may have are removed, and that there is no contamination of the results, analyses and inferences that may be made with the data. As with all research in the interpretative paradigm, the use of other data for triangulation assists in reducing the confounding influence of subjectivity. System refers to the process of ensuring that the text is systematically analysed and accounted for, again removing the possibility that only material that supports the researchers hypotheses is selected. This does not mean that all of the document(s) have to be analysed, but that there has to be a systematic way of including or excluding content or categories. Finally, generality means that the

findings have to have some relevance to theoretical underpinnings, for without this they are of little value.

Numerical data can be generated by a technique known as contingency analysis, which marks the presence or absence of a particular attribute in the content, and it was this approach that was adopted for this study.

As this form of content analysis is used for drawing inferences based on the analysed text, issues of reliability and validity are quintessentially important. Issues of validity are inextricably related to the design of the study, the dependability of the interpretation, and to reliability. Reliability is a necessary but not sufficient condition for validity. Modern interpretations of validity require that evidence be assembled to support the interpretations that will be made from the data in the specific study, and these interpretations may not be valid in another setting or context. This includes evidence of careful construction of the study, appropriate scoring/coding, adequate score reliability, accurate recording, and careful attention to issues of equity and fairness (after American Educational Research Association et al., 2002, p. 17). These issues are clearly laid out in the following description of this study.

To conduct this contingency analysis, there were four main steps: Identify the coding categories prior to searching for these attributes in the text; select the sample to be coded, and the unit of analysis; count or systematically log the frequency with which the categories occur in the text; and interpret the results of the analysis (Ezzy, 2002, p. 84).

Research questions

With reference back to the Focus Group Protocols, the questions that this study was intended to answer were:

> Is there a consensus amongst the focus group participants about whether the AYA/Mathematics Standards can be generally applied in New Zealand?
>
> What modifications would be required to the Standards to make them more suited to teachers in New Zealand?

Are there any issues regarding the proposed use of the Standards that should be considered?

To answer the first of these questions, participants' comments were coded using simple contingency analysis – they either felt that the Standards were applicable or they did not. The opening gambit for each focus group was for the participants to state what they believed about the applicability of the Standards to the New Zealand context. Statements in favour of or opposed to the applicability of each paragraph unit were scored and recorded.

The coding categories for the first question were designed to "reflect the purposes of the research, be exhaustive, be mutually exclusive, independent and be derived from a single classification principle" Holsti (1969, p. 95). Participants in their opening and concluding statements addressed the issue of applicability, and these were coded as: agreement (without qualification); agreement (with qualification); disagreement (with explanatory comments); and, rejection of this unit.

Where qualifications were made, these were recorded and revisited in the third iteration below. A default coding of "agreement without qualification" was taken for any unit that was passed over in discussion. To answer the second question, notes were made about suggested modifications and whether participants agreed or disagreed with the suggestion.

For the second question, the analysis first looked for statements that suggested changes to the Standards, and then for what these suggested changes were. Following that a second iteration was made to gauge the feelings of the participants to the Standards, the merit of the particular standard under discussion, and whether or not they felt they could meet the Standards, and what it would take to do so.

Finally, to address the third research question, a third iteration looked for broad categories and emerging themes (for example, statements that were concerned about the assessment of teachers against these Standards) that are central to an analysis of the research questions. The transcription, participant copies of the Standards and field notes provided the evidence to address this third question.

Results and discussion

Acceptability of the Standards

Among the participants, there was general consensus that the Standards (with the modifications suggested during discussion) would make a suitable framework for professional standards in mathematics teaching in New Zealand. As with any consensus, there was variation of opinion within the groups, but the teachers involved felt that accomplished teaching in New Zealand would be adequately described in these terms. Table 2 show the percentage of participants and statements/units that addressed this issue.

Table 2 Applicability of Standards classified by number and percent of participants and analysed units

	Number of participants (%)		Number of analysed units (%)	
Applicable without qualification.	1	(8.3%)	100	(77.5%)
Applicable with qualification.	10	(83.3%)	24	(18.6%)
Not applicable (but with comment).	1	(8.3%)		
Not applicable.			5	(3.9%)
Total	12	(100%)	129	(100%)

One teacher was happy to adopt these Standards as fully descriptive of accomplished teaching in New Zealand. With one exception, the remaining teachers agreed that a modified version of the Standards could be used in New Zealand.

One of the teachers felt that the NBPTS AYA/Mathematics Standards could not be applied in New Zealand, and it was clear that the stumbling block for this teacher was not the Standards *per se*, but the planned use of them, particularly for the purpose of developing a student evaluation instrument. This objection centred on whether students

could be asked to evaluate teachers on all of the material contained in the Standards (specific reference was made to whether students were competent to evaluate a teachers contribution to the professional community, the teachers mathematical knowledge and expertise), and finally the validity of any student evaluation of teaching performance. The focus group protocols did not provide an opportunity to fully respond to these concerns, but these concerns were followed up at a later date with this teacher in line with the literature reviewed in Chapter 2.

In addressing the applicability of the Standards, several themes emerged from the discussions.

1. The Standards were a philosophical statement, and a somewhat idealistic goal. This theme was almost universally expressed. From focus group one:

> **G**: It seems to me like it is like Mom and apple pie, or king and country – how could you argue against it … the ideal. There is very little here in principle you could argue against.
>
> **Ca**: I certainly agree that it's aimed at the ideal, and there is not much you can argue with.
>
> **L**: But I do have my reservations about what is an ideal and what is reality in a real world … do we set a desirable standard, and then we have like the general standards and have excellence above it or do we just have the standard excellence.

and from focus group two:

> **M**: I thought it was very comprehensive but also very idealist, and you have to wonder whether anyone can really measure up to these Standards.
>
> **J**: I thought that it was really just as you've said, this is looking at an ideal and I couldn't take exception to any of it, but I also, it seemed to me to be rather unrealistic, unrealistically ideal.

2. The Standards may be unachievable. This is closely allied to the previous point. If the Standards are idealistic, then they are going to seem unachievable, and for many teachers this puts them beyond their expectations.

> **T**: And if we think of standards as being something that is achievable and maybe measurable then I thought it was quite heady and often not clear.
>
> **L**: But I do have my reservations about what is an ideal and what is reality in a real world … do we set a desirable standard, and then we have like the general standards and have excellence above it or do we just have the standard excellence.
>
> **P**: It's more of a reminder, if somebody said to me, tomorrow I'm going to check you on this, there's no way that, because it's not that I don't do it, I don't do it all the time, or I might have done it a few years ago and I've forgotten about it and moved on …

The idealistic and unattainable qualities that these teachers refer to in these first two themes are reminiscent of what one teacher has called the Mother Theresa Charter – "saintly qualities, which for anyone, would be an unrealistic counsel to perfection" (MacBeath, 1999, p. 60). The teachers in the focus groups were reflecting on what seemed like the superhuman nature of the Standards, and how they felt humbled to think of themselves as being considered worthy of inclusion in this discussion – they were after all, selected on the basis of their perceived high quality teaching. However, good teachers do not exist in a vacuum, and the Standards have to describe a reality that is "out there". P took the view that the Standards are more of a reminder, and that teachers do not behave like this all day, every day. The Standards are definitely attainable and, to date, 1430 teachers in the USA have achieved NBCT status in AYA/Mathematics. A comment that is frequently heard from successful NBCT candidates is the pleasure they had in achieving what they thought to be unattainable and that the certification process is the best professional development experience they have had (Bohen, 2001; Clehouse, 2000;

Croshaw, 1999; Frazier, 1999; Haynes, 1995; Iovacchini, 1998; Rotberg, Futrell, & Lieberman, 1998; G. A. Taylor, 2000; L. M. Wilson, 1998).

3. The Standards are wordy and needed editing down. L, who captured the essence of what many of the other participants also said, summed it up in this way:

> **L**: Very wordy … at first I thought it was just because they were saying in twenty words instead of six … but in fact it was also very repetitive and I don't think it is manageable in this size … maybe a one page introduction and each standard on one page …… get the lesson out … asking myself is it too prescriptive and too detailed? No, its not too detailed it just keeps repeating itself all the time that was its major problem and we all found it very heavy to read and yet we're an elite group you picked out and it wouldn't [be] manageable for the majority of teachers …

In developing the items for the student evaluation instrument, the repetitive nature of the Standards was also noted. In some cases, almost identical items were drafted from two or three different units of the Standards. While the criticism and advice were apposite, the purpose of the exercise was to retain the integrity of the Standards by making as few amendments as possible. If a team of mathematics specialists were to engage in writing a similar set of Standards for highly accomplished New Zealand teachers, this advice should be heeded and acted upon.

4. There is a need to align the knowledge of the curriculum section of the Standards to the Mathematics in the New Zealand Curriculum document.

> **Co**: It was just the maths part really.

> **J**: I couldn't take exception to any of it, the only things that I wanted to comment on were some of the mathematical content that was detailed, that I felt wasn't relevant to the curricula that we are currently teaching, or things that I never covered when I was at University getting my maths degree.

The curricula statements that guide mathematics teaching in the USA and New Zealand have many similarities, but also some significant differences. The participants felt that the content strands of the New Zealand curriculum should be substituted for the sections of the Standards that address core mathematical knowledge (Units 129-135). In particular, almost all of the section on Discrete Mathematics (Unit 133) should be deleted, and descriptions of accomplished teaching in Number and Measurement added in its stead. Where the taught content was the same, the existing Standard should remain intact. As an aside, it was also noted that there is a fundamental difference in the way the mathematics curriculum is delivered in each country – in the USA, students generally take each strand of the curriculum as distinct subjects (for example, Algebra II, Geometry I, Pre-Calculus) taught by different teachers, whereas New Zealand students have a comprehensive course that integrates all strands of the curriculum, taught by the one teacher.

5. Concern about using the Standards as the foundation document for a student evaluation of teaching instrument.

> **Co**: The approach you showed us of extracting the guts from it and re aligning it sounds like the right way to go to approach something that's usable. ***My feeling is even stronger now that it's not usable.***

The last sentence was verbally emphasised, and the speaker went on to outline how, in preparing for the focus group, he had discussed the Standards with a colleague, and the colleague was daunted by the length and detail of them.

Similar sentiments were expressed in the second focus group by another speaker, who felt that some of the Standards described aspects of teaching that student could not be expected to have a realistic knowledge of.

> **J**: … and the other question that kept coming to me all the time was how would the students know that the teacher had this knowledge, I'm thinking about the purpose for which you, I think you have in mind for this. How would students know the breadth and depth of some of these things? But I couldn't take exception to any of it …

and again later in the discussion, when considering the role of teachers in family and local community as well as the professional community (units 209 to 225), the participants engaged in a dialogue with the researcher regarding this very issue.

> **V**: I just feel that something should be said earlier on. This is going to be for student's assessment of teachers? Students assess teachers by this criteria, is that correct?
>
> **EI**: Well, that's where I'm going.
>
> **V**: Exactly what J said earlier on. How would students know who belongs to what organisation, what work they do, outside of this classroom, etc., etc., etc. But then this isn't relevant is it? This isn't relevant if this is about student assessment of teachers, this is not relevant
>
> **M**: I suppose what V's saying is, that at the end of the day, the amount of this very substantial and thick document that is going to be relevant to the students, in which the students are even capable of commenting is in fact very small. You don't agree with that?
>
> **EI**: I don't think it would be very small, but it certainly won't be the full set of things which would be in here.
>
> **M**: Yeah couldn't be.
>
> **EI**: No. But sometimes there are ways of asking the question which might get the response to something which you feel sceptical about.

The points made by these latter teachers are entirely pertinent: (1) students should not be asked to evaluate the breadth and depth of a teacher's mathematical content knowledge, as this is something they are not competent to judge; and, (2) students are not well placed to comment on the professional involvement and contribution of a teacher outside of the classroom. The SEAT-M instrument to be developed in this thesis does not contain any items related to these areas. However, the first teacher was implacably opposed to the use of student evaluation instruments. As noted in the review of SETs in Chapter 2, the use of student evaluation instruments in high schools classrooms is relatively rare. Whenever teacher evaluation in this form is introduced, teachers feel

threatened for several reasons – the supposed immaturity and capriciousness of their students, the lack of perspective that the students will have, the reliability and validity of their judgments, and especially at school level, the differing relationship that teachers have with their students (they have a greater disciplinary role than do tertiary teachers) and the impact that this can have on teacher-student relationships and therefore on possible reprisals through the "ballot box" of an evaluation form. In spite of the evidence to the contrary, teachers tend to fear the worst, as they perhaps realise that the students are the very people who know their teaching best, and that they may have to confront an unpleasant reality when they receive their student ratings.

6. That not all teachers would make suitable candidates for certification as exemplary teachers. In particular, many of the Standards in the documents require a teacher to develop and have a degree of confidence in their own ability (and confidence in their students) for them to be considered eligible for certification. An interesting conversation occurred around this point when considering teachers who were prepared to take risks in their teaching and provide a safe arena/environment for ideas to be expressed by their students (Units 163 to 176).

   J: Can you see under 166 though, can you see somebody who's prepared to be a learner alongside the kids and to say, I don't actually know but let's work it out together. There's always the fear if you do that too often that the students, you can do it sometimes, but if you do it too often and you do it from the beginning the students can lose confidence in you. I think you can only do that once the students trust you personally, that's my understanding on that one. You can do it too soon.

   M: But that's different to saying that you don't know. I think there has to be a degree of trust between the student and the teacher and they have to feel confident, that sure you may not have come this problem before, but you're in control of the situation and even if you can't answer it, you'll find out how to answer it, maybe along side them, maybe you'll do it as a group activity, but I think it's a very risky situation if you start letting the students think in fact that they know more than you too soon, that's just my own feeling.

and later:

> **M**: But what I mean is, the teacher who does that has to feel themselves, confident in their own knowledge and their own ability to be able to handle that situation, it's not something that a teacher who is not secure in their subject knowledge is able to do readily I don't think.
>
> **EI**: Well, are we talking about a situation where you … walk into one and don't know it, or just finding that sometimes because of the way somebody asks a question … "hang on, I'm going to have to think about that one first, I don't have the answer immediately"?
>
> **M**: Sure but we all know that if a teacher lacks confidence then they tend to be book driven and they tend to be task driven, I'm not saying it's not good, I think it is good, I'm just saying that this particular skill requires teachers who are confident, both in their knowledge of the subject area and in their ability to relate to their students. Whereas some of the other people who are maybe not as experienced could …
>
> **EI**: The classic thing … perhaps the introduction of new technology … where we've all had to learn in one way or another and often the student knowledge may well have been a little advanced from where you are … and somebody says "Oh, you could have done it that way"?
>
> **J**: I think the thing that I'm thinking about in terms of 166 is us modelling mathematical problem solving, because we have got such confidence in the mathematics we can actually do everything and we can actually do it quite quickly and it's really easy for us to give our students the impression that maths is quick and simple and it just boom, boom, boom, but in fact some of it took hundreds of years to develop and a great deal of time and a lot of thought and I'm wondering if that's the section, I can't remember now, where we modelled that sometimes we're problem solvers, that I'm going to go home and worry away at this, doesn't mean I don't know what I'm doing but I'm actually showing you that perseverance and some of those qualities that we're trying to teach our students are useful for us too and I'm taking on board what you're saying about doing it ….

> **M:** Basically what I'm saying is if you were going to have a tick list of this is what I have to do to be an excellent teacher then I would think this would be towards the end of that tick list. I think there are other objectives which are easier to achieve, I think this is one of the harder, that's one of the harder objectives for a teacher to achieve…

National Board certification is not intended for all teachers. This is a voluntary advanced certification process, for teachers with a minimum of three years teaching experience. Teachers will only apply if they feel confident in their own ability, and need to have ability to demonstrate (and capture on video) highly accomplished and effective teaching and learning experiences (Covert, 2000; Iovacchini, 1998). In addition, they need to have the ability to successfully engage in the discourse of intense self-reflection and analysis required to fulfil the requirements of the portfolio and assessment centre exercises (Burroughs, 2001; Burroughs et al., 2000; King, 1994). Even then, the chances of success are not substantial. By November 2003, approximately 32,000 teachers had been certificated out of 65,000 applicants, while in North Carolina, one of the states with the greatest number of NBCTs, the annual certification rate hovers between 41 and 52 percent (Goldhaber & Anthony, 2004; Goldhaber, Perry, & Anthony, 2003). In the mind of the last focus group speaker (M), teacher confidence and the ability to provide a safe cognitive/intellectual environment for students (and the teacher) are amongst the harder requirements to achieve, such that they act as the ultimate gatekeepers to a teacher's decision regarding preparation for certification.

7. That there may be some fundamental differences regarding assessment and the teaching environment between the USA and NZ. The introduction of a new national assessment system in New Zealand (Unit Standards and Achievement 2001) conflicts with some of the expectations expressed in the NBPTS Standards. This was clearly highlighted in a interplay between several of the focus group members which centred on whether exemplary teachers allow students to be autonomous learners, and to arrive at and present their solutions to problems in different ways (paragraphs 191 to 194):

J: They come to us at high school as autonomous learners but somehow this gets changed.

M: Hormones have got something to do with that.

V: They go through rebirth.

J: It gets quite, whether we encourage them or not, when we get to year 13 and they're preparing for Bursary, it is like the tunnel's getting narrower and narrower and there's only one way to prepare for Bursary and I'm going to do that.

P: Yeah, but going to the seminars with the School C markers, I think they actually have to teach them to jump through certain hoops, if they don't write v-o-l instead of v they get penalised.

General: That's right.

P: So they've got to learn to jump through hoops and you've got to be as narrow as that. …

M: There are two things we're doing, we're trying to teach them mathematics, but the bottom line is we're also trying to teach them to pass exams and we have to meld those two together, so sometimes we do have to say to them, well sorry folks I know you could do this on your graphic calculator, but for the exam you've got to play the game and you've got to do this

M: Of course, unit standards … unit standards don't give the students the freedom to choose their method.

V: That's one of my beefs … the actual practice is …

M: It seems strange to me that in fact we seem to be making a backward step I think in assessment, we're removing the right of the student to answer questions in the form that they want.

J: Excepting that, you can phrase them in a certain way that you're asking them to show that they can provide a variety of methods, so you could give them questions as long as they covered the variety you ask, they wouldn't necessarily have to use Pythagoras or trig …

M: But they could end up with no unit standards if they don't use the right method, in the right place.

P: Yeah, it depends on how you write the assessment, but you can get, we had trouble with the trig one particularly, because they can do it by a

number of methods, we just have to go through and match up which ones they've used. We just write a note, make sure you show us at least this, this and this. Some of it can be very …

M: It seems to be that the assessments that we're being asked to give them now actually removes this freedom, that excellent teachers are being asked to give them…

The teachers clearly value the freedom to reach the solution to a problem through multiple pathways, but feel hampered by certain very specific assessment requirements. The distorting effects of high stakes assessment on teaching practices have been well documented (see, for instance, Amrein & Berliner, 2002). Teachers are in a dilemma when considering teaching the full mathematics curriculum and meeting the restrictive prescriptions of a high stakes national assessment, and this is reflected in these comments. Although there are no national assessment programmes in the USA, this dilemma is not unique to New Zealand. Statewide testing requirements, plus the "adequate yearly progress" of the *No Child Left Behind Act 2001* in the USA, leave American teachers in the same position. The classroom teacher has to decide whether they can take their class on rich mathematical journeys, following different paths from those described by the examination prescription, and still ensure that their students are prepared adequately for the assessment. In mathematical problem solving, a good, 'elegant' solution is also noted for its economy of exposition, but an elegant solution is not the only way of achieving a successful outcome, and accomplished teachers are capable of managing multiple goals and achieving multiple outcomes. Multiplicity is one of the consistent themes in the Standards document ("… instructional repertoire multiple paths to the subjects" Unit 15; "…employ multiple methods for ensuring student growth" Unit 20; "… reason and take multiple perspectives to be creative and take risks" Unit 22; and "… multiple solutions can be both useful and interesting" Unit 191). In addition, the assessment system used to certify NBCTs has to "accommodate multiple profiles of excellence across the Standards" (Unit 64).

At the time of the focus groups, schools were engaged in extensive professional development work regarding the introduction of NCEA (or Achievement 2001, as it was then called). There was heightened sensitivity to issues surrounding the introduction of a new national achievement based assessment system, which was replacing long-standing norm-referenced external examinations. Several schools had already experimented with Unit Standards and found that some aspects of the assessment criteria were very pernickety. As one of the participants noted, the careful construction of assessment tasks can help to overcome some of the more detailed and frustrating requirements regarding the way in which students can respond.

8. The place of technology and things 'new' in mathematics.

> **V**: The whole ICT thing, the whole you know ICT into the curriculum, that is forcing … we have a requirement at school that every curriculum level in year nine and ten must have an ICT component to it, and for some people that's a big movement. Just even thinking about it, let alone … I would think the mathematics teachers would be more technically online, but some of mine I'm still struggling to get them in front of a computer, let alone … sure they can write a letter or whatever, but get them using spreadsheets and stuff. Which you'd think that a mathematician would intuitively be able to do.
>
> **M**: I'd just like to comment on 208, the sentence that says they have respect for both new and old thinking about mathematics teaching. I actually must be getting quite old, cause I actually find a lot of the so called new is in fact old and I seem to get a lot of stuff presented to me now as new, which I seemed to have come across 30 years ago and it's simply being recycled. So I'm beginning to feel like the person who thinks … policemen are getting younger, I feel as if I'm definitely getting older.

J: I think it's very hard to say what is old and what is new thinking, because as I say I find a lot of the so-called new material and new thinking is in fact just rehashed old.

V: It goes out of vogue and it comes back in again and goes out…it's like your wardrobe**.**

There are two threads here – new technology (for instance, computers and graphic calculators), and changing fashions in what we should teach as part of the curriculum. Neither was intended to act as a prompt to modify the Standards, but rather to acknowledge the changing nature of the mathematics teachers' world. Change is accepted as inevitable, and highly accomplished mathematics teachers need to be on top of this aspect of their teaching as much as they are on top of their core mathematical content knowledge. New mathematics has been and gone, and it is easy for teachers to develop "techno-cynicism" just as easily as they could become "techno-zealots" (Boshier & Onn, 2000) in order to protect themselves from constant change in technology, as well as the turmoil they were facing over the introduction of a new national assessment system.

9. That there are important differences between the USA and New Zealand in the expectations of schools regarding the interaction between the teacher and the students' family and support systems.

P: One would have to say that part way down, 211, about homework assignments. Families are not necessarily impressed about that, teachers often design homework assignments involving activities that will encourage family discussion. I'm not sure that that is necessarily seen as a positive by families.

M: I'm mindful that all students may not have families to assist them. Mindful that all students do not have families that want to assist them, or that it's within the capabilities to assist them and sometimes I think that parents have a very negative influence on children's attitudes to mathematics. I sometimes just want to scream when another parent sits

down in front of me and says I was no good at maths, when I was at school.

P: And certainly the last sentence, 'recognising the potential benefits of learning, they invite family members and members to visit the classroom', that just doesn't happen in New Zealand secondary schools. I might be wrong in making that statement, but I don't ever … well it sounds like a nice ideal and it would be wonderful, I think they have more fear and that would just increase the fear.

J: I think that people should be able to have other people in the classroom and I think if they're an excellent teacher they should be comfortable with that, but I can't imagine families coming in and being part of that.

V: Another thing is about 211, not only must we be mindful that students don't have parents to help them for a variety of reasons but also some students have tutors that will do the work for them, so that's not really a valid thing of giving them work to do and that includes family members, it's not the family, it's the tutor that does it and they don't necessarily learn anything by it.

P: There's also, some of our girls are at home and they are the oldest or one of the oldest girls in a large family and they have responsibilities, and not even the time for homework …it's different

V: 213's okay though because it's all about what schools do really isn't it?

J: Welcoming environment, working with families, supporting maths events, involving, careers. That sort of thing, that's not the same issue I don't think.

The crucial role of the teacher in communicating with families and the wider community reflects one of the greatest challenges of the post-modern age, according to Hargreaves (2000). The implications go beyond the concerns raised by these participants. While liaison with families has always been a part of the teachers' duties, this was often restricted to traditional activities such as parent-teacher interviews and volunteer help at school, but the current drive for more inclusive, open schools places additional demands, emotional and social, on teachers. Parents as first teachers, and then as active supporters of later learning at school, place the parent alongside the teacher in mapping out the learning programme for their child. Teachers, however, prefer to maintain their autonomy and control over the curriculum and teaching methods with parents in a supportive but subsidiary role (Crozier, 1999; Todd & Higgins, 1998). Matters of assessment, curriculum, teaching methods and discipline are all contentious, such that teachers act defensively when parents are in the class, feeling threatened and under surveillance. The presence of parents strips away the mystique of professionalism. Teachers can feel just as uncertain in the presence of parents, as parents experience in their role as parents. However, the National Board contention is that parents are an invaluable source of information about their own children, and teachers need to move beyond a position of professional superiority to a genuine partnership and alliance for better learning and teaching. Not only do accomplished teachers seek information about students "strengths, interests, dispositions, habits and home life" (National Board for Professional Teaching Standards, 1996, p. 35), but they provide information about school programmes, assignments, study skills, and the services that the school and community has to offer. The information channels work both ways. As a consequence, Hargreaves argues, through the active support of an informed parent body and the wider community, teachers are in a better position to defend political attacks and argue for changes that will work to enhance teaching and learning.

Modifications to the Standards

In examining the text for statements regarding modifications to the Standards, it was noticeable how few suggestions were made concerning this, apart from matters relating to the content of the mathematics curriculum.

The most common suggestions for alterations occurred in relation to the core mathematical knowledge. At first, it was noted that the entire Number and Measurement strands were missing, and that almost all of the content described in Discrete Mathematics (Unit 133) was not taught in New Zealand. It was suggested that the names of the MiNZC curriculum strands – algebra, number, statistics, geometry (including trigonometry), and measurement (including calculus) - replace the detail of the units concerning the core mathematical knowledge (Units 129-135). The sixth MiNZC strand, mathematical processes, consists of three elements (problem solving, communicating mathematical ideas, and developing reasoning and logic), each of which is well defined in the Standards. These sections of the Standards (Units 136 to 144) were left intact.

The second area of concern was the observation that the Standards omitted one important component of teaching – classroom management and discipline especially in relation to teaching in diverse classrooms. Good classroom management is seen as a vital component of the training that we undertake as teachers, and without it, teachers are unable to teach and students unable to learn. We can see this concern expressed in the following statement when discussing Standard VI: Learning Environments (Units 177 to 183):

> **Pe**: As a teacher told me years ago, if you have an enthusiastic programme you don't need to concentrate on discipline. Discipline looks after itself. Now, with such a diverse range of students, across the diverse social milieu I think we do need to be concentrating on strategies to develop a whole range of students. I don't see that … I just feel it hasn't really answered the diverse range of things that are in our classroom today in New Zealand.

This was supported in the annotated copies, where L had added the words "and disciplinarian" to the different roles that the teacher assumes (Unit 168) under Standard V: The Art of Teaching.

The determination was that a simple statement about the need for good classroom management and disciplined inquiry should be included (without elaboration) in Unit

168 which refers to the roles that teachers assume to accomplish the complex task of teaching.

Apart from the curriculum amendments and the matter of classroom management and discipline, participants suggested a total of 27 different changes in units 102 to 231, all of which involved minor amendments to the wording. One of these changes was repeated four times, to amend the same text (deletion of the words "state" and district", and replacement by the word "national" to describe curricula and governance). These minimal changes could have been a result of the protocols for the focus group, which asked for changes to be kept to a minimum so that the focus was on the big ideas in the document.

For instance, in Unit 112, the NBPTS standard reads:

> They recognize and work to overcome barriers which can prevent women, minorities, or any students, including those with disabilities, from achieving success in mathematics.

The focus group felt that the intent of this paragraph would better suit teaching in New Zealand if, in keeping with the terminology used here and to be more inclusive by removing the specified groups, it simply read:

> They recognise and work to overcome barriers to learning which can prevent students from achieving success in mathematics.

The term "barriers to learning" was in current use, as a consequence of the National Education Guidelines (1989) and Education Review Office requirements for schools, and the way in which they catered for and reported the achievement of diverse groups in their school.

A second example of the minor nature of these changes, and the rationale for it is found in Unit 107 where the word "like" was replaced by "care deeply about" such that the sentence now reads, "They care deeply about their students and find mathematics and

the teaching of it a lively and enjoyable experience". The term "like" was regarded as too general, whereas the replacement term conveyed a sense of professional care and duty.

Concluding Statement

In summary, the two focus groups were agreed that the NBPTS AYA/Mathematics Standards are comprehensive, demanding, somewhat idealistic and theoretical, frequently repetitious, and express some important differences, especially in the expectations about teacher-parent interaction. It was also noted that some aspects of the Standards are not appropriate material for students to evaluate their teachers. Most importantly, however, the members of the focus groups felt that the Standards are applicable to mathematics teachers in New Zealand, with relatively few minor changes, including the need to align the Standards with the Mathematics in the New Zealand Curriculum document.

This has major implications for New Zealand as it too seeks to define the concept of a highly accomplished teacher. The master teacher concept has been on the table since at least the 1960's, but has not gained any currency. In the US, the two major teacher associations (the NEA and AFT) have been involved in the NBPTS from its inception, and one of the requirements of the Board is that classroom teachers must comprise at least 50% of all decision making panels. This has lead to accusations of provider capture (Ballou & Podgursky, 1998b; Finn & Wilcox, 1999; Wilcox, 1999) and the purported deleterious effects that this has on the credibility of National Board certification. While these critics support the notion of identifying and rewarding highly accomplished teachers, they want only one criterion for certification – demonstrable, measurable improvements in student learning. There is emerging evidence that measurably improved student learning is occurring in NBCT classes compared to non-NBCT classes (Goldhaber & Anthony, 2004; Vandervoort et al., 2004), but there is a need for replication of studies such as this to increase the validity of the claim that NBCTs improve student learning. In the USA, the NEA and AFT have been intimately involved in the development of the NBPTS Standards and the assessment processes that identify NBCTs, and there is widespread (although not universal) acceptance among teachers of the work of the Board and the rewards for those teachers who can

demonstrate their expertise. To gain acceptance with teachers in New Zealand, the active involvement of the New Zealand Educational Institute (NZEI) and the New Zealand Post Primary Teachers Association (NZPPTA) will be vital. Without the teacher unions in the fold, the development and use of Standards for highly accomplished teaching will not become reality.

The teachers in the focus groups seemed unconcerned about the political and industrial ramifications of the highly accomplished teacher model, responding more critically to the proposed study methodology (the use of SETs) than they did to the proposed model of accomplished teaching. Given the overall positive response in which they responded to the Standards, a model similar to that of the NBPTS for highly accomplished mathematics teachers in Years 11-13 would stand up well in New Zealand.

Chapter Four: Study Two

The Students Evaluating Accomplished Teaching – Mathematics (SEAT-M) instrument is designed to assess highly accomplished mathematics teaching from a student perspective. While there are instruments designed to measure student attitudes to mathematics (for example, Fennema & Sherman, 1976; Holly, 1971), no student instruments exist specifically to identify those characteristics that mark out highly accomplished mathematics teachers. This chapter will describe the process of converting the amended NBPTS AYA/Mathematics Standards into 470 questionnaire items, the two trials in New Zealand schools that lead to the final selection of 51 items from the pool of 191 items, and the way in which Item Response Theory and Classical Test Theory (particularly factor analysis) were used to make the final selection for the SEAT-M. The purpose of the analysis of these Forms was to select an optimal set of items for inclusion in the final questionnaire form, to represent the sub domains and the overall domain of exemplary teaching as specified in the NBPTS AYA/Mathematics Standards.

Instrument development

Marsh and Hocevar (1991, p. 11) suggested that the general procedure for the careful design of a SET instrument should follow these steps: the development of a large pool of items (from literature reviews, existing instruments, interviews with students and teachers), trials involving these items with students providing feedback about the items on the pilot instruments, and consideration of the psychometric qualities of the items during several revisions. Berk (1979) went further and asserted that the framework for the development of SET instruments must have the specification of the domain of interest as a crucial first step. In addition, Berk outlined the classical test statistics that are typically used. As will be demonstrated below, analyses using Item Response Theory now enhance the ability of the test developer to analyse items and determine their suitability for inclusion in the proposed instrument.

The major step added before those outlined by Marsh and Hocevar is that the items can be developed by referencing to an established set of principles for accomplished teachers, thus providing a method for ascertaining the validity of the final set of items. As noted in the previous chapter, the amended NBPTS AYA Mathematics Standards were adopted as the appropriate description of the domain of accomplished mathematics teaching in secondary schools. The focus groups of accomplished New Zealand teachers had found these Standards to be suitable as a foundation for defining accomplished mathematics teaching. From this foundation, the SET instrument would be constructed, following the suggested processes above.

From the amended NBPTS AYA Mathematics Standards, 470 statements were drafted to closely reflect both the wording and intent of the Standards. The statements were numbered with reference to the paragraph of the Standard they referred to, in order to keep track of the origin and development of the statement. Thus, 182/5 was the fifth statement drafted from the text of paragraph 182. Wherever possible the wording of the items was taken directly from the Standards, though some alternative wording changes were made to use language that might be more accessible to students, rather than using the language of teachers. For example, Item 182/9 uses the words of the Standards but Item 182/10 adopts a more colloquial wording to address the same behaviour. The following two examples illustrate the way in which paragraphs from the Standard were translated into draft statements.

Table 3 Sample of paragraph from Standards and drafted statements

	NBPTS AYA Mathematics paragraph	Drafted statements *My mathematics teacher …*
119	In order to gauge their students' strengths, needs and interests, teachers insightfully observe and listen to their students in whatever setting students use to express themselves, be it a formal classroom setting, an individual conference or informal conversation. These insights, including their ability to identify students with disabilities, exceptional needs or talents, enable	• Observes and listens to the class members in a variety of settings where students express themselves(119/1) • Identifies students who have exceptional talents in mathematics and helps/supports them (119/2) • Identifies students who have particular difficulty or need in

	teachers to adapt their practice to support all kinds of students. They work collaboratively with specialists, as necessary, and modify their plans and materials to support different kinds of student, including those whose primary language is not the language of instruction.	maths and helps/supports them (119/3) • Gives special help and support to students whose main language is not English (119/4) • Uses their knowledge about us to meet each of our needs (119/5) • Works collaboratively with other specialists to modify their teaching plans to support every student (119/6) • Identifies students who have disabilities and supports them in maths (119/3)
182	The creation and maintenance of such learning environments require skill and planning, a variety of instructional methods, flexibility, good judgment and discretion. Teachers, considering the needs, interests and working styles of their students and the mathematics they are studying, create a climate in which students learn to value mathematics and experience success in doing worthwhile mathematics. They lead by example and convey to students the delight that comes with command of a mathematical tool or principle. They continue the development of social skills through a combination of group and individual work and help students develop the ability to work both independently and collaboratively on mathematics.	• Creates and maintains a learning environment by being well planned (182/1) • Creates and maintains a learning environment by using a variety of methods in teaching (182/2) • Creates and maintains a learning environment by being flexible (182/3) • Creates and maintains a learning environment by displaying good judgment (182/4) • Creates and maintains a learning environment by displaying discretion (182/5) • is considerate of our needs/ interests/ working styles (182/6) • creates a climate where we learn to value maths (182/7) • creates a climate where we experience success in doing worthwhile maths (182/8) • leads by example (182/9) • practises what s/he preaches (182/10) • takes pleasure in having command of a mathematical tool (182/11) • helps to develop independent work habits (182/12)

	• helps to develop collaborative work habits (182/13)

This unpacking of the Standards was designed to convert the complex sentences of the Standards into statements that reflected a single teacher behaviour that could then be assessed by students without interference from other behaviours.

As noted by the teachers involved in the focus groups, there was considerable repetition and overlap in the Standards document and this was reflected in the drafted statements. Amalgamating repetitive statements and removing items referring to aspects that the NZ focus groups claimed were not relevant to the NZ context reduced the 470 items reduced to 191 items for trial. The process for doing this involved collecting statements with common threads from across the Standards and writing an item that reflected this commonality. In the following example (Table 4), the drafted statements are centred on NBPTS Core Proposition 1 (Teachers are committed to students and their learning), and describe the way in which the teacher takes a variety of steps to ensure that the students learn and succeed in mathematics. The original drafted statements are on the left, and the synthesised items are on the right with the statement numbers recorded afterwards, plus the questionnaire and questionnaire number they first appeared in. In this way, 32 statements became 9 usable items.

Table 4 Example of the synthesis of Standards statements to items for trial

Drafted statements *My mathematics teacher …*	Synthesised items *My mathematics teacher …*
• takes extra steps to ensure that students learn (107/5) • does everything possible to help us learn mathematics (107/6) • tries to help everyone even if they don't know the maths taught in previous years, both in and outside the classroom (111/1) • tries to help the students who are weak in maths to catch up (111/2) • works (helps) to get us back into the math'al mainstream (111/3) • gets parents, counsellors,	• takes extra steps to ensure that all students, (regardless of their ability) learn and achieve success in mathematics (107/5/6, 110/12/13, 111/1/2/7/8/9 - Item B54) • identifies and helps students with special needs or special abilities in mathematics and provides help for them (111/3, 119/2/3/4 – Item B25, modified slightly) • involves families, counsellors, administrators and others in the school and community to help and support students to learn and persevere in maths (111/4, 112/6, 119/6, 202/1/2, 210/1/3,

administrators and others in the school community to help a student to learn in maths (111/4) • enlists support from families and other school personnel to provide support and assistance when we/I are/am having difficulty in maths (112/6) • helps only the fast kids in the class (110/12) • ignores the students who are having difficulty with maths (110/13) • makes it possible for all students in our class to achieve success in maths (111/7) • works to make it possible for all students to achieve success in maths (111/8) • recognises the barriers to learning that prevent any student from achieving success in maths, and works to overcome them (111/9) • identifies students who are exceptionally good at mathematics and helps them (119/2) • identifies students who have particular difficulty in maths and helps them (119/3) • gives special help to students whose main language is not English (119/4) • works collaboratively with other specialists to modify their teaching plans to support every student (119/6) • involves our families in supporting our learning (202/1) • involves members of the community in supporting our learning (202/2) • Involves my family in supporting my education (210/1)	211/1/2/7 – Item B58, modified slightly)
• Plays a part in keeping the community up to date with what is happening in mathematics (210/2)	• Plays a part in keeping the community up to date with what is happening in mathematics (210/2 – Item B42)

- tries to involve the community in supporting mathematics teaching in the school (210/3)
- views our family as a partner in our learning, growth and development (211/1)
- gets our family to encourage us to persevere in maths (211/2)
- looks to our families for information about our strengths, interests, dispositions, habits and home life (211/3)
- keeps our family informed about the maths programme/ significance of test scores and grade/ consequences of taking or not taking certain courses/ reasons for group or class assignments/ benefits of planning for future education (211/4)
- has homework assignments that will encourage family discussion of school subjects (211/5)
- realises that not all students in the class have families who can assist them (211/6)
- works with our family to help us develop good study habits, complete homework, set goals and improve performance (211/7)
- invites our family to participate in the classroom (211/8)
- Provides support and encouragement when I am doing well (212/1)
- Makes a real effort to help me when I am not doing well (212/2)
- gets other people to help if I am not doing well in maths (212/3)

- Looks to my family for information about my strengths, interests … (211/3/7 – Item B55)

- Keeps my family informed about my progress in maths (211/4, 212/5 – Item B44)

[Not a NZ standard according to focus group]

- Realises that not all students in the class have families who can assist them (211/6 – Item B12)

[Not a NZ standard according to focus group]

- Provides support and encouragement to me (to all of the class) all of the time (212/1/2/3 – Item B23, modified slightly)

This process of synthesis continued for different clusters of statements, until 191 items were drafted. These 191 items were then mapped back on to the Standards document to ensure that there was complete coverage of all of the Standards.

The readability of the 191 items was assessed using the Flesch-Kincaid Grade level formula at 9.7. This indicates that the items were at an appropriate level for the intended student level of Grades 10-12 in the USA. Years 11-13 are the equivalent of these grades in New Zealand.

To minimise respondent fatigue and to ensure completion within approximately 40 minutes of class time, the items were divided into three different forms. The forms were not intended to be parallel forms. The grouping of items tended to follow the grouping of the NBPTS Standards, although the overlap between items written for two distinct parts of the Standards meant that this distinction was not always possible. There was only one item in common to the three forms – "*My mathematics teacher, compared with all other mathematics teachers I have had, is the best"*. This was the last item on each of the questionnaire forms. Global items such as this have been criticised when used for personnel decision-making (summative purposes) for asking the students to make comparisons with other teachers, or ask the student whether they would recommend the course to a friend with similar interests (Scriven, 1995). However, Peterson, Wahlquist and Bone (2000) argued that a single global item can be useful as a summative report, provided the global item can be shown to well represent the other items on the scale.

Certain elements in the NBPTS Standards for which students may not be expected to observe the teacher in a particular role - for example, Standard IX (paragraph 208) states that exemplary teachers keep abreast of the latest changes in mathematics and mathematical pedagogy by reading professional journals, attending conferences, and participating in professional organisations - were omitted from the questionnaire. It is not reasonable to assume that students can be expected to know this about their teacher.

In prefacing the Standards Format, the Board recognised that one of the essential tensions in translating a holistic picture of exemplary teaching into statements that describe this teaching is that the statements themselves will be discrete and atomistic. While this is inevitable for the purposes of recording the Standards, aspects of teaching were not to be seen as discrete activities that could be separately measured in the classroom. This tension is re-iterated in Paragraph 68 of the Standard, for example, reference is made to the difficulties inherent in listing discrete duties, such as managing

the classroom and designing learning activities, whereas the Standards endeavour to view teaching as a seamless activity in which the teacher makes dozens of decisions in which all of their knowledge, understandings and experiences are utilised. Feedback from the focus groups indicated that classroom management and maintaining student discipline had received virtually no attention in the Standards. To address this concern, five items were taken and adapted from an earlier survey instrument developed by the researcher (Irving, 1996). These items were all in Form C, but none of them were selected for the November or USA questionnaires. These five items were:

C3 focuses all of the students on their work
C21 motivates us to do our best work
C36 uses a variety of techniques to maintain control of the students in this class
C47 provides enough work to keep all students in the class working
C51 keeps the interest of all the students in the class

A six-point Likert type intensity scale was used. As the purpose of the instrument is to dependably discriminate teachers who are at the top of the scale, the anchors used have to discriminate more at the top than at the lower ends of the scale. For this reason, a positively packed response scale (G. T. L. Brown, 2004; Lam & Klockars, 1982) was adopted, with two points for disagreement and four for agreement. An additional reason for this choice is that research has shown that students typically respond in a positive way when evaluating their teachers (Bendig, 1952b; Centra, 1973a; Miklich, 1969; K. D. Peterson et al., 2000; W. R. Wilson, 1999). For both of these reasons, Bendig (1952a) had made just such an adjustment in a student rating scale by dropping a scale point at the negative end and adding one to the positive end of the scale. To achieve this discrimination, the wording of these anchor points was critical. The two endpoints were the commonly used "Strongly agree" and Strongly disagree", with four intermediate anchor points. These intermediate points were chosen to create an equal-appearing interval scale between the two endpoints, and were selected from an adverbial list of anchors with magnitude estimation values assigned to each anchor (Cliff, 1959). The following anchors were used for the questionnaires:

1 *Strongly disagree*
2 *Tend to disagree*
3 *Slightly agree*

4 *Somewhat agree*

5 *Usually agree*

6 *Strongly agree*

The students responding to the questionnaires were asked to indicate a degree of agreement or disagreement in their responses to each item, and no neutral category was provided. This avoided the possible ambiguities that arise with the interpretation of a neutral or not applicable position on the scale.

Trial One

Setting

Four schools were chosen from Greater Auckland, the largest metropolitan area in New Zealand. Sampling was a two stage process – firstly the selection of schools, and secondly, the selection of classes.

For Round One, four schools were selected from the 73 state or integrated high schools in the Auckland urban area. They were selected to give a cross section of schools though ultimately agreement to participate was the deciding factor. The socio-economic status (SES) decile ratings for the four schools were 3, 7, 8 and 10. Each state school has a decile rating assigned by the New Zealand Ministry of Education which is a composite socio-economic indicator based on census data for the residential areas of a random sample of students in the school. The decile scale ranges from 1 (low SES) to 10 (high SES).

The four schools agreed to participate anonymously, and have been given pseudonyms. The initial approach was made to the Principal to seek approval in principle. The Head of Mathematics Department was then approached and consent gained to survey students in mathematics classes in Years 11, 12 and 13. The Head of Department and mathematics teachers in each school determined which classes were surveyed, and arranged a timetable for the day on which the surveys were completed. The researcher had no role in this selection. For each school, the administration took place on one day, and occupied one lesson of class time. Informed consent was also obtained from each of the teachers and each student who participated. Provision was made for any students

who did not wish to participate to be supervised in another room. This provision was unnecessary as every student in each class agreed to participate. The classroom teacher left the classroom after introducing the researcher.

All of the schools were coeducational state schools, spread through the Greater Auckland urban area. Table 5 shows the roll, decile rating, ethnic makeup (New Zealand European, NZE; Maori, M; Pacific Island, PI; Asian, A; and Other, O) and gender balance (Female, F; Male, M) for each of the four schools, plus the mean for all Greater Auckland high schools. The diverse ethnic makeup of the schools reflects the catchment area for each school. The names of schools have been changed because the schools had agreed to participate anonymously, as required by the Human Participants Ethics Committee (Ref: 2000/090).

Students from 16 classes completed one of the three versions of the questionnaire. Form A (Appendix One) was completed by 138 students, Form B (Appendix Two) by 160 students and Form C (Appendix Three) by 154 students. The three different forms were randomly allocated to members of the same class. The classes were not randomly selected, but they included classes taking academic mathematics and non-academic mathematics programmes at these levels. Completed questionnaires were received from a total of 452 students (100% response from the students in the 16 selected classes)

Table 5 School Descriptives for Trial One

School	Roll	Decile (SES)	Ethnicity (%)					Gender (%)	
			NZE	M	PI	A	O	F	M
Ashcroft	1400	7	75	13	2	8	2	53	47
Eruera	2000	3	48	17	13	14	2	55	45
Hemi	1300	10	75	5	0	16	4	47	53
Meadow	1500	8	43	6	4	36	5	52	48
Auckland Mean	1002	5	51	11	15	14	8	50	50

The students (Table 6) were representative of the students in their school, with one major exception. Maori and Pacific Island students were under-represented, and New Zealand European and Asian students over-represented in the sample. This skewed distribution reflects the composition of classes at the senior level in New Zealand schools. In the March 2000 School Statistics (Ministry of Education, 2000b), 35% of Maori students and almost 27% of Pacific Island students left school without a formal qualification, which is the entrée into the senior school classes that participated in the survey. There were no Year 11 students from Hemi High School, as these students were preparing for school examinations at the time of the survey.

Table 6 Participant Descriptives for Trial One

School	N	Year (N)			Ethnicity (N)					Gender (N)	
		11	12	13	NZE	M	PI	A	O	F	M
Ashcroft	86	19	24	43	36	3	2	34	11	33	53
Eruera	123	83	23	17	58	8	23	21	13	67	56
Hemi	100	-	56	44	56	1	1	39	4	40	60
Meadow	143	21	54	68	47	1	4	61	30	61	82
Totals	452	123	157	172	197	13	30	155	58	201	251
Percent		27	35	38	44	3	7	34	13	45	55

Analysis

The completed questionnaires were scanned using Remark Office OMR 4.0 (Principia Products Inc, 1997), and saved as SPSS and Excel files. On the few occasions where a student had three consecutive responses showing for an item, the mean was entered, but if the three responses were not consecutive the response for that item was entered as a blank. Where two adjacent responses were made, a random number was generated. If the random number was even, the higher of the two responses was recorded. If the random number was odd, the lower of the two responses was recorded. Non-adjacent responses were entered as a blank. Students who had completed fewer than 50% of the items were deleted from the analysis.

Two approaches to statistical analysis were employed – factor analysis, and IRT methodologies. There are three competing tensions in using these approaches that have to be resolved: (1) identifying and selecting items which have desirable IRT characteristics; (2) achieving a parsimonious extraction of factors ensuring that the factors are interpretable and meaningful; and, (3) ensuring that the selected items map the domain of exemplary teaching as articulated by the Standards. An item which has good IRT parameters but which contributes nothing to a factor, or an item that weighs well on a factor but which has poor IRT parameters may be rejected. These methodologies are briefly outlined, and the selection of items from the different Forms follow this discussion.

Classical Test Theory and Factor Analysis

Nunnally (1967) outlined four essential considerations in the process of test assembly: content analysis and validity; item difficulty; item-total correlation; and, factor analysis. However, none of these could stand-alone. In this case, the content of the instrument is derived directly from the amended NBPTS AYA Mathematics Standards. At each stage during the development and field trials, the items were referenced back to the Standards to ensure a complete mapping of the domain. This step is common to both CTT and IRT test construction methodologies. In CTT, item difficulty is usually measured by *p*-values, the probability that a student correctly answers the item. As all responses to a questionnaire item have value (for polytomous items there is no right or wrong answer), the item mean is representative of the p-value, with the item's skewness indicating the asymmetry of the distribution of responses. The item mean represents the average assessed proficiency of the teacher on the item. The item-total correlation is an indicator of the strength of the relationship between the item and the domain of the questionnaire. This is the typical CTT discrimination index, and items with low or negative correlations need to be checked for ambiguous wording, and revised, trialled again or removed from the item pool.

Exploratory factor analysis seeks to reveal the structure of the domain of interest, and provide a more meaningful explanatory framework to understand the complex network of relationships measured by the items. There has been considerable debate in the

literature about the merits of various methods of extraction, factor rotation and the number of factors to extract, with clear-cut rules postulated for an optimal solution but these rules have not been universally accepted. Mechanistic rules lack the flexibility of more subjective methods of making decisions in this area, and psychometricians differ in their preferences about how to obtain an optimal solution (see Fabrigar, Wegener, MacCallum, & Strahan, 1999 for a discussion of factor analysis in psychological research). The most widely used model fitting method or factor extraction procedure is maximum likelihood (ML), with oblique rotations. According to Fabrigar et al (p. 291), this procedure "provides a much better simple structure, more interpretable results, and more theoretically plausible representations of the data" than principal components extraction with orthogonal rotation. For this study, ML with direct oblimin rotations was used. Missing cases were deleted pairwise.

To determine the number of factors, multiple methods were employed, including the use of eigenvalues greater than one, a scree test and as the ultimate test, the interpretability of the derived factors. Items were being constantly deleted from the analysis, so as an additional guide, items which had factor loadings of less than .30 were deleted unless there were other reasons for maintaining them in the analysis. These reasons included adequate mapping of a domain, and the desire to over-determine a factor. Over-determining involves retaining more items than might be considered necessary (that is, the additional items provide no additional information) to ensure that a factor is adequately represented. As a general guideline, at least three to five items were required to carry a factor (and the associated items) through to the next stage of analysis, as these produce more stable factors (Fabrigar et al., 1999; S. M. Harris & Halpin, 2002). The pattern matrix and correlations among the factors are reported for each Form. Coefficients of internal consistency (Cronbach's alpha, α) were also computed for each Form.

Although CTT has served the test community well for the first half of the twentieth century, its limitations have been well documented (Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991). There are advantages in using CTT for instrument assembly (viz., it is easy to meet the underlying assumptions; the use of relatively small sample sizes; and, straight-forward statistical analyses), but these have to be weighed against the difficulties associated with the model. For example, in CTT, item statistics

are sample dependent - correlations are calculated between each item and the total scale, as well as average item scores, and these statistics are only meaningful for that specific sample of people. A different sample of examinees would produce a different mean item score as well as a different correlation between the item and the rest of the scale. A consequence of this is that field-testing items for test development using CTT becomes very difficult, as the item statistics are group dependent and not generalisable to other samples. Furthermore, examinee scores are test-dependent – a different score would result from another test purportedly measuring the same "trait".

A further problem identified by Cattell (1973), is the possible existence of a "bloated specific" factor. Bloated specific factors occur with highly correlated items that effectively state the same thing through item overlap, sharing a method component or a response bias. In order to avoid a "bloated specific" single-variable factor, the table of correlations was inspected for item pairs with relatively high correlations, ($r > .6$), and where appropriate one of the two items deleted from the factor analysis. For example, in Form A, A1 (My mathematics teacher encourages all students to participate fully in class) and A6 (My mathematics teacher makes sure that all students participate in class irrespective of their gender, ethnicity, cultural background, prior experience and expectations) had a correlation of .65. Item A6 had a low discrimination value, and was deleted from further analysis.

Item Response Theory item selection and test construction

In psychological and educational testing, IRT has been used for a variety of purposes including test construction to obtain a test with pre-determined measurement properties from a pool of calibrated items (Stahl, Shumway, Bergstrom, & Fisher, 1997), computer adaptive testing and administration which enables tailored tests to be administered in accordance with a person ability level, θ (Bergstrom & Lunz, 1992; Sykes & Ito, 1997; Sykes & Yen, 2000), test scoring (Ludlow & O'Leary, 1999; Stocking, 1996), test equating to enable different forms of a test to be reported on the same scale (Baker, 1993; L. L. Cook & Eignor, 1991; Glas & Beguin, 1996; Zeng & Kolen, 1995), attitude and opinion measurement (Chow & Winzer, 1992; Cochran, 1997; Roberts & Laughlin, 1996), detection of item bias or differential item functioning (Camilli & Congdon, 1999; Kim & Cohen, 1998; V. S. L. Williams, 1997) and

diagnostic assessment (Birenbaum & Tatsuoka, 1993; Cooke & Michie, 1997; Gumpel, Wilson, & Shalev, 1998).

When used for ability testing, IRT provides a theoretical model that assumes that there is a continuous variable (some characteristic of the person, or latent trait, such as ability or in the current case proficiency) that relates the probability that a person (examinee) will correctly respond to an item. As proficiency increases, the probability of the various responses along the Likert scale to the item increases. This probability is expressed as a function of the variable and the purpose is to locate both the item and the respondent values as points on the same scale. The examinee statistic is expressed along a scale that is similar to the standard normal variate, or z-scale. In essence, the researcher is determining the amount of the latent proficiency that the examinee has. In this research, the trait of interest ishighly accomplished teaching of mathematics. Students respond on a variable scale using Likert-type scales, and their responses indicate the amount of the trait that the student believes the teacher has. Therefore, as the reported teacher proficiency increases, the probability of a student assigning the teacher a given rating increases.

IRT has emerged as a powerful tool in psychological and educational testing. This came about because of concern about the discontinuity between the role of items and test scores that is the basis of classical test theory. The classical model was based on the characteristics of the items themselves, rather than on the score that resulted from that test (Baker, 1992). The main advantage of IRT over Classical Test Theory (CTT) is that the item statistics (parameters) are theoretically independent of the sample of examinees to whom the instrument is administered – that is, parameter invariance. What this means is that once the items have been calibrated, the item statistics are expected to stay the same regardless of the distribution of ability of any new sample answering those items. IRT also provides ability statistics (scores) that are independent of the instrument used, and well-defined standard errors which make it possible to calculate an accuracy index of an individual's ability as estimated by the instrument.

There are three IRT models, all of which are variants of the three-parameter model. The three-parameter (3PL) model provides estimates for the slope parameter (*a*, or discrimination), location parameter (*b*, or item difficulty) and a pseudo-guessing

parameter (*c*, which provides an estimate of the possibility of guessing the correct answer). The formula for the three-parameter model is

$$P_{xi}(\theta) = c_i + (1 - c_i)\frac{e^z}{1+e^z}$$
$$\text{where } z = Da_i(\theta - b_{xi}).$$

The variable θ represents the latent trait, a_i is the discrimination index for item *i*, b_{xi} is the difficulty parameter for category score *x* in item *i*, c_i is the pseudo-guessing parameter for item *i*, and the scaling constant D is equal to 1.702 (to make the logistic and normal ogive form). A two-parameter (2PL) model sets the *c* value to 0, and is particularly appropriate for research of this kind, where each response has value and there is no guessing. The 2PL model assumes that each item differs in both difficulty and discrimination, and provides estimates for both. The one-parameter (1PL) model (Rasch model) assumes that the items all discriminate equally and differ only in their difficulty, with no guessing. In this latter case, the only parameter that is estimated is the *b* or difficulty parameter. The Rasch Unidimensional Measurement Model (RUMM) software used with the 1PL model assumes a fixed discrimination parameter.

CTT and earlier IRT models were originally concerned with dichotomous models – the examinee had the answer right or wrong. In this study, a Likert–type scale is employed to measure the respondent's level of agreement or disagreement with a statement. These measures are not dichotomous (yes or no, right or wrong) but of varying degree. The model has to assess information from all category responses, so a polytomous IRT model is appropriate. In addition, we wish to know whether examinees who agree with a particular statement vary in the extent to which they agree. Each of the six response options for each item is explicitly included in the mathematical model so that it is possible to determine for an examinee with a particular status on the trait what the probability is of observing each response option.

There are four common polytomous models – partial credit, rating scale, nominal response and graded response. The partial credit (PC) model (G. N. Masters, 1982) is employed when an examinee is to be awarded some credit for a response that is partially correct – partially solving an equation in mathematics, for example, and

gaining two out of three marks available. The rating scale (RS) model (Andrich, 1978a,, 1978b) is derived from the PC and is appropriate for use with rating scales such as Likert-type response. In this model, the threshold or boundary concept is utilised. This means that there is a boundary or threshold above which an examinee is expected to respond in category *k* or higher as opposed to a category lower than *k.* In this model, the boundaries remain constant across all items. The nominal response (NR) model has more of a diagnostic function. Consider the test item (taken from Tatsuoka, 1983):

-6-(-10) = ?

A -16

B -4

C 4

In this case, responses are either correct or incorrect (which would initially indicate a dichotomous model), but each of the incorrect responses (A and B) represents an erroneous application of the rules regarding the subtraction of signed numerals. Tatsuoka was able to show that incorrect responses such as these could augment our estimates of an examinee's ability by providing information about the level of understanding, rather than proficiency. That is, for the responses A and B the examinee performed a binary operation involving 6 and 10, even if it was not the correct one. The graded response (GR) model (Samejima, 1969) is similar to the RS model, as it also utilises a boundary perspective. However, it differs in that the boundaries can vary across items, whereas in RS the boundaries are fixed. In a comparison of the PC and GR models, (De Ayala, Dodd, & Koch, 1992) found that the GR model was able to fit substantially more items than the PC model. Therefore there were two reasons for choosing the GR model - its ability to measure the extent to which examinees who agree with a particular statement vary in the extent to which they agree, and its ability to provide a better person fit to more items.

Decisions in constructing SEAT-M were based on Samejima's (1969) Graded Response (GR) model, a polytomous Item Response Theory (IRT) model. The GR models two parameters, *a* and *b*, and these are related to item discrimination and item difficulty in CTT terminology, although they are not equivalent in all respects. The *a* parameter is often referred to as the slope parameter, and is constant across each item's response

categories. It measures the steepness of the curve at the point of inflexion. Discrimination indicates the capacity of the item to detect differences between examinees or to distinguish between candidates of differing proficiency as measured by the instrument. That is, it measures the strength of the relationship between the item and the trait. In CTT, difficulty indicates how easy or hard it is for examinees to agree with the statement, but in IRT *b* has been called the location or threshold parameter. It represents the θ-level at the point of inflexion and indicates the point at which an individual is likely to be in that or a higher category as opposed to a lower one.

The formula for the GR model is given by

$$P_{xi}(\theta) = \frac{e^{z}}{1+e^{z}}$$
$$\text{where } z = Da_i(\theta - b_{xi}).$$

where $P_{xi}(\theta)$ is the probability that an examinee with ability θ receives a category score of x_i or higher on item *i*. As category 0 is the lowest category, the probability of scoring in that category or higher, $P_0(\theta)$, is defined as unity. Given that there are six response categories in the questionnaires used in this research, $P_2(\theta)$ is the probability that an examinee with ability θ responds in categories 2, 3, 4, or 5 rather than in categories 0 or 1. That is, $P_2(\theta)$ is a cumulative probability. To find the probability (p_{xi}) that an examinee will respond in a specific category, we can use the difference between the cumulative probabilities of the adjacent categories – for example, the probability of responding in category 2, $p_2(\theta)$, is given by:

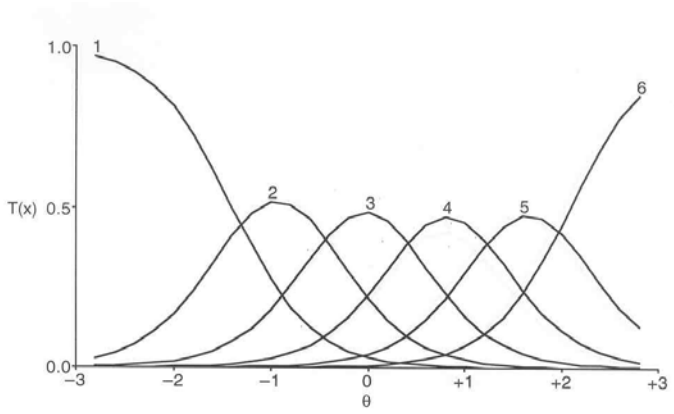
$$p_2(\theta) = P_2(\theta) - P_3(\theta)$$

Estimation of the parameters is referred to as calibration and for most models is extremely difficult (if not impossible) by hand. Computer programmes employing either joint maximum likelihood or marginal maximum likelihood approaches have been developed including Logist (Wingersky, Barton, & Lord, 1982), Bilog (Mislevy & Bock, 1990) and Multilog (Thissen, 1991a). I will only refer to the latter as the others are for dichotomous items.

The item parameters were estimated using Samejima's (1969) graded response IRT model. This was done using Multilog (Thissen, 1991a) to calculate the parameters and Plotlog (Thissen, 1991b) to draw the trace lines (also called category response curves or item characteristic curves), information curves and test characteristic curves derived from the parameters. As there were six response categories, there were five threshold or *b* parameters for each item (b_1, b_2, b_3, b_4, and b_5). The first location parameter, b_1, represents the threshold between the "strongly disagree" and "tend to disagree" categories on the scale, b_2 the threshold between Tend to Disagree and Slightly Agree categories, and so on across the scale.

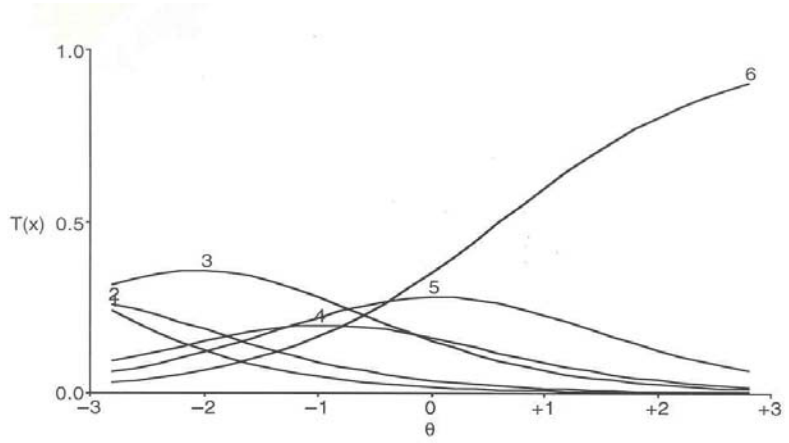
The slope parameter and threshold parameters determine the shape and location of the trace lines. Higher values of *a,* the slope parameter, will generate narrower and more peaked trace lines, which indicates that the response categories differentiate among trait levels well. The threshold parameters, b_i, indicate the points on the latent scale where respondents have a 0.50 probability of responding above that threshold. The ideal graph for an item would have a high value of *a*, and the values of b_i would be reasonably close together, producing a narrow, peaked set of trace lines, such as for item N37 (Figure 1). Low values of *a*, and extreme or asymmetric values of b_i will produce flat or extended trace lines that indicate random responses to the response categories, such as item A60 (Figure 2).

Figure 1 Item Characteristic Curves for an item with ideal characteristics, Item N37



Note: $a = 2.37$, $b_1 = -1.41$, $b_2 = -0.39$, $b_3 = 0.48$, $b_4 = 1.31$, $b_5 = 2.17$

Figure 2 Item Characteristic Curves for an item with poor characteristics, Item A60



Note: a = 1.30, b_1 = -3.05, b_2 = -2.11, b_3 = -0.91, b_4 = -0.26, b_5 = 0.70

Item Information

Pearson and Garavaglia (1997) enumerate several ways in which item information can be viewed in decision making. Noting that an additional item can 'add value' and hence information, they outline five conceptions of the function of new information. The first is technical. Item information is a measure of the psychometric precision of a measurement. In IRT, this leads to increased confidence in decisions about the ability level on the underlying trait, as it provides "maximal capacity to discriminate among individuals with differing levels of the ability in question" (p.3). The second function of item information leads to the formulation of new constructs – it provides something new to our understanding of the construct. A third way of viewing item information is as psychological support in construct validity. Psychometric measurements provide the technical information we require for construct validity, but there are occasions when the test constructor increases the sample of items to provide a greater sense of trust that the domain has been adequately represented. A fourth view of new information is that it provides a second perspective on a task that has already been examined from a different perspective. Such would be the case where, for example, an essay that has been examined from the perspective of its content is re-examined from the perspective of its ability to communicate effectively to an audience. The final conception provides another perspective on a decision that has already been made.

In this thesis, item information is defined in the first sense as a measure of psychometric precision - "a quantity inversely proportional to the squared length of the confidence interval around an estimate of the examinee's ability" (Birnbaum, 1968) – and means that when the information available at an ability level is high, the standard error of estimation is small. High information means the confidence intervals are small, increasing our confidence in decisions about the ability level on the underlying trait, as it provides "maximal capacity to discriminate among individuals with differing levels of the ability in question" (Pearson & Garavaglia, 1997, p. 3).

Plotlog also produces an item information curve (IIC) for each item. Figure 3 shows the IIC for an item (A36 "*My mathematics teacher provides time for us to be involved in peer tutoring*"), that provides very little information almost uniformly across the ability scale.

Figure 3 Item Information Curve for low information item, A36

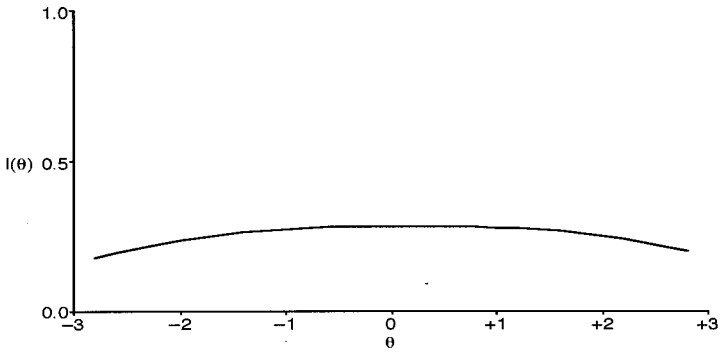


Figure 4 Item Information Curve for high information item, A34

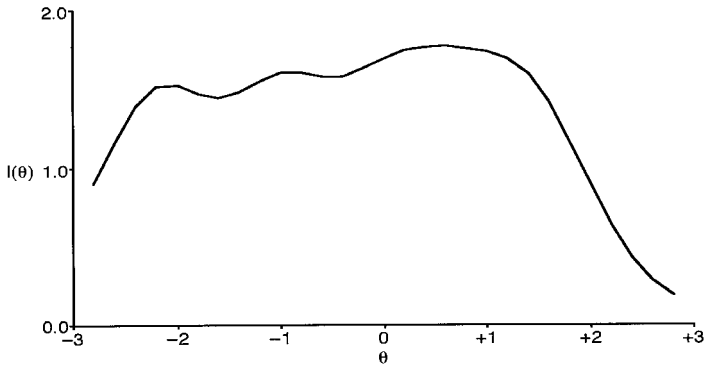


Figure 4 shows item A34 ("My mathematics teacher encourages us to test mathematical ideas and discover mathematical principles") which has high information across the scale from θ = -2 to θ = +2, then tapers sharply downwards as θ increases beyond +2. As the purpose of this research is to identify teachers who display high levels of mathematics teaching proficiency, ideal items would continue to have high information values for large positive values on the ability scale. In practice, very few items of the 191 items on all three forms had this characteristic. In part, this is a consequence of the relatively low frequency with which students selected the two extreme points on the rating scale, resulting in a higher value of the SE for b_1 and b_5.

Results

Form A

Descriptive statistics

The number of responses, mean, standard deviation, skew and kurtosis statistics for each of the 65 items in Form A are shown in Table 7.

As only one item, A09, had a mean rating exceeding 5, on the six point scale used, then the mean ratings for teachers evaluated using this instrument were not very high -- perhaps an indication that students do not give high ratings to their teachers indiscriminately. The mean rating on individual items ranged from a high of 5.05 (Item A09 "My mathematics teacher seems to have a broad and deep understanding of the concepts, principles, techniques and reasoning methods of maths") to a low of 2.81 (Item A16 "My mathematics teacher has introduced us to a variety of new topics like fractals, linear programming, cracking codes and technology based numerical methods"), with a mean rating for all items on Form A of 3.91. Cronbach's alpha reliability for this questionnaire form was α=0.97, which suggests that it is meaningful to interpret scores on the total scale.

Factor analysis

The 65 items of Form A were subjected to maximum likelihood factor analysis, with oblimin rotation. The Kaiser-Meyer-Olkin measure of sampling adequacy value was a "meritorious" .89 (Kaiser, 1974, p. 35). It was possible to identify five interpretable factors, explaining 52.8% of the total variance. The pattern matrix for these five factors is shown in Table 8.

Factor One, Mathematical Thinking and Problem Solving, (eigenvalue = 25.38) accounts for 39.0% of the common variance; Factor Two, Relates Mathematics to the Real World, (eigenvalue = 2.92) accounts for 4.5% of the variance; Factor Three, Becoming Mathematical Learners (eigenvalue = 2.29) accounts for 3.5% of the shared variance; Factor Four, Student Engagement with the Curriculum (eigenvalue = 2.05) accounts for 3.15% of the variance; and, Factor Five, Language and Processes of Mathematics (eigenvalue = 1.73) accounts for 2.7% of the common shared variance.

The correlations between factors range from .28 to .44. These correlations are relatively small, and indicate that the factors measure a distinct (but somewhat overlapping) aspect of exemplary mathematics teaching. The goodness of fit statistic $\chi^2(1765) = 2081.52$, $p < .01$ indicates very good specification of the five factor model.

Item Response Theory

The *a* and *b* threshold item parameters for Form A are shown in Table 9. The table also shows the CTT item-total (point biserial) correlations.

The *a* parameters for the 64 items on Form A were generally high, ranging from a low of .53 to a high of 2.40, with a mean *a* parameter for the 64 items of 1.50.

The ten items with the highest discrimination indices (*a* values) were A34 (2.40), A03 (2.10), A47 (2.08), A55 (2.07), A30 (2.04), A39 (2.02), A11 (1.97), A07 (1.94), A33 (1.91), and A32 (1.90). These items also had excellent item-total correlations, all in excess of .67. Items A11, A07 and A32 load on Factor Two, Relates Mathematics to the Real World, and the others load on Factor One, Mathematical Thinking and Problem

Table 7 Descriptive Statistics for Form A

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
A01	*Encourages all students to participate fully in class.*	138	4.48	.11	1.31	-.51	.21	-.81	.41
A02	*provides time to apply maths to a broad range of interesting subjects and applications.*	136	3.97	.11	1.33	-.08	.21	-.93	.41
A03	*encourages us to explore, confront and challenge new ideas presented in maths.*	137	4.07	.11	1.34	-.32	.21	-.41	.41
A04	*uses technology, activities and physical models to help us recognise the connections among different ways of representing ideas in maths.*	137	2.95	.13	1.52	.56	.21	-.74	.41
A05	*helps us to make links between the different strands of maths and other aspects of our lives.*	137	3.59	.12	1.40	.19	.21	-.84	.41
A06	*makes sure that all students participate in class regardless of their gender, ethnicity, cultural background, prior experience and expectations.*	137	4.66	.13	1.47	-1.07	.21	.33	.41
A07	*helps us to see the "big picture" by relating the themes in maths.*	137	3.79	.12	1.42	-.20	.21	-.70	.41
A08	*recognises settings in the real world where mathematical solutions are worthwhile.*	137	3.91	.12	1.41	-.27	.21	-.75	.41
A09	*seems to have a broad and deep understanding of the concepts, principles, techniques and reasoning methods of maths.*	138	5.05	.10	1.17	-1.44	.21	1.99	.41
A10	*shows us how to use indirect methods (like testing extreme cases, organised searches , etc.) to solve problems.*	132	3.86	.12	1.41	-.28	.21	-.88	.42
A11	*helps us to understand & appreciate the powerful relationships between mathematical ideas and problems.*	138	3.98	.11	1.30	-.44	.21	-.33	.41

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
A12	*regards technology as an essential component of teaching maths.*	138	3.21	.12	1.35	.19	.21	-.76	.41
A13	*uses a variety of processes to describe patterns in different kinds of data.*	137	3.81	.11	1.24	-.17	.21	-.56	.41
A14	*asks us to explain our solutions to problems and justify our conclusions.*	138	4.03	.13	1.58	-.35	.21	-1.05	.41
A15	*shows us how we can use geometry to solve problems in the real world.*	136	3.46	.13	1.49	-.05	.21	-1.02	.41
A16	*has introduced us to a variety of new topics like fractals, linear programming, cracking codes and technology based numerical methods.*	135	2.87	.13	1.52	.46	.21	-.71	.41
A17	*helps us to apply our growing knowledge in both pure and applied settings.*	134	3.69	.12	1.35	-.03	.21	-.85	.42
A18	*shows and challenges us to discover and describe patterns in visual, numerical and symbolic data.*	134	3.70	.12	1.34	.05	.21	-.81	.42
A19	*helps us to understand mathematical concepts rather than routine computational procedures and proofs.*	135	4.10	.13	1.50	-.48	.21	-.78	.41
A20	*teaches us about the role that maths has in the history of problem-solving and decision-making across time and cultures.*	137	2.99	.14	1.63	.42	.21	-1.08	.41
A21	*helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	137	3.13	.13	1.47	.18	.21	-.94	.41
A22	*has a classroom where we are engaged in learning.*	136	4.20	.13	1.51	-.56	.21	-.62	.41
A23	*teaches us that proof provides a standard of precision that sets maths apart from other subjects.*	131	3.62	.13	1.47	.00	.21	-1.01	.42

	My mathematics teacher ...	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
A24	*organises tasks that help us see the relationship between different ways of representing mathematical ideas.*	137	3.68	.12	1.39	-.05	.21	-.66	.41
A25	*involves us in maths competitions, fairs (e.g., Mathex) and other events that allow us to demonstrate our mathematical knowledge and skills.*	138	2.99	.14	1.68	.35	.21	-1.10	.41
A26	*shows us how we can use measurement to solve problems in the real world.*	137	3.63	.12	1.44	-.05	.21	-.79	.41
A27	*provides time to develop problem solving skills that we can use both in maths and outside the classroom.*	137	3.72	.13	1.54	-.08	.21	-1.04	.41
A28	*invites us to question ideas, offer ideas of our own, and argue in support of them.*	138	4.01	.14	1.64	-.35	.21	-1.05	.41
A29	*provides problems and applications to develop the maths we have learned.*	138	4.64	.11	1.25	-.72	.21	-.32	.41
A30	*shows us interesting and useful ways of solving problems.*	137	4.18	.12	1.37	-.51	.21	-.33	.41
A31	*encourages us to try different techniques to solve problems.*	138	4.05	.12	1.43	-.18	.21	-.85	.41
A32	*helps us to effectively apply ideas in maths to solving problems in the everyday world (e.g., the scientific, technical, arts, music worlds).*	138	3.01	.12	1.44	.40	.21	-.59	.41
A33	*helps us to build our own broad and deep understanding of maths.*	137	3.93	.12	1.44	-.10	.21	-1.15	.41
A34	*encourages us to test mathematical ideas and discover mathematical principles.*	137	3.80	.12	1.39	.14	.21	-.93	.41
A35	*shows us how we can use statistics to solve problems in the real world.*	138	3.70	.13	1.53	-.22	.21	-.98	.41
A36	*provides time for us to be involved in peer tutoring.*	138	3.54	.15	1.76	-.05	.21	-1.34	.41

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
A37	*uses a variety of teaching methods to represent, solve and make decisions about real problems.*	138	3.79	.11	1.29	-.01	.21	-.50	.41
A38	*uses basic skills to solve more complex problems.*	136	4.42	.12	1.34	-.62	.21	-.27	.41
A39	*weaves together the pieces of maths to form a comprehensive and flowing mathematical experience.*	133	3.68	.12	1.36	-.05	.21	-.84	.42
A40	*distinguishes between different ways of solving a problem to illustrate the most efficient method.*	134	4.25	.12	1.43	-.50	.21	-.61	.42
A41	*teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.*	136	3.05	.12	1.34	.17	.21	-.80	.41
A42	*shows us how the different strands of maths are linked together.*	138	3.86	.11	1.34	-.17	.21	-.66	.41
A43	*provides frequent opportunity for us to reflect on our own learning.*	137	3.80	.13	1.48	-.11	.21	-1.00	.41
A44	*encourages us to seek more than one solution to problems.*	137	3.81	.12	1.43	.05	.21	-.92	.41
A45	*shows us how we can use calculus to solve problems in the real world.*	134	3.25	.14	1.62	.04	.21	-1.22	.42
A46	*provides tasks that help us to see the many different ways of representing mathematical ideas & problems.*	137	3.77	.11	1.31	.07	.21	-.80	.41
A47	*Helps us to communicate better in maths.*	136	3.82	.12	1.43	-.31	.21	-.68	.41
A48	*tries out different ways of involving us in our learning of maths.*	138	3.64	.12	1.41	.02	.21	-.84	.41
A49	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	137	3.01	.12	1.35	.58	.21	-.18	.41
A50	*teaches us the fundamental processes of mathematical thinking – exploration, inference, interpretation, representation, modelling, conjecture and analysis.*	132	3.37	.12	1.36	.03	.21	-.66	.42

	My mathematics teacher ...	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
A51	*encourages us to question our peers when discussing new ideas, and solving problems.*	138	3.51	.13	1.54	.08	.21	-.98	.41
A52	*helps us construct an understanding of the language and processes of maths.*	138	3.88	.11	1.33	-.19	.21	-.70	.41
A53	*provides time for us to reflect on and talk about the maths we are learning.*	138	3.57	.13	1.54	.02	.21	-1.05	.41
A54	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	133	3.71	.12	1.32	.02	.21	-.48	.42
A55	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	138	3.86	.12	1.36	.02	.21	-.80	.41
A56	*uses a variety of methods to collect, organise, represent and summarise collections of data.*	138	3.62	.12	1.41	-.01	.21	-.90	.41
A57	*conveys to the class the idea that maths relates to the real world.*	137	3.61	.13	1.51	.03	.21	-1.01	.41
A58	*provides time for us to develop our own personal interests by formulating and solving our own problems.*	138	3.57	.14	1.59	.03	.21	-1.08	.41
A59	*uses rules to prove theorems and draw conclusions.*	137	4.55	.12	1.38	-.72	.21	-.35	.41
A60	*teaches us to use calculators and computers effectively for both routine and complex problems.*	136	4.58	.12	1.41	-.70	.21	-.50	.41
A61	*encourages us to question and discuss the mathematical ideas and concepts we are taught.*	138	3.96	.13	1.46	-.22	.21	-.84	.41
A62	*uses a variety of activities to involve each of us in our learning of maths.*	138	3.33	.12	1.40	.19	.21	-.80	.41
A63	*shows us how we can use algebra to represent patterns and solve problems in the real world.*	132	3.86	.14	1.58	-.23	.21	-1.02	.42
A64	*involves students in decisions about their learning of maths.*	138	3.77	.13	1.48	-.16	.21	-.85	.41

	My mathematics teacher ...	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
A65	*compared with all other maths teachers I have had, is the best.*	138	3.75	.16	1.85	-.16	.21	-1.45	.41
	Valid N (listwise)	91							

Table 8 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form A

		Factor loading					Used in	
Item	My mathematics teacher …	1	2	3	4	5	N	US
A28	*invites us to question ideas, offer ideas of our own, and argue in support of them.*	**.66**	.14	.04	-.07	-.01		
A33	*helps us to build our own broad and deep understanding of maths.*	**.56**	.08	-.11	.08	-.38		
A40	*distinguishes between different ways of solving a problem to illustrate the most efficient method.*	**.56**	-.00	-.06	.20	-.20		
A31	*encourages us to try different techniques to solve problems.*	**.55**	.07	.25	.08	.09	N26	US19
A30	*shows us interesting and useful ways of solving problems.*	**.54**	.15	.11	.20	.03	N09	US04
A06	*makes sure that all students participate in class regardless of their gender, ethnicity, cultural background, prior experience and expectations.*	**.53**	.02	.29	-.11	-.01		
A65	*compared with all other maths teachers I have had, is the best.*	**.50**	-.04	.28	.00	-.15		
A27	*provides time to develop problem solving skills that we can use both in maths and outside the classroom.*	**.49**	.18	.03	.32	.24	N04	
A09	*seems to have a broad and deep understanding of the concepts, principles, techniques and reasoning methods of maths.*	**.47**	.19	-.11	.11	-.03		
A14	*asks us to explain our solutions to problems and justify our conclusions.*	**.47**	.08	.15	-.04	-.08		
A01	*encourages all students to participate fully in class.*	**.46**	.04	.36	-.11	-.10		
A29	*provides problems and applications to develop the maths we have learned.*	**.45**	.00	-.01	.15	-.20		
A38	*uses basic skills to solve more complex problems.*	**.43**	.08	-.20	.30	-.14		
A39	*weaves together the pieces of maths to form a comprehensive and flowing mathematical experience.*	**.41**	.32	-.13	.09	-.28		
A47	*helps us to communicate better in maths.*	**.40**	-.15	.26	.33	-.25	N02	
A55	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	**.40**	-.07	.09	.34	-.30	N63	US48
A03	*encourages us to explore, confront and challenge new ideas presented in maths.*	**.39**	.33	.24	-.17	-.18		
A19	*helps us to understand mathematical concepts rather than routine computational procedures and proofs.*	**.38**	.18	-.04	.08	-.11		
A34	*encourages us to test mathematical ideas and discover mathematical principles.*	**.37**	.16	.09	.34	-.11	N61	US46

Item	My mathematics teacher …	Factor loading					Used in	
		1	2	3	4	5	N	US
A24	*organises tasks that help us see the relationship between different ways of representing mathematical ideas.*	**.34**	.17	.22	.05	-.08		
A22	*has a classroom where we are engaged in learning.*	**.34**	.04	.14	.04	-.15		
A60	*teaches us to use calculators and computers effectively for both routine and complex problems.*	**.28**	.02	.11	.23	-.11		
A05	*helps us to make links between the different strands of maths and other aspects of our lives.*	.09	**.78**	-.03	-.06	.02	N15	US09
A20	*teaches us about the role that maths has in the history of problem-solving and decision-making across time and cultures.*	-.03	**.68**	.07	.05	.05		
A21	*helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	-.11	**.57**	-.06	.24	-.10	N50	US38
A08	*recognises settings in the real world where mathematical solutions are worthwhile.*	.28	**.57**	.00	.02	.03		
A07	*helps us to see the "big picture" by relating the themes in maths.*	.28	**.54**	.04	-.04	-.13		
A18	*shows and challenges us to discover and describe patterns in visual, numerical and symbolic data.*	.16	**.46**	.14	.05	-.06		
A13	*uses a variety of processes to describe patterns in different kinds of data.*	.14	**.45**	.17	-.06	-.19		
A11	*helps us to understand & appreciate the powerful relationships between mathematical ideas and problems.*	.33	**.43**	.04	-.02	-.16		
A15	*shows us how we can use geometry to solve problems in the real world.*	-.15	**.40**	.16	.35	-.03	N12	US06
A45	*shows us how we can use calculus to solve problems in the real world.*	-.01	**.39**	.06	.25	-.15	N19	US13
A32	*helps us to effectively apply ideas in maths to solving problems in the everyday world (e.g., the scientific, technical, arts, music worlds).*	.09	**.37**	.10	.18	-.25		
A02	*provides time to apply maths to a broad range of interesting subjects and applications.*	.29	**.35**	.20	-.04	-.06		
A17	*helps us to apply our growing knowledge in both pure and applied settings.*	.24	**.32**	.06	.03	-.28	N62	US47
A10	*shows us how to use indirect methods (like testing extreme cases, organised searches , etc.) to solve problems.*	.19	**.31**	.26	.15	.05		
A23	*teaches us that proof provides a standard of precision that sets maths apart from other subjects.*	.17	**.26**	.05	.15	-.12		
A16	*has introduced us to a variety of new topics like fractals, linear programming, cracking codes and technology based numerical methods.*	-.13	**.25**	.14	.21	-.17		

Item	My mathematics teacher …	Factor loading					Used in	
		1	2	3	4	5	N	US
A62	*uses a variety of activities to involve each of us in our learning of maths.*	-.05	-.00	**.66**	.25	-.01		
A58	*provides time for us to develop our own personal interests by formulating and solving our own problems.*	.14	.06	**.56**	-.09	-.21		
A61	*encourages us to question and discuss the mathematical ideas and concepts we are taught.*	.25	.11	**.48**	.07	-.05	N32	
A04	*uses technology, activities and physical models to help us recognise the connections among different ways of representing ideas in maths.*	-.09	.39	**.46**	-.03	-.09		
A48	*tries out different ways of involving us in our learning of maths.*	.23	-.03	**.44**	.27	-.02		
A36	*provides time for us to be involved in peer tutoring.*	.00	.07	**.43**	-.04	-.19		
A43	*provides frequent opportunity for us to reflect on our own learning.*	.33	-.01	**.41**	.17	-.10		
A44	*encourages us to seek more than one solution to problems.*	.38	.06	**.38**	.00	-.02	N21	US15
A64	*involves students in decisions about their learning of maths.*	.20	.03	**.38**	.15	-.18		
A25	*involves us in maths competitions, fairs (e.g., Mathex) and other events that allow us to demonstrate our mathematical knowledge and skills.*	.05	.12	**.27**	.12	.15		
A35	*shows us how we can use statistics to solve problems in the real world.*	.10	.26	-.17	**.62**	.02	N51	US39
A46	*provides tasks that help us to see the many different ways of representing mathematical ideas & problems.*	.14	-.02	.34	**.57**	.04		
A63	*shows us how we can use algebra to represent patterns and solve problems in the real world.*	-.07	.15	.14	**.54**	-.02	N65	US50
A50	*teaches us the fundamental processes of mathematical thinking – exploration, inference, interpretation, representation, modelling, conjecture and analysis.*	.01	-.15	.14	**.52**	-.28	N08	US03
A49	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	-.11	.24	.13	**.49**	-.23	N57	US42
A26	*shows us how we can use measurement to solve problems in the real world.*	.31	.42	-.20	**.43**	.19		
A57	*conveys to the class the idea that maths relates to the real world.*	.00	.33	.14	**.42**	-.09	N20	US14
A42	*shows us how the different strands of maths are linked together.*	.26	.13	.10	**.42**	-.03		
A56	*uses a variety of methods to collect, organise, represent and summarise collections of data.*	.10	.05	.23	**.40**	-.07	N40	US30
A37	*uses a variety of teaching methods to represent, solve and make decisions about real problems.*	.28	.15	.11	**.38**	-.03		
A59	*uses rules to prove theorems and draw conclusions.*	.12	-.02	-.11	**.37**	-.33		

Item	My mathematics teacher …	Factor loading 1	2	3	4	5	Used in N	US
A41	*teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.*	-.05	.31	.08	**.33**	-.15	N18	US12
A52	*helps us construct an understanding of the language and processes of maths.*	.24	-.03	.06	.14	**-.68**	N16	US10
A54	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	.13	.13	-.05	.19	**-.51**	N43	US32
A53	*provides time for us to reflect on and talk about the maths we are learning.*	.22	.16	.20	-.09	**-.45**	N24	US17
A12	*regards technology as an essential component of teaching maths.*	-.23	.30	.20	.01	**-.40**	N07	
A51	*encourages us to question our peers when discussing new ideas, and solving problems.*	.28	.12	.26	-.03	**-.28**		

Factor correlations	**1**	**2**	**3**	**4**	**5**
Factor 1: Develops mathematical thinking and problem solving	--				
Factor 2: Relates mathematics to the real world	.44	--			
Factor 3: Becoming mathematical learners	.39	.39	--		
Factor 4: Presents different representations of mathematics	.38	.48	.28	--	
Factor 5: Language and processes of mathematics	-.39	.36	-.35	-.34	--

N and US. Items prefixed with US were in the final instrument (SEAT-M) used in the USA. Prefix N indicates that they were further trialled in the November instrument, but did not make the SEAT-M instrument.

Explains 52.8% of variance

Cronbach's alpha = .974

Table 9 Classical Test Theory and Item Response Theory Item Statistics for 64 Form A Items

Scale	Item #	Item Mean *(SE)*	Item-total correlation r_{pbs}	a *(SE)*	b_1 *(SE)*	b_2 *(SE)*	b_3 *(SE)*	b_4 *(SE)*	b_5 *(SE)*
	A01	4.48 (0.11)	.586	1.45 (0.22)	-3.99 (1.07)	-1.96 (0.39)	-0.72 (0.24)	-0.03 (0.20)	1.16 (0.21)
	A02	3.91 (0.12)	.612	1.60 (0.28)	-3.19 (0.75)	-1.20 (0.27)	-0.21 (0.18)	0.69 (0.18)	1.68 (0.28)
	A03	4.04 (0.12)	.673	2.10 (0.31)	-2.04 (0.40)	-1.28 (0.22)	-0.41 (0.14)	0.52 (0.15)	1.35 (0.17)
	A04	2.93 (0.13)	.588	1.26 (0.25)	-1.46 (0.36)	0.08 (0.21)	0.92 (0.24)	1.58 (0.34)	2.60 (0.51)
US09	A05	3.57 (0.12)	.531	1.45 (0.27)	-2.56 (0.55)	-0.93 (0.25)	0.11 (0.20)	1.09 (0.18)	1.98 (0.36)
	A06	4.63 (0.13)	.528	1.34 (0.28)	-2.45 (0.56)	-1.81 (0.41)	-1.27 (0.33)	-0.38 (0.22)	0.65 (0.21)
	A07	3.76 (0.12)	.692	1.94 (0.31)	-1.86 (0.30)	-1.07 (0.22)	-0.16 (0.14)	0.69 (0.17)	1.66 (0.26)
	A08	3.88 (0.12)	.589	1.66 (0.27)	-2.19 (0.41)	-1.04 (0.23)	-0.36 (0.18)	0.69 (0.18)	1.73 (0.29)
	A09	5.05 (0.10)	.583	1.37 (0.31)	-3.23 (0.75)	-2.79 (0.63)	-1.77 (0.39)	-0.91 (0.25)	0.32 (0.21)
	A10	3.70 (0.14)	.638	1.59 (0.28)	-2.30 (0.44)	-1.05 (0.25)	-0.15 (0.18)	0.55 (0.19)	1.97 (0.36)
	A11	3.98 (0.11)	.690	1.97 (0.30)	-2.08 (0.33)	-1.35 (0.24)	-0.38 (0.17)	0.56 (0.14)	1.82 (0.28)
N07	A12	3.21 (0.11)	.474	0.96 (0.23)	-2.51 (0.72)	-0.79 (0.36)	0.66 (0.28)	1.81 (0.44)	3.85 (0.91)
	A13	3.78 (0.11)	.625	1.54 (0.31)	-2.83 (0.63)	-1.37 (0.30)	-0.26 (0.19)	0.90 (0.19)	2.34 (0.43)
	A14	4.03 (0.13)	.550	1.16 (0.25)	-2.62 (0.60)	-1.18 (0.33)	-0.50 (0.26)	0.37 (0.23)	1.49 (0.35)
US06	A15	3.41 (0.13)	.609	1.25 (0.25)	-1.95 (0.47)	-0.73 (0.26)	-0.12 (0.22)	1.08 (0.27)	2.56 (0.52)
	A16	2.81 (0.13)	.482	0.83 (0.23)	-1.50 (0.54)	-0.21 (0.32)	1.13 (0.28)	2.44 (0.65)	3.53 (0.99)
US47	A17	3.58 (0.12)	.653	1.61 (0.32)	-2.36 (0.52)	-1.09 (0.23)	0.02 (0.17)	0.80 (0.15)	2.20 (0.38)
	A18	3.59 (0.12)	.622	1.46 (0.29)	-2.68 (0.54)	-1.15 (0.28)	0.05 (0.19)	0.91 (0.22)	2.14 (0.39)
	A19	4.01 (0.14)	.574	1.34 (0.26)	-2.43 (0.57)	-1.29 (0.31)	-0.60 (0.24)	0.25 (0.19)	1.48 (0.33)
	A20	2.97 (0.14)	.525	1.17 (0.27)	-1.28 (0.40)	0.03 (0.22)	0.64 (0.25)	1.38 (0.22)	2.58 (0.52)
US38	A21	3.11 (0.13)	.564	1.13 (0.24)	-1.71 (0.45)	-0.45 (0.27)	0.50 (0.24)	1.64 (0.36)	3.06 (0.66)
	A22	4.14 (0.13)	.462	1.10 (0.25)	-2.72 (0.69)	-1.65 (0.44)	-0.83 (0.31)	0.30 (0.23)	1.43 (0.36)
	A23	3.43 (0.14)	.539	1.19 (0.24)	-2.54 (0.61)	-1.00 (0.31)	0.10 (0.22)	0.97 (0.25)	2.33 (0.48)
	A24	3.65 (0.12)	.598	1.52 (0.29)	-2.11 (0.42)	-0.99 (0.23)	-0.09 (0.18)	1.09 (0.17)	1.92 (0.36)
	A25	2.99 (0.14)	.302	0.53 (0.23)	-1.80 (0.95)	-0.36 (0.51)	1.28 (0.64)	2.71 (1.16)	4.57 (2.01)
	A26	3.60 (0.12)	.581	1.38 (0.28)	-2.13 (0.44)	-0.99 (0.28)	0.05 (0.18)	1.03 (0.25)	2.06 (0.40)
N04	A27	3.70 (0.13)	.601	1.57 (0.25)	-1.91 (0.40)	-0.73 (0.22)	0.03 (0.16)	0.73 (0.17)	1.59 (0.29)
	A28	4.01 (0.14)	.605	1.46 (0.26)	-1.93 (0.40)	-1.13 (0.28)	-0.24 (0.21)	0.39 (0.18)	1.21 (0.24)
	A29	4.64 (0.11)	.559	1.47 (0.28)	-3.83 (1.18)	-2.06 (0.43)	-1.09 (0.25)	-0.26 (0.18)	0.98 (0.23)
US04	A30	4.15 (0.12)	.696	2.04 (0.34)	-2.05 (0.33)	-1.46 (0.26)	-0.52 (0.17)	0.35 (0.14)	1.31 (0.21)
US19	A31	4.05 (0.12)	.571	1.69 (0.29)	-2.45 (0.50)	-1.31 (0.24)	-0.37 (0.18)	0.59 (0.17)	1.27 (0.24)
	A32	3.01 (0.12)	.707	1.90 (0.29)	-1.12 (0.23)	-0.21 (0.15)	0.64 (0.17)	1.57 (0.24)	2.23 (0.34)
	A33	3.90 (0.13)	.704	1.91 (0.32)	-2.64 (0.49)	-0.90 (0.19)	-0.13 (0.17)	0.54 (0.17)	1.50 (0.23)
US46	A34	3.77 (0.12)	.744	2.40 (0.38)	-2.20 (0.37)	-0.99 (0.17)	0.05 (0.13)	0.68 (0.14)	1.39 (0.21)

Table 9. Continued

Scale	Item #	Item Mean *(SE)*	Item-total correlation r_{pbs}	*a (SE)*	b_1 *(SE)*	b_2 *(SE)*	b_3 *(SE)*	b_4 *(SE)*	b_5 *(SE)*
US39	A35	3.70 (0.13)	.554	1.27 (0.24)	-2.00 (0.45)	-1.01 (0.27)	-0.17 (0.22)	0.76 (0.19)	2.12 (0.42)
	A36	3.54 (0.15)	.458	0.93 (0.21)	-1.75 (0.53)	-0.64 (0.31)	0.01 (0.27)	0.92 (0.24)	2.04 (0.50)
	A37	3.79 (0.11)	.612	1.81 (0.28)	-2.21 (0.40)	-1.37 (0.25)	-0.05 (0.15)	0.84 (0.17)	1.84 (0.28)
	A38	4.36 (0.12)	.550	1.35 (0.25)	-2.97 (0.64)	-1.97 (0.64)	-1.06 (0.27)	0.07 (0.20)	1.19 (0.29)
	A39	3.55 (0.13)	.733	2.02 (0.34)	-2.10 (0.33)	-0.82 (0.17)	-0.04 (0.16)	0.84 (0.17)	1.89 (0.29)
	A40	4.13 (0.13)	.588	1.69 (0.32)	-2.34 (0.49)	-1.50 (0.29)	0.58 (0.21)	0.23 (0.16)	1.17 (0.24)
US12	A41	3.01 (0.12)	.671	1.38 (0.24)	-1.59 (0.38)	-0.34 (0.23)	0.57 (0.19)	1.68 (0.29)	3.31 (0.68)
	A42	3.86 (0.11)	.684	1.65 (0.31)	-2.37 (0.45)	-1.27 (0.25)	-0.24 (0.18)	0.75 (0.18)	1.87 (0.30)
	A43	3.77 (0.13)	.697	1.74 (0.29)	-2.08 (0.41)	-0.93 (0.20)	-0.09 (0.18)	0.68 (0.16)	1.63 (0.25)
US15	A44	3.78 (0.12)	.583	1.47 (0.28)	-2.69 (0.55)	-1.09 (0.26)	-0.09 (0.19)	0.95 (0.30)	1.61 (0.32)
US13	A45	3.15 (0.14)	.541	1.33 (0.27)	-1.25 (0.35)	-0.40 (0.23)	0.16 (0.21)	1.14 (0.18)	2.45 (0.48)
	A46	3.75 (0.11)	.689	1.73 (0.29)	-2.79 (0.55)	-1.08 (0.24)	-0.04 (0.16)	0.90 (0.17)	1.87 (0.29)
N02	A47	3.77 (0.13)	.738	2.08 (0.34)	-1.70 (0.31)	-1.02 (0.20)	-0.19 (0.16)	0.67 (0.15)	1.70 (0.21)
	A48	3.64 (0.12)	.651	1.59 (0.27)	-2.35 (0.44)	-0.77 (0.23)	-0.07 (0.17)	1.11 (0.22)	1.89 (0.33)
US42	A49	2.99 (0.12)	.652	1.62 (0.27)	-1.69 (0.32)	-0.28 (0.19)	0.80 (0.17)	1.75 (0.29)	2.32 (0.38)
US03	A50	3.22 (0.13)	.578	1.06 (0.24)	-2.36 (0.61)	-1.08 (0.36)	0.26 (0.25)	1.60 (0.38)	3.20 (0.74)
	A51	3.51 (0.13)	.629	1.57 (0.26)	-1.67 (0.34)	-0.76 (0.21)	0.21 (0.19)	0.94 (0.19)	1.77 (0.30)
US10	A52	3.88 (0.11)	.680	1.82 (0.30)	-2.30 (0.39)	-1.16 (0.21)	0.27 (0.16)	0.71 (0.18)	1.90 (0.27)
US17	A53	3.57 (0.13)	.595	1.50 (0.26)	-1.85 (0.36)	-0.72 (0.21)	0.05 (0.18)	0.87 (0.19)	1.80 (0.30)
US32	A54	3.57 (0.13)	.553	1.56 (0.25)	-2.27 (0.46)	-1.29 (0.27)	0.03 (0.19)	1.13 (0.23)	1.93 (0.36)
US48	A55	3.86 (0.12)	.738	2.07 (0.34)	-2.37 (0.39)	-1.08 (0.20)	-0.13 (0.14)	0.75 (0.15)	1.43 (0.24)
US30	A56	3.62 (0.12)	.663	1.30 (0.24)	-2.48 (0.59)	-0.98 (0.26)	0.03 (0.21)	1.04 (0.23)	2.32 (0.40)
US14	A57	3.59 (0.13)	.718	1.79 (0.30)	-1.80 (0.35)	-0.73 (0.19)	0.11 (0.16)	0.80 (0.18)	1.70 (0.27)
	A58	3.57 (0.14)	.658	1.45 (0.24)	-1.76 (0.38)	-0.66 (0.23)	0.15 (0.18)	0.93 (0.19)	1.69 (0.31)
	A59	4.51 (0.12)	.436	1.06 (0.24)	-3.59 (1.04)	-2.45 (0.63)	-1.14 (0.37)	-0.21 (0.24)	1.03 (0.24)
	A60	4.51 (0.13)	.530	1.30 (0.28)	-3.05 (0.74)	-2.11 (0.46)	-0.91 (0.25)	-0.26 (0.21)	0.70 (0.24)
N32	A61	3.96 (0.12)	.722	1.72 (0.32)	-2.03 (0.42)	-1.15 (0.22)	-0.23 (0.17)	0.60 (0.19)	1.45 (0.19)
	A62	3.33 (0.12)	.623	1.16 (0.25)	-2.31 (0.55)	-0.71 (0.28)	0.43 (0.23)	1.50 (0.34)	2.87 (0.62)
US50	A63	3.70 (0.15)	.546	1.15 (0.27)	-2.27 (0.62)	-1.26 (0.37)	-0.18 (0.24)	0.59 (0.23)	1.70 (0.40)
	A64	3.77 (0.13)	.620	1.56 (0.27)	-1.99 (0.43)	-1.12 (0.26)	-0.12 (0.18)	0.77 (0.20)	1.79 (0.29)
	A65	3.75 (0.16)	.590						

Note: SE is in parentheses

Solving. These specific items have an emphasis on helping and encouraging the students in their mathematical endeavours.

Item selection from Form A

Through successive iterations, items were eliminated. These items either had inadequate IRT characteristics, or weak loading on the interpretable factors. In addition, the criterion was to have a minimum of five items loading on each factor, although after IRT analysis only two items were retained in Factor 3, and four items in Factor 5.

Table 10 shows the set of twenty-four items selected from Form A with their original factor loadings and IRT characteristics. In this table, the items are arranged by factor, with loadings below .30 suppressed.

Table 10 Factor loadings and IRT parameters for 24 items selected from Form A

Item	Factor loadings					IRT parameters					
	1	2	3	4	5	*a*	b_1	B_2	b_3	b_4	b_5
A31	**.55**					1.69	-2.45	-1.31	-.37	.59	1.27
A30	**.54**					2.04	-2.05	-1.46	-.52	.35	1.31
A27	**.49**					1.57	-1.91	-.73	-.03	.73	1.59
A47	**.40**				.33	2.08	-1.70	-1.02	-.19	.67	1.70
A55	**.40**				.34	2.07	-2.37	-1.08	-.13	.75	1.43
A34	**.37**			.34		2.40	-2.20	-.99	.05	.68	1.39
A05		**.78**				1.45	-2.56	-.93	.11	1.09	1.98
A21		**.57**				1.13	-1.71	-.45	.50	1.64	3.06
A15		**.40**		.35		1.25	-1.95	-.73	-.12	1.08	2.56
A45		**.39**				1.33	-1.25	-.40	.16	1.14	2.45
A17		**.32**				1.61	-2.36	-1.09	.02	.80	2.20
A61			**.48**			1.72	-2.03	-1.15	-.23	.60	1.45
A44	.38		**.38**			1.47	-2.69	-1.09	-.09	.95	1.61
A35				**.62**		1.27	-2.00	-1.01	-.17	.76	2.12
A63				**.54**		1.15	-2.27	-1.26	-.18	.59	1.70
A50				**.52**		1.06	-2.36	-1.08	.26	1.60	3.20

Item	Factor loadings					IRT parameters					
	1	2	3	4	5	*a*	b_1	B_2	b_3	b_4	b_5
A49				**.49**		1.62	-1.69	-.28	.80	1.75	2.32
A57		.33		**.42**		1.79	-1.80	-.73	.11	.80	1.70
A56				**.40**		1.30	-2.48	-.98	.03	1.04	2.32
A41		.31		**.34**		1.38	-1.59	-.34	.57	1.68	3.31
A52					**.68**	1.82	-2.30	-1.16	.27	.71	1.90
A54					**.51**	1.56	-2.27	-1.29	.03	1.13	1.93
A53					**.45**	1.50	-1.85	-.72	.05	.87	1.80
A12					**.40**	.96	-2.51	-.79	.66	1.81	3.85

Form B

Descriptive statistics

The number of responses, mean, standard deviation, skew and kurtosis statistics for each of the 63 items in Form B are shown in Table 11.

The mean rating on individual items ranged from a high of 4.82 (Item B41) to a low of 1.77 (Item B15), with a mean rating for all items on Form B of 3.72. Cronbach's alpha reliability for the whole questionnaire was $\alpha = .98$, indicating high internal consistency among the items on the Form. This means that for Form B, 97.5% of the observed score variance is due to differences in the performance of individuals, while 2.5% will be due to error. None of the items had a mean rating in excess of 5, providing further strength to the argument that students do not award high ratings capriciously.

Factor analysis

The 63 items of Form B were subjected to maximum likelihood factor analysis, with oblimin rotation. The Kaiser-Meyer-Olkin value was a "marvelous" .93, (Kaiser, 1974, p.35). Three interpretable factors were extracted, explaining 51.6% of the total variance. The goodness of fit statistic $\chi^2(1767) = 2205.6$, $p < .01$ indicates very good specification of the three factor model. The pattern matrix for these three factors is shown in Table 12.

Factor One has a strong Commitment to Students and Their Learning orientation; Factor Two makes links between the Family and the Community and their role in helping each student succeed in Mathematics; and, Factor Three describes Teaching for Student Engagement.

Factor One, Commitment to Students and Their Learning, (eigenvalue = 25.03) accounts for 39.7% of the common variance; Factor Two, Family and Community, (eigenvalue = 4.81) accounts for 7.6% of the variance; and, Factor Three, Teaching for Student Engagement (eigenvalue = 2.60) accounts for 4.1% of the common shared variance.

The absolute value of the correlations between factors range from .28 to .55. The correlation between the Factors One and Three exceeds .5. While the use of an oblique rotation allows for the factors to be correlated, this value is quite high. These two factors do not describe unique dimensions of exemplary mathematics teaching, but given that the items were all derived from statements from the same or overlapping sections of the Standards as was clearly noted by the participants in the Focus Groups in Study 1, this may have been expected. The correlations of Factor Three with both Factor One and Two are relatively small, indicating that when considered as pairs these factors measure a distinct (but somewhat overlapping) aspect of exemplary mathematics teaching.

Item Response Theory

The *a* and *b* threshold item parameters for Form B are shown in Table 13. The table also shows the CTT item-total (point biserial) correlations.

The *a* parameters for Form B were generally high, ranging from .35 (Item B15) to 2.30 (Item B51), with a mean *a* parameter for the 63 items of 1.38. The ten items with the highest discrimination indices are B51 (2.30), B13 (2.18), B22 (2.05), B48 (2.04), B23 (2.02), B61 (1.94), B47 (1.94), B54 (1.87), B14 (1.83), and, B33 (1.77). These items load on Factors One and Three, and reflect a disposition on the part of the teacher to build a positive attitude towards mathematics in all students, and make mathematics

Table 11 Descriptive Statistics for Form B

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
B01	*cares about and values each individual in the class.*	158	4.28	.10	1.29	-.51	.19	-.56	.38
B02	*shares ideas in an open and positive way.*	159	4.38	.10	1.21	-.39	.19	-.81	.38
B03	*is alert and sensitive to our individual differences.*	155	3.74	.10	1.40	-.16	.20	-.95	.39
B04	*respects the contributions we make in our maths class.*	158	4.39	.10	1.31	-.50	.19	-.52	.38
B05	*makes all topics in maths interesting.*	159	3.60	.12	1.50	-.16	.19	-.84	.38
B06	*has good judgment and displays discretion.*	156	4.22	.10	1.32	-.72	.19	-.02	.39
B07	*often has new teachers visit our classroom to observe their teaching.*	157	1.90	.11	1.31	1.62	.19	1.94	.39
B08	*demonstrates their concern for us through their actions and words.*	157	3.64	.10	1.19	.14	.19	-.70	.39
B09	*creates and maintains a learning environment by being flexible.*	156	4.14	.10	1.27	-.25	.19	-.90	.39
B10	*knows the students in this class really well.*	157	4.16	.10	1.43	-.50	.19	-.67	.39
B11	*determines and builds on each student's existing mathematical knowledge and understanding.*	159	4.05	.10	1.26	-.13	.19	-.84	.38
B12	*realises that not all students in the class have families who can assist them.*	157	3.69	.12	1.44	-.05	.19	-.83	.39
B13	*helps us to be confident in learning, doing and understanding maths.*	159	4.24	.12	1.45	-.53	.19	-.67	.38
B14	*understands and caters for students with different abilities in maths.*	157	3.99	.12	1.54	-.13	.19	-1.13	.39
B15	*often has teacher trainees in our classroom.*	159	1.77	.10	1.23	1.83	.19	2.76	.38
B16	*recognises that each student can obtain increased knowledge in maths.*	157	4.48	.10	1.27	-.67	.19	-.17	.39
B17	*teaches maths in a lively and enjoyable way.*	158	3.80	.13	1.61	-.32	.19	-.95	.38

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
B18	*uses interesting materials and resources that appeal to different people in the class.*	157	3.12	.12	1.49	.18	.19	-1.02	.39
B19	*knows what I can and can't do in maths.*	158	3.75	.12	1.50	-.09	.19	-1.04	.38
B20	*understands the impact that home life, cultural background, community expectations, and student attitudes can have on our learning.*	157	3.53	.10	1.39	-.01	.19	-.84	.39
B21	*allows us to make mistakes without feeling bad.*	158	4.52	.12	1.53	-.99	.19	.05	.38
B22	*creates a positive atmosphere in class where we feel part of a team of learners.*	156	3.94	.12	1.45	-.25	.19	-.75	.39
B23	*provides support and encouragement to all of the class.*	155	4.14	.12	1.50	-.33	.20	-.98	.39
B24	*holds my interest in class.*	157	3.53	.10	1.42	-.16	.19	-.99	.39
B25	*identifies and helps students with special needs or special abilities in maths*	157	4.15	.10	1.39	-.38	.19	-.73	.39
B26	*believes that all of the students in the class can learn and use significant mathematics.*	156	4.34	.11	1.31	-.55	.19	-.47	.39
B27	*empowers students to think through and solve problems both independently and together as a group.*	158	3.88	.11	1.38	-.17	.19	-.92	.38
B28	*recognises the beliefs and attitudes towards maths that each of us brings to the classroom.*	155	3.81	.10	1.22	-.21	.20	-.39	.39
B29	*uses their knowledge about each of us to create problems that are interesting and worth solving.*	156	3.36	.11	1.37	.06	.19	-.87	.39
B30	*involves us and our family in exploring career opportunities.*	156	2.23	.11	1.33	1.09	.19	.33	.39
B31	*focuses on the students in the class and their learning in mathematics.*	156	4.11	.11	1.38	-.39	.19	-.76	.39
B32	*understands and teaches according to the way that students learn maths.*	157	3.97	.12	1.48	-.51	.19	-.59	.39
B33	*is enthusiastic and enjoys teaching us maths.*	157	4.57	.12	1.45	-.96	.19	.07	.39

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
B34	*stimulates our learning by varying the way we are taught to allow for the strengths and weaknesses of the people in the class.*	155	3.57	.12	1.43	-.08	.20	-.74	.39
B35	*uses a variety of ways to encourage and involve our families in our maths work.*	158	2.39	.11	1.38	.97	.19	.14	.38
B36	*chooses approaches to teaching that work for all students in the class.*	158	3.46	.11	1.34	.10	.19	-.91	.38
B37	*works with other subject teachers to provide for every student in the class.*	154	2.74	.11	1.36	.64	.20	-.13	.39
B38	*allows us to learn maths in different ways*	158	3.70	.12	1.54	-.20	.19	-.94	.38
B39	*creates a welcoming environment that opens the class to family members and members of the community.*	157	2.82	.12	1.48	.57	.19	-.56	.39
B40	*is committed to the learning of all the students in the class.*	156	4.30	.11	1.33	-.45	.19	-.65	.39
B41	*expects students to respect the contributions of other students in the class.*	158	4.82	.09	1.18	-.86	.19	.23	.38
B42	*plays a part in keeping the community up to date with what is happening in maths.*	153	2.93	.12	1.51	.50	.20	-.82	.39
B43	*is committed to the principle of equity/fairness in the way they treat all people.*	155	4.23	.12	1.43	-.58	.20	-.44	.39
B44	*looks to my family for information about my strengths, interests, habits and home life.*	158	2.20	.11	1.40	1.12	.19	.36	.38
B45	*is able to explain something in different ways to help us understand.*	158	4.20	.12	1.55	-.54	.19	-.85	.38
B46	*provides a variety of options to allow for individual interests, aptitudes, knowledge and ways of learning.*	156	3.53	.11	1.37	-.12	.19	-.71	.39
B47	*makes maths come alive in the classroom.*	155	3.36	.13	1.56	.11	.20	-1.03	.39
B48	*makes learning maths satisfying and stimulating.*	155	3.40	.12	1.46	-.02	.20	-.96	.39

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
B49	*identifies and helps students with special abilities and special needs in maths, including those whose first language is not English.*	156	4.01	.12	1.54	-.35	.19	-.99	.39
B50	*listens to what the students have to say.*	158	4.62	.12	1.45	-.96	.19	-.04	.38
B51	*enables us to develop confidence and self esteem in maths.*	149	3.97	.11	1.35	-.41	.20	-.59	.40
B52	*understands the impact our individual backgrounds have on our learning.*	155	3.48	.11	1.36	.04	.20	-.81	.39
B53	*provides time to build on previous knowledge, interests and understandings.*	156	3.79	.11	1.34	-.23	.19	-.77	.39
B54	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.*	157	4.08	.12	1.44	-.35	.19	-.82	.39
B55	*keeps my family informed about my progress in maths.*	156	2.87	.11	1.42	.35	.19	-.99	.39
B56	*chooses a variety of approaches to teaching that work for the wide range of students in the class.*	157	3.32	.11	1.41	.01	.19	-.82	.39
B57	*has a classroom where students are respected and feel safe to participate.*	158	4.53	.11	1.35	-.76	.19	-.19	.38
B58	*involves families, administrators and teachers in the school and community to help and support student to learn and continue in maths.*	157	2.71	.11	1.34	.49	.19	-.69	.39
B59	*creates and maintains a learning environment by being well planned.*	158	4.51	.10	1.30	-.71	.19	-.11	.38
B60	*often attends and contributes to meetings of maths teachers.*	149	4.11	.11	1.37	-.31	.20	-.61	.40
B61	*makes everyone in the class believe that maths is for them.*	157	3.76	.12	1.53	-.17	.19	-1.09	.39
B62	*creates-an-environment-for us to become self-directed and capable of learning maths on our own.*	158	4.03	.10	1.31	-.28	.19	-.64	.38
B63	*compared with all other maths teachers I have had, is the best.*	158	3.81	.15	1.87	-.32	.19	-1.36	.38
	Valid N (listwise)	91							

Table 12 Summary of Items and Factor Loadings for Oblimin Three-Factor solution for Form B

		Factors			Used in	
Item	My mathematics teacher …	1	2	3	N	US
B50	*listens to what the students have to say.*	**.89**	-.01	.13		
B21	*allows us to make mistakes without feeling bad.*	**.82**	.03	.17		
B57	*has a classroom where students are respected and feel safe to participate.*	**.76**	-.03	.02		
B41	*expects students to respect the contributions of other students in the class.*	**.74**	-.11	-.02		
B04	*respects the contributions we make in our maths class.*	**.72**	-.02	.06		
B23	*provides support and encouragement to all of the class.*	**.71**	-.02	-.17		
B25	*identifies and helps students with special needs or special abilities in maths*	**.70**	.00	-.02		
B43	*is committed to the principle of equity/fairness in the way they treat all people.*	**.70**	.11	.11		
B26	*believes that all of the students in the class can learn and use significant mathematics.*	**.68**	-.14	.01		
B22	*creates a positive atmosphere in class where we feel part of a team of learners.*	**.67**	-.08	-.23	N13	US07
B51	*enables us to develop confidence and self esteem in maths.*	**.66**	.08	-.20	N11	US05
B13	*helps us to be confident in learning, doing and understanding maths.*	**.65**	.01	-.23		
B40	*is committed to the learning of all the students in the class.*	**.64**	-.07	-.15	N28	US20
B31	*focuses on the students in the class and their learning in mathematics.*	**.61**	.10	-.13		
B59	*creates and maintains a learning environment by being well planned.*	**.59**	-.13	-.14		
B01	*cares about and values each individual in the class.*	**.59**	.14	-.10		
B06	*has good judgment and displays discretion.*	**.57**	-.17	-.25		
B03	*is alert and sensitive to our individual differences.*	**.57**	.10	-.02		
B49	*identifies and helps students with special abilities and special needs in maths, including those whose first language is not English.*	**.56**	.17	-.14		
B54	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.*	**.56**	.01	-.26	N48	US36
B14	*understands and caters for students with different abilities in maths.*	**.55**	.09	-.26		
B61	*makes everyone in the class believe that maths is for them.*	**.54**	.16	-.21		
B08	*demonstrates their concern for us through their actions and words.*	**.54**	.11	.07		
B12	*realises that not all students in the class have families who can assist them.*	**.54**	.08	.09		

Item	My mathematics teacher …	Factors			Used in	
		1	2	3	N	US
B02	*shares ideas in an open and positive way.*	**.53**	.08	-.19		
B11	*determines and builds on each student's existing mathematical knowledge and understanding.*	**.50**	-.10	-.30		
B33	*is enthusiastic and enjoys teaching us maths.*	**.50**	-.15	-.40		
B52	*understands the impact our individual backgrounds have on our learning.*	**.50**	.29	-.04		
B20	*understands the impact that home life, cultural background, community expectations, and student attitudes can have on our learning.*	**.48**	.10	-.06	N10	
B16	*recognises that each student can obtain increased knowledge in maths.*	**.48**	-.13	-.26		
B60	*often attends and contributes to meetings of maths teachers.*	**.48**	.10	.21		
B27	*empowers students to think through and solve problems both independently and together as a group.*	**.46**	-.12	-.34	N25	US18
B53	*provides time to build on previous knowledge, interests and understandings.*	**.44**	.06	-.28		
B19	*knows what I can and can't do in maths.*	**.43**	.00	-.28		
B62	*creates-an-environment-for us to become self-directed and capable of learning maths on our own.*	**.41**	.05	-.31		
B28	*recognises the beliefs and attitudes towards maths that each of us brings to the classroom.*	**.41**	-.01	-.20		
B09	*creates and maintains a learning environment by being flexible.*	**.41**	-.02	-.29		
B63	*compared with all other maths teachers I have had, is the best*	**.40**	.08	-.35		
B30	*involves us and our family in exploring career opportunities.*	-.16	**.77**	-.13	N36	US27
B58	*involves families, administrators and teachers in the school and community to help and support student to learn and continue in maths.*	.13	**.73**	.04	N29	US21
B44	*looks to my family for information about my strengths, interests, habits and home life.*	.09	**.66**	-.10	N55	US41
B07	*often has new teachers visit our classroom to observe their teaching.*	-.08	**.65**	-.05		
B35	*uses a variety of ways to encourage and involve our families in our maths work.*	-.12	**.60**	-.31		
B55	*keeps my family informed about my progress in maths.*	.15	**.57**	.01	N64	US49
B15	*often has teacher trainees in our classroom.*	-.10	**.54**	-.03		
B37	*works with other subject teachers to provide for every student in the class.*	.19	**.54**	-.02	N53	US40
B39	*creates a welcoming environment that opens the class to family members and members of the community.*	.24	**.47**	-.01	N47	US35
B42	*plays a part in keeping the community up to date with what is happening in maths.*	.16	**.37**	-.16		
B17	*teaches maths in a lively and enjoyable way.*	-.01	.10	**-.81**		

Item	My mathematics teacher …	Factors 1	2	3	Used in N	US
B18	*uses interesting materials and resources that appeal to different people in the class.*	-.14	.30	**-.70**	N42	US31
B34	*stimulates our learning by varying the way we are taught to allow for the strengths and weaknesses of the people in the class.*	.05	.11	**-.68**	N37	
B47	*makes maths come alive in the classroom.*	.12	.20	**-.67**	N01	US01
B45	*is able to explain something in different ways to help us understand.*	.24	-.05	**-.65**	N30	US22
B05	*makes all topics in maths interesting.*	.09	.03	**-.65**		
B46	*provides a variety of options to allow for individual interests, aptitudes, knowledge and ways of learning.*	.00	.26	**-.63**		
B48	*makes learning maths satisfying and stimulating.*	.13	.25	**-.63**	N23	US16
B38	*allows us to learn maths in different ways*	.04	.22	**-.61**	N45	
B24	*holds my interest in class.*	.22	.05	**-.60**	N22	
B29	*uses their knowledge about each of us to create problems that are interesting and worth solving.*	.12	.13	**-.56**	N54	
B32	*understands and teaches according to the way that students learn maths.*	.28	.07	**-.49**		
B36	*chooses approaches to teaching that work for all students in the class.*	.22	.10	**-.49**		
B56	*chooses a variety of approaches to teaching that work for the wide range of students in the class.*	.30	.23	**-.41**		
B10	*knows the students in this class really well.*	.28	.01	**-.38**		

Factor correlations	**1**	**2**	**3**
Factor 1: Commitment to Students and their Learning	--		
Factor 2: Family and Community	.28	--	
Factor 3: Teaching for Student Engagement	-.55	-.36	--

N and US. Items prefixed with US were in the final instrument (SEAT-M) used in the USA. Prefix N indicates that they were further trialled in the November instrument, but did not make the SEAT-M instrument.

Explains 51.6% of variance

Cronbach's alpha = ..975

Table 13 Classical Test Theory and Item Response Theory Item Statistics for 63 Form B Items

Scale	Item #	Item Mean	Item=total correlation r_{pbs}	*a* (SE)	b_1 (SE)	b_2 (SE)	b_3 (SE)	b_4 (SE)	b_5 (SE)
	B01	4.28 (0.10)	.665	1.57 (0.24)	-3.40 (0.61)	-1.95 (0.30)	-0.85 (0.19)	-0.06 (0.16)	1.44 (0.24)
	B02	4.38 (0.10)	.683	1.53 (0.24)	-7.27 (**)	-2.27 (0.61)	-1.06 (0.23)	-0.08 (0.21)	1.44 (0.23)
	B03	3.74 (0.11)	.619	1.25 (0.22)	-3.00 (0.61)	-1.41 (0.29)	-0.35 (0.23)	0.65 (0.22)	2.27 (0.42)
	B04	4.39 (0.10)	.600	1.21 (0.22)	-4.00 (0.87)	-2.49 (0.46)	-1.13 (0.28)	0.02 (0.20)	1.28 (0.28)
	B05	3.60 (0.12)	.647	1.41 (0.21)	-1.90 (0.33)	-1.08 (0.23)	-0.05 (0.18)	0.96 (0.22)	2.00 (0.35)
	B06	4.22 (0.11)	.660	1.52 (0.23)	-2.63 (0.44)	-1.92 (0.30)	-0.91 (0.22)	0.11 (0.16)	1.65 (0.27)
	B07	1.90 (0.10)	.317	0.53 (0.22)	0.33 (0.43)	2.59 (1.12)	3.87 (1.60)	5.18 (2.22)	6.78 (3.14)
	B08	3.64 (0.10)	.461	1.01 (0.21)	-4.97 (1.47)	-1.81 (0.45)	-0.14 (0.26)	1.38 (0.34)	3.11 (0.71)
	B09	4.14 (0.10)	.609	1.29 (0.23)	-4.66 (1.16)	-2.07 (0.36)	-0.78 (0.23)	0.24 (0.20)	1.79 (0.33)
	B10	4.16 (0.11)	.535	1.11 (0.23)	-3.34 (0.70)	-1.87 (0.39)	-0.93 (0.25)	0.13 (0.23)	1.68 (0.37)
	B11	4.05 (0.10)	.659	1.46 (0.24)	-3.82 (0.77)	-1.99 (0.31)	-0.52 (0.19)	0.46 (0.19)	1.78 (0.30)
	B12	3.69 (0.11)	.425	0.86 (0.20)	-3.42 (0.85)	-1.84 (0.46)	-0.24 (0.28)	1.07 (0.36)	2.59 (0.63)
	B13	4.24 (0.11)	.799	2.18 (0.30)	-2.42 (0.30)	-1.52 (0.20)	-0.76 (0.15)	0.02 (0.15)	1.07 (0.17)
	B14	3.99 (0.12)	.750	1.83 (0.29)	-2.53 (0.36)	-1.31 (0.19)	-0.37 (0.17)	0.40 (0.15)	1.04 (0.19)
	B15	1.77 (0.10)	.215	0.35 (0.23)	1.26 (1.01)	4.59 (2.92)	5.93 (3.64)	8.69 (5.94)	10.78 (7.80)
	B16	4.48 (0.10)	.595	1.21 (0.23)	-3.98 (0.88)	-2.56 (0.49)	-1.44 (0.30)	-0.24 (0.21)	1.26 (0.29)
	B17	3.80 (0.13)	.701	1.73 (0.23)	-1.76 (0.26)	-1.03 (0.21)	-0.50 (0.18)	0.63 (0.18)	1.42 (0.22)
US31	B18	3.12 (0.12)	.592	1.24 (0.22)	-1.74 (0.37)	-0.41 (0.22)	0.37 (0.21)	1.48 (0.30)	2.79 (0.57)
	B19	3.75 (0.12)	.607	1.33 (0.23)	-2.67 (0.46)	-1.15 (0.23)	-0.20 (0.20)	0.73 (0.20)	1.74 (0.32)
N10	B20	3.53 (0.11)	.506	1.10 (0.19)	-2.74 (0.54)	-1.33 (0.29)	0.06 (0.23)	1.08 (0.28)	2.76 (0.53)
	B21	4.52 (0.12)	.614	1.37 (0.23)	-2.33 (0.40)	-1.94 (0.32)	-1.23 (0.24)	-0.43 (0.19)	0.77 (0.23)
US07	B22	3.94 (0.12)	.726	2.05 (0.28)	-2.25 (0.31)	-1.38 (0.22)	-0.39 (0.15)	0.55 (0.14)	1.29 (0.18)
	B23	4.14 (0.12)	.758	2.02 (0.29)	-2.47 (0.33)	-1.46 (0.20)	-0.44 (0.15)	0.18 (0.16)	1.06 (0.16)
N22	B24	3.53 (0.11)	.747	1.73 (0.26)	-2.05 (0.31)	-0.92 (0.18)	-0.12 (0.16)	0.78 (0.18)	2.29 (0.37)
	B25	4.15 (0.11)	.599	1.45 (0.25)	-3.13 (0.59)	-1.78 (0.30)	-0.81 (0.21)	0.24 (0.20)	1.46 (0.26)
	B26	4.34 (0.11)	.522	1.13 (0.20)	-3.80 (0.75)	-2.50 (0.45)	-1.03 (0.27)	-0.08 (0.23)	1.57 (0.35)
US18	B27	3.88 (0.11)	.624	1.40 (0.23)	-3.22 (0.56)	-1.44 (0.27)	-0.42 (0.21)	0.56 (0.19)	1.85 (0.32)
	B28	3.81 (0.10)	.511	1.06 (0.22)	-3.54 (0.71)	-2.19 (0.44)	-0.40 (0.25)	1.02 (0.29)	2.91 (0.61)
N54	B29	3.36 (0.11)	.648	1.38 (0.23)	-2.38 (0.41)	-0.78 (0.22)	0.06 (0.19)	1.29 (0.24)	2.58 (0.47)
US27	B30	2.23 (0.11)	.357	0.54 (0.19)	-1.20 (0.60)	1.67 (0.69)	2.95 (1.02)	4.36 (1.45)	7.56 (2.94)
	B31	4.11 (0.11)	.744	1.67 (0.24)	-2.89 (0.43)	-1.58 (0.24)	-0.65 (0.18)	0.23 (0.16)	1.47 (0.23)
	B32	3.97 (0.12)	.731	1.68 (0.26)	-2.11 (0.33)	-1.38 (0.25)	-0.63 (0.18)	0.37 (0.16)	1.54 (0.25)
	B33	4.57 (0.12)	.666	1.77 (0.28)	-2.53 (0.40)	-1.75 (0.28)	-1.23 (0.20)	-0.36 (0.15)	0.70 (0.17)
N27	B34	3.57 (0.12)	.707	1.58 (0.23)	-2.08 (0.37)	-1.14 (0.24)	-0.13 (0.17)	0.99 (0.21)	1.97 (0.30)
	B35	2.39 (0.11)	.456	0.77 (0.21)	-1.24 (0.47)	0.87 (0.35)	1.87 (0.52)	3.26 (0.87)	4.56 (1.33)
	B36	3.46 (0.11)	.596	1.53 (0.22)	-2.55 (0.37)	-0.93 (0.21)	0.19 (0.17)	0.94 (0.20)	2.45 (0.40)

Scale	Item #	Item Mean	Item=total correlation r_{pbs}	a (SE)	b_1 (SE)	b_2 (SE)	b_3 (SE)	b_4 (SE)	b_5 (SE)
US40	B37	2.74 (0.11)	.441	0.77 (0.21)	-2.17 (0.66)	-0.19 (0.31)	1.50 (0.45)	3.09 (0.83)	4.12 (1.13)
N45	B38	3.70 (0.12)	.627	1.44 (0.22)	-2.31 (0.38)	-1.07 (0.22)	-0.26 (0.20)	0.67 (0.19)	1.80 (0.31)
US35	B39	2.82 (0.12)	.506	0.84 (0.22)	-1.91 (0.58)	-0.11 (0.29)	0.94 (0.33)	2.49 (0.62)	3.41 (0.87)
US20	B40	4.30 (0.11)	.685	1.70 (0.27)	-3.32 (0.54)	-1.92 (0.31)	-0.87 (0.19)	0.00 (0.16)	1.18 (0.21)
	B41	4.82 (0.09)	.614	1.41 (0.25)	-3.92 (0.80)	-3.18 (0.56)	-1.68 (0.28)	-0.62 (0.20)	0.64 (0.20)
	B42	2.93 (0.12)	.543	0.91 (0.19)	-2.07 (0.53)	-0.15 (0.27)	0.88 (0.32)	1.82 (0.45)	3.33 (0.82)
	B43	4.23 (0.11)	.592	1.24 (0.21)	-2.84 (0.52)	-2.07 (0.37)	-0.83 (0.23)	0.07 (0.20)	1.40 (0.30)
US41	B44	2.20 (0.11)	.498	0.95 (0.25)	-0.38 (0.29)	0.98 (0.31)	1.92 (0.47)	2.76 (0.65)	4.04 (1.06)
US22	B45	4.20 (0.12)	.719	1.75 (0.27)	-2.47 (0.34)	-1.30 (0.22)	-0.74 0.18)	0.01 (0.15)	1.08 (0.20)
	B46	3.53 (0.11)	.667	1.52 (0.23)	-2.22 (0.35)	-1.07 (0.21)	-0.11 (0.19)	1.09 (0.22)	2.29 (0.42)
US01	B47	3.36 (0.13)	.764	1.94 (0.27)	-1.53 (0.22)	-0.64 (0.15)	0.17 (0.17)	0.93 (0.16)	1.73 (0.24)
US16	B48	3.40 (0.12)	.776	2.04 (0.27)	-1.71 (0.27)	-0.75 (0.15)	0.02 (0.16)	0.94 (0.16)	2.00 (0.27)
	B49	4.01 (0.12)	.686	1.72 (0.23)	-2.31 (0.36)	-1.26 (0.20)	-0.45 (0.17)	0.20 (0.16)	1.29 (0.22)
	B50	4.62 (0.12)	.738	1.75 (0.27)	-2.57 (0.40)	-1.79 (0.26)	-1.02 (0.18)	-0.51 (0.15)	0.61 (0.18)
US05	B51	3.97 (0.11)	.758	2.30 (0.31)	-2.42 (0.27)	-1.44 (0.20)	-0.56 (0.16)	0.40 (0.14)	1.62 (0.22)
	B52	3.48 (0.11)	.604	1.35 (0.22)	-2.61 (0.43)	-1.03 (0.23)	-0.03 (0.19)	1.15 (0.25)	2.47 (0.43)
	B53	3.79 (0.11)	.699	1.53 (0.24)	-2.63 (0.42)	-1.50 (0.24)	-0.23 (0.19)	0.62 (0.20)	2.20 (0.37)
US36	B54	4.08 (0.11)	.685	1.87 (0.24)	-2.56 (0.33)	-1.44 (0.24)	-0.48 (0.16)	0.32 (0.16)	1.31 (0.19)
US49	B55	2.87 (0.11)	.409	0.73 (0.19)	-2.16 (0.63)	-0.15 (0.31)	1.18 (0.43)	2.25 (0.63)	5.76 (1.57)
	B56	3.32 (0.11)	.749	1.72 (0.26)	-1.79 (0.29)	-0.82 (0.19)	0.12 (0.15)	1.09 (0.20)	2.29 (0.36)
	B57	4.53 (0.11)	.683	1.49 (0.24)	-3.08 (0.55)	-2.21 (0.34)	-1.05 (0.21)	-0.29 (0.17)	0.94 (0.21)
US21	B58	2.71 (0.11)	.453	0.75 (0.21)	-2.07 (0.64)	0.08 (0.30)	1.35 (0.44)	2.93 (0.82)	5.62 (1.93)
	B59	4.51 (0.10)	.602	1.30 (0.22)	-3.49 (0.65)	-2.44 (0.40)	-1.29 (0.26)	-0.20 (0.20)	1.11 (0.27)
	B60	4.11 (0.11)	.294	0.55 (0.20)	-6.06 (2.23)	-3.87 (1.50)	-1.43 (0.66)	0.79 (0.51)	2.80 (1.08)
	B61	3.76 (0.12)	.756	1.94 (0.27)	-2.11 (0.31)	-0.91 (0.18)	-0.13 (0.15)	0.50 (0.15)	1.55 (0.21)
	B62	4.03 (0.10)	.634	1.53 (0.21)	-3.07 (0.51)	-1.84 (0.31)	-0.58 (0.19)	0.41 (0.18)	1.71 (0.28)
	B63	3.81 (0.15)	.705	1.53 (0.27)	-1.31 (0.28)	-0.72 (0.21)	-0.35 (0.19)	0.30 (0.19)	1.11 (0.21)

Note: SE is in brackets

a lively subject. These items also had the excellent item-total correlations (point biserial correlations) in excess of .67.

The *b* parameters had the following ranges: $-7.27<b_1<1.26$; $-3.87<b_2<4.59$; $-1.68<b_3<5.93$; $-0.62<b_4<8.69$; $0.61<b_5<10.78$. Item B15 accounted for the all of the unusually large values for the upper limits of b_i. With Item B15 removed, the upper bound of the *b* parameters was as follows: $b_1 <0.33$; $b_2 <2.59$; $b_3 <3.87$; $b_4 <5.18$; and, $b_5 <7.56$. Apart from the upper limit for b_5, all of these upper limits were for Item B07.

Item selection from Form B

Through successive iterations, items were eliminated. These items either had inadequate IRT characteristics, or a weak loading on the interpretable factors.

Table 14 shows the set of twenty items selected from Form B with their original factor loadings and IRT characteristics. The items are arranged according to the factor analysis.

Table 14 Factor loadings and IRT parameters for 20 items selected from Form B

Item	Factor loadings			IRT parameters					
	1	2	3	*a*	b_1	b_2	b_3	b_4	b_5
B22	**.67**			2.05	-2.25	-1.38	-.39	.55	1.29
B51	**.66**			2.30	-2.42	-1.44	-.56	.40	1.62
B40	**.64**			1.70	-3.32	-1.92	-.87	.00	1.18
B54	**.56**			1.87	-2.56	-1.44	-.48	.32	1.31
B20	**.48**			1.10	-2.74	-1.33	.06	1.08	2.76
B27	**.46**			1.40	-3.22	-1.44	-.42	.56	1.85
B30		**.77**		.54	-1.20	1.67	2.95	4.36	7.56
B58		**.73**		.75	-2.07	.08	1.35	2.93	5.62
B44		**.66**		.95	-.38	.98	1.92	2.76	4.04
B55		**.57**		.73	-2.16	-.15	1.18	2.25	5.76
B37		**.54**		.77	-2.17	-.19	1.50	3.09	4.12
B39		**.47**		.84	-1.91	-.11	.94	2.49	3.41
B18			**.70**	1.24	-1.74	-.41	.37	1.48	2.79

B34	**.68**	1.58	-2.08	-1.14	-.13	.99	1.97
B47	**.67**	1.94	-1.53	-.64	.17	.93	1.73
B45	**.65**	1.75	-2.47	-1.30	-.74	.01	1.08
B48	**.63**	2.04	-1.71	-.75	.02	.94	2.00
B38	**.61**	1.44	-2.31	-1.07	-.26	.67	1.80
B24	**.60**	1.73	-2.05	-.92	-.12	-.78	2.29
B29	**.56**	1.38	-2.38	-.78	.06	1.29	2.58

The most notable feature of the selected Form B items is the relatively poor IRT characteristics (especially the low discrimination) of the items in Factor Two (Family and Community). A strong message that emerged from the focus groups in Study One was that this was one aspect of the NBPTS Standards that was not strongly developed in New Zealand schools. Teachers had indicated this in their response to the Standards, and this was borne out by the students in their response to the items in the questionnaire. These items would have been eliminated on the basis of their IRT characteristics, but the low correlation between this factor and the other two are an indication that this is a distinct factor worthy of inclusion. Furthermore, these items represent an important element of the Standards, and any mapping of the retained items against the Standards would have revealed this gap in the questionnaire. The six items that have been retained represent the best items in terms of their loading on the factor, their IRT characteristics, and the mapping of the questionnaire items against the Standards. Any revision of the SEAT-M for New Zealand would need to reconsider the relevance of these items to a local definition of highly accomplished teaching.

All of the retained items have strong loadings on their respective factor (all greater than .46) and with the exception of Factor Two, have very strong IRT parameters as well.

Form C

Descriptive statistics

The number of responses, mean, standard deviation, skew and kurtosis statistics for each of the 68 items in Form C is shown in Table 15.

Table 15 Descriptive Statistics for Form C

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
C01	*makes maths meaningful for me.*	154	3.90	.10	1.26	-.29	.20	-.60	.39
C02	*places a high value on learning maths.*	153	4.39	.09	1.16	-.59	.20	-.19	.39
C03	*focuses all of the students on their work.*	150	4.06	.11	1.37	-.61	.20	-.22	.39
C04	*is able to use many different ways to get mathematical ideas across, like words, stories, numbers, diagrams, graphs and symbols.*	154	3.99	.12	1.52	-.42	.20	-.81	.39
C05	*chooses imaginative examples, problems and situations that motivate us.*	153	3.67	.12	1.46	-.12	.20	-.97	.39
C06	*uses group investigations to assess us.*	152	2.51	.11	1.40	.67	.20	-.45	.39
C07	*uses examples from a wide range of fields to show how maths is related and useful.*	154	3.55	.11	1.42	-.13	.20	-.98	.39
C08	*explores ideas with us even if the answer is not known in advance.*	152	3.99	.10	1.28	-.40	.20	-.42	.39
C09	*integrates the goals of the curriculum and their knowledge of the students in the class.*	141	3.95	.11	1.30	-.30	.20	-.37	.41
C10	*gives us the chance to sensitively assess other students' work.*	152	2.80	.12	1.48	.36	.20	-.98	.39
C11	*is not afraid of failure.*	152	4.20	.12	1.45	-.42	.20	-.82	.39
C12	*applies concepts in realistic settings.*	153	4.01	.10	1.24	-.39	.20	-.28	.39
C13	*presents new ideas they have found in journals and at conferences and meetings to help us expand our learning of maths.*	151	3.12	.12	1.52	.31	.20	-.93	.39
C14	*uses examples from other school subjects and the outside world to help us understand new ideas in maths.*	153	3.24	.11	1.41	.19	.20	-.97	.39

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
C15	*consistently makes decisions about their teaching that will further our learning.*	152	3.66	.10	1.22	-.16	.20	-.59	.39
C16	*teaches us high quality, important and meaningful maths.*	154	4.44	.10	1.22	-.71	.20	.39	.39
C17	*encourages us to advance in maths as far as possible,*	154	4.52	.10	1.28	-.58	.20	-.56	.39
C18	*is fair in the way they assess each student in the class.*	153	4.67	.10	1.29	-.71	.20	-.41	.39
C19	*tells us that we are expected to do well in maths.*	154	4.08	.12	1.48	-.42	.20	-.78	.39
C20	*uses a wide variety of resources (e.g., speakers, historical material, the library, museum visits , etc.) to help us reach our mathematical goals.*	153	2.14	.10	1.19	1.00	.20	.32	.39
C21	*motivates us to do our best work.*	154	4.10	.11	1.37	-.26	.20	-.98	.39
C22	*teaches us meaningful and important maths.*	153	4.42	.10	1.27	-.44	.20	-.65	.39
C23	*provides the inspiration for student investigations.*	145	3.35	.11	1.28	-.11	.20	-.66	.40
C24	*uses items that are in the news and relates them to our classwork.*	152	2.78	.12	1.53	.61	.20	-.67	.39
C25	*gathers information from us and uses it to improve their teaching.*	154	2.82	.11	1.42	.34	.20	-.87	.39
C26	*checks for student understanding before and at the end of each lesson.*	154	4.03	.13	1.59	-.45	.20	-.92	.39
C27	*helps us experience success in doing worthwhile maths.*	154	3.94	.11	1.35	-.28	.20	-.83	.39
C28	*models their own mathematical reasoning in all tasks, actions and discussions.*	154	3.97	.10	1.26	-.10	.20	-.83	.39
C29	*knows and caters for the problems we commonly encounter in learning new topics.*	152	4.32	.10	1.23	-.44	.20	-.27	.39
C30	*provides tasks that challenge us to think..*	154	4.77	.09	1.16	-1.11	.20	1.00	.39
C31	*ensures that all students take courses that lead to increased mathematical knowledge.*	152	3.47	.11	1.38	-.06	.20	-.77	.39
C32	*encourages us to place a high value on maths.*	154	4.00	.10	1.29	-.28	.20	-.72	.39
C33	*encourages us to set high goals for ourselves in maths.*	152	4.22	.11	1.40	-.54	.20	-.59	.39

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
C34	*recognises and overcomes the barriers that prevent students from achieving success in maths.*	152	3.88	.11	1.33	-.41	.20	-.54	.39
C35	*emphasises the points we are expected to understand and learn.*	152	4.69	.10	1.20	-.81	.20	.17	.39
C36	*uses a variety of techniques to maintain control of the students in this class.*	154	3.73	.11	1.41	-.26	.20	-.89	.39
C37	*provides useful feedback after each assessment.*	154	4.17	.10	1.29	-.32	.20	-.61	.39
C38	*takes calculated risks with the way a lesson might develop if the outcome might be beneficial.*	149	3.36	.11	1.34	-.16	.20	-.85	.40
C39	*skillfully combines their knowledge of adolescents, mathematics and how we learn to help us be successful in maths.*	150	3.51	.11	1.36	-.10	.20	-.82	.39
C40	*prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.*	152	3.39	.11	1.36	.05	.20	-.81	.39
C41	*helps us to use various performance measures to monitor our progress in maths.*	149	3.34	.10	1.26	.32	.20	-.33	.40
C42	*uses well defined goals to assess our work and learning.*	150	3.67	.11	1.28	-.26	.20	-.46	.39
C43	*asks questions and uses skilful probing to help classroom discussion and thinking.*	150	4.04	.11	1.35	-.31	.20	-.69	.39
C44	*expects us to learn maths even if we have different backgrounds and previous learning experiences.*	152	4.55	.11	1.33	-.72	.20	-.27	.39
C45	*illustrates the way that different cultures have contributed to the development of mathematics.*	148	2.94	.13	1.58	.31	.20	-1.02	.40
C46	*seems to modify their plans for the lesson if something interesting comes up.*	151	3.50	.12	1.51	-.11	.20	-1.07	.39
C47	*provides enough work to keep all students in the class working.*	152	4.96	.10	1.27	-1.26	.20	.81	.39
C48	*gives us time to understand new ideas and progress to the next level of understanding.*	152	4.02	.13	1.56	-.43	.20	-.90	.39

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
C49	*intervenes when appropriate to help a student gain better understanding.*	148	4.30	.11	1.39	-.58	.20	-.52	.40
C50	*uses a blend of new and traditional methods to teach us.*	152	3.68	.11	1.39	-.18	.20	-.68	.39
C51	*keeps the interest of all the students in the class.*	149	3.52	.12	1.51	-.26	.20	-.93	.40
C52	*uses cooperative learning strategies and group work to help us learn and tackle substantial mathematical issues.*	150	3.01	.11	1.40	.24	.20	-.75	.39
C53	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.*	152	4.12	.11	1.32	-.54	.20	-.31	.39
C54	*does not claim to have all of the answers.*	149	3.87	.13	1.57	-.22	.20	-1.07	.40
C55	*uses assessment results to provide extra help/extension to appropriate students.*	149	3.82	.12	1.49	-.28	.20	-.85	.40
C56	*identifies individual strengths and weaknesses after each assessment.*	151	3.64	.12	1.47	-.07	.20	-1.04	.39
C57	*uses examples from the history of mathematics to illustrate its development.*	151	2.92	.12	1.49	.51	.20	-.61	.39
C58	*teaches us equally well in all strands of the mathematics curriculum (algebra, number, measurement, geometry, etc).*	151	4.65	.11	1.32	-.78	.20	.09	.39
C59	*uses an appropriate range of formal and informal assessments to monitor individual and class progress.*	149	3.93	.10	1.27	-.22	.20	-.53	.40
C60	*adjusts the lesson if we experience difficulties in learning.*	152	3.84	.12	1.50	-.24	.20	-1.04	.39
C61	*tells us what the purpose of each lesson is.*	152	3.55	.12	1.50	-.15	.20	-1.02	.39
C62	*gets us to think about the nature and quality of our work.*	151	3.68	.11	1.34	-.14	.20	-.74	.39
C63	*teaches us how to evaluate progress towards our goals.*	152	3.48	.11	1.36	.00	.20	-.72	.39
C64	*uses examples that help us to understand and learn new ideas.*	152	4.47	.11	1.39	-.72	.20	-.34	.39
C65	*makes good use of time to optimise learning.*	152	4.33	.11	1.33	-.69	.20	-.12	.39
C66	*is fair in the way they assess all students.*	152	4.76	.10	1.27	-1.04	.20	.64	.39
C67	*encourages us to take risks and make mistakes.*	151	3.91	.12	1.46	-.32	.20	-.84	.39

	My mathematics teacher …	N	Mean		Std. Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
C68	*compared with all other maths teachers I have had, is the best.*	151	3.83	.14	1.73	-.32	.20	-1.19	.39
	Valid N (listwise)	101							

The mean rating on individual items ranged from a high of 4.96 (Item C47) to a low of 2.14 (Item C20), with a mean rating for all items on Form C of 3.83. Cronbach's alpha reliability for the whole test was $\alpha = .97$, indicating high internal consistency among the items on the Form. This means that for Form A, 97.3% of the observed score variance is due to differences in the performance of individuals, while 2.7% will be due to error. As occurred with the other two Forms trialled at this time, none of the items had a mean rating in excess of 5, which provides support to the argument that students do not award high ratings capriciously, and award their ratings judiciously.

Factor analysis

The 68 items of Form C were subjected to maximum likelihood factor analysis, with oblimin rotation. The Kaiser-Meyer-Olkin measure of sampling adequacy value was a "meritorious" .89 (Kaiser, 1974, p. 35). Five interpretable factors were extracted, explaining 50.2% of the total variance. The goodness of fit statistic $\chi^2(1948) = 2233.76$, $p < .01$ indicates very good specification of the five factor model. The pattern matrix for these five factors is shown in Table 16.

Factor One indicates that the teacher Varies the Lesson; Factor Two describes how the teacher Manages the Learning Environment; Factor Three refers to Commitment to Students and Their Learning; Factor Four describes Reasoning and Thinking Mathematically; and, Factor Five describes how the teacher Relates Mathematics to the Real World.

Factor One, Varies the Lesson, (eigenvalue = 23.96) accounts for 35.2% of the common variance; Factor Two, Manages the Learning Environment, (eigenvalue = 3.83) accounts for 5.6% of the variance; Factor Three, Commitment to Students and Their Learning (eigenvalue = 2.75) accounts for 4.0% of the common shared variance; Factor Four, Reasoning and Thinking Mathematically (eigenvalue = 1.94) accounts for 2.9% of the common shared variance; and, Factor Five, Relates Mathematics to the Real World (eigenvalue = 1.67) accounts for 2.5% of the shared variance.

The absolute value of the correlations between factors range from .14 to .53. The correlation between the Factors One and Five exceeds -.5. While the use of an oblique

rotation allows for the factors to be correlated, this value is quite high. These two factors do not describe unique dimensions of exemplary mathematics teaching, but given that the items were all derived from statements from the same or overlapping sections of the Standards as was clearly noted by the participants in the Focus Groups in Study 1, this may have been expected. The remaining correlations are relatively small, indicating that when considered as pairs these factors measure a distinct (but somewhat overlapping) aspect of exemplary mathematics teaching.

Item Response Theory

Table 17 shows the IRT parameters for the 67 items in Form C. The *a* parameters for From C were high, ranging from a minimum value of .42 (Item C19) to a maximum of 2.89 (Item C27), with a mean *a* parameter for the 67 items of 1.70. The *a* value of 2.89 for C27 was the highest of any item on the three forms.

The discrimination indices for the items on this form were higher than on the other two forms, with the *a* values for fifteen items exceeding two. They were C27 (2.89), C50 (2.70), C63 (2.62), C42 (2.52), C34 (2.49), C55 (2.35), C39 (2.33), C21 (2.29), C05 (2.28), C64 (2.28), C60 (2.25), C53 (2.23), C23 (2.12), C15 (2.09), and, C52 (2.06). These items are spread across the first four of the factors, without a consistent pattern. These items also had excellent item-total correlations, all in excess of .66.

Item selection from Form C

The same process was used to select items as in the previous two forms. Items with poor IRT characteristics, or a weak loading on the interpretable factors were considered for elimination. At the same time, the deletion of items was moderated by the need to maintain an adequate mapping of the NBPTS Standards document.

Table 18 shows the set of twenty-one items selected from Form C with their original factor loadings and IRT characteristics. The items are arranged according to the factor analysis, with factor loadings shown only for those above .30.

Table 16 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form C

		Factors					Used in	
Item	My mathematics teacher …	1	2	3	4	5	N	US
C20	*uses a wide variety of resources (e.g., speakers, historical material, the library, museum visits , etc.) to help us reach our mathematical goals.*	**.72**	-.06	.04	.04	.02		
C25	*gathers information from us and uses it to improve their teaching.*	**.62**	.06	-.09	.07	.03		
C10	*gives us the chance to sensitively assess other students' work.*	**.60**	-.18	.04	.11	-.24		
C24	*uses items that are in the news and relates them to our classwork.*	**.54**	-.09	-.02	.12	-.22		
C23	*provides the inspiration for student investigations.*	**.53**	.12	-.07	.15	-.11	N37	
C52	*uses cooperative learning strategies and group work to help us learn and tackle substantial mathematical issues.*	**.52**	.08	-.00	.13	-.20		
C57	*uses examples from the history of mathematics to illustrate its development.*	**.51**	.19	-.15	-.17	-.02		
C06	*uses group investigations to assess us.*	**.48**	-.05	.05	-.01	-.23		
C45	*illustrates the way that different cultures have contributed to the development of mathematics.*	**.47**	.10	-.17	-.26	-.14		
C27	*helps us experience success in doing worthwhile maths.*	**.46**	.21	-.16	.32	-.02	N56	
C67	*encourages us to take risks and make mistakes.*	**.41**	.16	-.09	.28	.06		
C38	*takes calculated risks with the way a lesson might develop if the outcome might be beneficial.*	**.40**	.18	-.13	.04	-.08		
C61	*tells us what the purpose of each lesson is.*	**.39**	.33	-.03	.02	-.05	N60	US45
C63	*teaches us how to evaluate progress towards our goals.*	**.37**	.24	-.36	.04	-.04	N38	US28
C41	*helps us to use various performance measures to monitor our progress in maths.*	**.37**	.03	-.15	.24	-.10		
C50	*uses a blend of new and traditional methods to teach us.*	**.36**	.25	-.08	.12	-.21		
C34	*recognises and overcomes the barriers that prevent students from achieving success in maths.*	**.29**	.24	-.25	.03	-.19		
C62	*gets us to think about the nature and quality of our work.*	**.27**	.12	-.26	.00	-.24	N59	US44
C58	*teaches us equally well in all strands of the mathematics curriculum (algebra, number, measurement, geometry, etc).*	-.18	**.70**	-.07	.05	-.05		
C66	*is fair in the way they assess all students.*	.07	**.67**	-.06	.14	.14		

Item	My mathematics teacher …	Factors					Used in	
		1	2	3	4	5	N	US
C65	*makes good use of time to optimise learning.*	.08	**.66**	.13	.00	-.22		
C18	*is fair in the way they assess each student in the class.*	-.01	**.63**	.02	.20	.16		
C49	*intervenes when appropriate to help a student gain better understanding.*	.00	**.55**	-.08	.04	-.20		
C48	*gives us time to understand new ideas and progress to the next level of understanding.*	.36	**.54**	.13	-.16	-.16		
C16	*teaches us high quality, important and meaningful maths.*	-.09	**.54**	-.21	.05	-.18		
C64	*uses examples that help us to understand and learn new ideas.*	.19	**.53**	.12	-.14	-.40		
C53	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.*	.12	**.51**	.08	-.04	-.38	N31	US23
C60	*adjusts the lesson if we experience difficulties in learning.*	.28	**.47**	-.14	-.08	-.08	N14	US08
C68	*compared with all other maths teachers I have had, is the best.*	.17	**.46**	-.13	-.04	.14		
C35	*emphasises the points we are expected to understand and learn.*	-.06	**.45**	-.35	.07	.05		
C29	*knows and caters for the problems we commonly encounter in learning new topics.*	-.04	**.44**	-.07	.22	-.24	N44	US33
C56	*identifies individual strengths and weaknesses after each assessment.*	.31	**.43**	-.24	.00	.10		
C01	*makes maths meaningful for me.*	.11	**.42**	-.20	-.03	-.21	N41	
C54	*does not claim to have all of the answers.*	.06	**.37**	.17	.25	-.04		
C22	*teaches us meaningful and important maths.*	.10	**.36**	-.32	-.12	-.17		
C59	*uses an appropriate range of formal and informal assessments to monitor individual and class progress.*	.13	**.35**	-.18	.09	.02	N03	
C47	*provides enough work to keep all students in the class working.*	-.18	**.34**	-.09	.18	.03		
C51	*keeps the interest of all the students in the class.*	.28	**.34**	.04	-.13	-.33		
C26	*checks for student understanding before and at the end of each lesson.*	.24	**.32**	.04	.32	.10		
C37	*provides useful feedback after each assessment.*	.08	**.29**	-.12	.21	-.18		
C32	*encourages us to place a high value on maths.*	.05	.01	**-.82**	.06	-.05	N46	US34
C33	*encourages us to set high goals for ourselves in maths.*	.06	-.01	**-.72**	-.07	-.13		
C17	*encourages us to advance in maths as far as possible,*	-.04	.17	**-.61**	.15	.03		
C44	*expects us to learn maths even if we have different backgrounds and previous learning experiences.*	-.16	-.05	**-.46**	.26	.05		
C40	*prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.*	.37	-.01	**-.45**	.04	-.05	N49	US37

Item	My mathematics teacher …	Factors					Used in	
		1	2	3	4	5	N	US
C42	*uses well defined goals to assess our work and learning.*	.35	.13	**-.44**	-.01	-.10	N52	
C19	*tells us that we are expected to do well in maths.*	.06	-.12	**-.43**	-.13	-.01		
C31	*ensures that all students take courses that lead to increased mathematical knowledge.*	.19	.11	**-.39**	.02	-.13	N06	
C21	*motivates us to do our best work.*	.31	.30	**-.37**	.04	.00		
C02	*places a high value on learning maths.*	-.27	.24	**-.36**	.10	-.28		
C39	*skillfully combines their knowledge of adolescents, mathematics and how we learn to help us be successful in maths.*	.29	.18	**-.30**	-.03	-.23		
C09	*integrates the goals of the curriculum and their knowledge of the students in the class.*	.06	.24	**-.24**	.19	-.22	N35	US26
C11	*is not afraid of failure.*	.24	.02	.15	**.54**	-.09		
C30	*provides tasks that challenge us to think..*	-.02	.06	-.16	**.49**	-.14		
C28	*models their own mathematical reasoning in all tasks, actions and discussions.*	-.04	.19	-.17	**.38**	-.24		
C55	*uses assessment results to provide extra help/extension to appropriate students.*	.28	.18	-.20	**.36**	-.09	N17	US11
C15	*consistently makes decisions about their teaching that will further our learning.*	.22	.20	-.16	.29	-.13	N33	US24
C03	*focuses all of the students on their work.*	-.00	.25	-.10	.27	-.24		
C14	*uses examples from other school subjects and the outside world to help us understand new ideas in maths.*	.24	.06	-.07	.25	-.23	N39	US29
C04	*is able to use many different ways to get mathematical ideas across, like words, stories, numbers, diagrams, graphs and symbols.*	-.05	-.05	-.00	.04	**-.75**		
C05	*chooses imaginative examples, problems and situations that motivate us.*	.17	.04	-.05	-.01	**-.68**		
C12	*applies concepts in realistic settings.*	.06	.11	.00	.36	**-.46**	N58	US43
C08	*explores ideas with us even if the answer is not known in advance.*	.03	.10	-.07	.25	**-.44**	N34	US25
C07	*uses examples from a wide range of fields to show how maths is related and useful.*	.35	-.06	-.08	.05	**-.40**		
C46	*seems to modify their plans for the lesson if something interesting comes up.*	.24	-.02	-.21	-.10	**-.39**		
C13	*presents new ideas they have found in journals and at conferences and meetings to help us expand our learning of maths.*	.37	-.09	-.05	.09	**-.38**		
C36	*uses a variety of techniques to maintain control of the students in this class.*	.04	.08	-.14	.15	**-.35**		
C43	*asks questions and uses skilful probing to help classroom discussion and thinking.*	.14	.09	-.27	.14	**-.27**	N05	US02

Item	My mathematics teacher …	Factors					Used in	
		1	2	3	4	5	N	US
	Factor correlations	1	2	3	4	5		
	Factor 1: Varies the lesson	--						
	Factor 2: Manages the learning environment	.36	--					
	Factor 3: Commitment to students and their learning	-.29	-.37	--				
	Factor 4: Reasoning and thinking mathematically	.14	.42	-.22	--			
	Factor 5: Relates mathematics to the real world	-.53	-.39	.37	-.22	--		

N and US. Items prefixed with US were in the final instrument (SEAT-M) used in the USA. Prefix N indicates that they were further trialled in the November instrument, but did not make the SEAT-M instrument.

Explains 50.2% of variance

Cronbach's alpha = .973

Factor 1: Varies the lesson
Factor 2: Manages the learning environment
Factor 3: Commitment to students and their learning
Factor 4: Reasoning and thinking mathematically
Factor 5: Relates lesson to real world

Table 17 Classical Test Theory and Item Response Theory Item Statistics for 67 Form C Items

Scale	Item #	Item Mean	Item-total correlation r_{pbs}	a (SE)	b_1 (SE)	b_2 (SE)	b_3 (SE)	b_4 (SE)	b_5 (SE)
N41	C01	3.90 (0.10)	.685	1.95 (0.37)	-2.19 (0.43)	-1.11 (0.23)	-0.12 (0.14)	0.64 (0.14)	1.86 (0.32)
	C02	4.39 (0.09)	.472	1.15 (0.28)	-3.96 (1.11)	-2.56 (0.66)	-0.94 (0.31)	0.03 (0.20)	1.95 (0.46)
	C03	4.06 (0.11)	.569	1.54 (0.30)	-1.92 (0.45)	-1.26 (0.27)	-0.52 (0.21)	0.51 (0.17)	1.74 (0.34)
	C04	3.99 (0.12)	.505	1.23 (0.27)	-2.17 (0.50)	-1.11 (0.30)	-0.32 (0.21)	0.58 (0.19)	1.67 (0.36)
	C05	3.67 (0.12)	.683	2.28 (0.36)	-1.49 (0.27)	-0.45 (0.15)	0.08 (0.12)	0.76 (0.13)	1.53 (0.24)
	C06	2.51 (0.11)	.481	1.33 (0.31)	-0.54 (0.24)	0.39 (0.18)	1.27 (0.28)	2.01 (0.44)	3.31 (0.80)
	C07	3.55 (0.11)	.632	1.66 (0.31)	-1.62 (0.33)	-0.56 (0.18)	0.21 (0.14)	0.82 (0.17)	2.19 (0.42)
US25	C08	3.99 (0.10)	.657	1.78 (0.23)	-2.15 (0.48)	-1.17 (0.25)	-0.23 (0.16)	0.63 (0.16)	1.79 (0.35)
US26	C09	3.95 (0.11)	.579	1.97 (0.34)	-1.95 (0.44)	-0.98 (0.24)	-0.22 (0.14)	0.74 (0.15)	1.54 (0.25)
	C10	2.80 (0.12)	.448	1.45 (0.31)	-0.73 (0.24)	0.10 (0.17)	0.78 (0.20)	1.60 (0.33)	2.97 (0.62)
	C11	4.20 (0.12)	.429	1.03 (0.26)	-3.19 (0.91)	-1.69 (0.50)	-0.54 (0.28)	0.33 (0.23)	1.56 (0.42)
US43	C12	4.01 (0.10)	.655	1.91 (0.34)	-2.20 (0.48)	-1.24 (0.24)	-0.36 (0.15)	0.60 (0.14)	1.71 (0.28)
	C13	3.12 (0.12)	.543	1.49 (0.32)	-1.21 (0.32)	-0.15 (0.18)	0.58 (0.18)	1.32 (0.27)	2.23 (0.43)
US29	C14	3.24 (0.11)	.553	1.44 (0.24)	-1.71 (0.33)	-0.25 (0.19)	0.52 (0.18)	1.29 (0.26)	2.61 (0.51)
US24	C15	3.66 (0.10)	.664	2.09 (0.33)	-1.92 (0.33)	-0.82 (0.18)	0.13 (0.12)	0.90 (0.15)	2.17 (0.38)
	C16	4.44 (0.10)	.609	1.99 (0.32)	-2.08 (0.42)	-1.70 (0.25)	-0.79 (0.17)	0.20 (0.13)	1.16 (0.19)
	C17	4.52 (0.10)	.486	1.39 (0.28)	-3.91 (1.17)	-1.75 (0.40)	-0.88 (0.23)	0.03 (0.17)	1.10 (0.24)
	C18	4.67 (0.10)	.431	1.10 (0.55)	-4.05 (1.34)	-2.54 (0.70)	-1.07 (0.34)	-0.21 (0.23)	0.93 (0.29)
	C19	4.08 (0.12)	.215	0.42 (0.33)	-6.58 (3.58)	-3.71 (2.15)	-1.47 (1.03)	0.72 (0.62)	3.61 (1.89)
	C20	2.14 (0.10)	.500	1.22 (0.26)	-0.34 (0.22)	1.06 (0.24)	1.85 (0.37)	2.86 (0.62)	4.63 (1.39)
	C21	4.10 (0.11)	.701	2.29 (0.38)	-2.59 (0.40)	-0.87 (0.18)	-0.23 (0.14)	0.43 (0.12)	1.23 (0.18)
	C22	4.42 (0.10)	.635	1.97 (0.29)	-3.05 (0.80)	-1.38 (0.27)	-0.58 (0.16)	0.26 (0.13)	1.03 (0.18)
N37	C23	3.35 (0.11)	.655	2.12 (0.36)	-1.35 (0.26)	-0.44 (0.15)	0.28 (0.13)	1.23 (0.18)	2.36 (0.39)
	C24	2.78 (0.12)	.606	1.62 (0.35)	-0.73 (0.21)	0.27 (0.15)	0.89 (0.18)	1.43 (0.26)	2.23 (0.45)
	C25	2.82 (0.11)	.545	1.71 (0.29)	-0.76 (0.20)	0.08 (0.15)	0.85 (0.17)	1.49 (0.25)	2.79 (0.60)
	C26	4.03 (0.13)	.529	1.20 (0.27)	-2.05 (0.56)	-1.14 (0.35)	-0.35 (0.23)	0.36 (0.20)	1.50 (0.36)
N56	C27	3.94 (0.11)	.801	2.89 (0.41)	-1.78 (0.36)	-0.70 (0.13)	-0.07 (0.11)	0.49 (0.10)	1.37 (0.16)
	C28	3.97 (0.10)	.585	1.64 (0.32)	-3.07 (0.70)	-1.30 (0.27)	-0.15 (0.16)	0.64 (0.17)	1.79 (0.34)
US33	C29	4.32 (0.10)	.692	1.81 (0.34)	-2.64 (0.60)	-1.64 (0.39)	-0.61 (0.18)	0.32 (0.14)	1.27 (0.23)
	C30	4.77 (0.09)	.534	1.15 (0.27)	-4.01 (1.20)	-2.41 (0.68)	-1.64 (0.47)	-0.58 (0.24)	1.20 (0.33)
N06	C31	3.47 (0.11)	.653	1.88 (0.27)	-1.49 (0.28)	-0.59 (0.17)	0.20 (0.14)	0.98 (0.17)	1.99 (0.33)
US34	C32	4.00 (0.10)	.604	1.74 (0.29)	-2.46 (0.50)	-1.19 (0.22)	-0.15 (0.15)	0.57 (0.15)	1.74 (0.28)
	C33	4.22 (0.11)	.549	1.53 (0.27)	-2.34 (0.49)	-1.26 (0.24)	-0.43 (0.19)	0.32 (0.17)	1.42 (0.27)
	C34	3.88 (0.11)	.742	2.49 (0.39)	-1.62 (0.24)	-0.70 (0.15)	-0.16 (0.12)	0.64 (0.12)	1.60 (0.23)
	C35	4.69 (0.10)	.576	1.28 (0.29)	-3.59 (0.95)	-2.41 (0.56)	-1.30 (0.34)	-0.24 (0.20)	1.04 (0.26)

Scale	Item #	Item Mean	Item *r*	*a* (SE)	b_1 (SE)	b_2 (SE)	b_3 (SE)	b_4 (SE)	b_5 (SE)
	C36	3.73 (0.11)	.583	1.47 (0.26)	-1.95 (0.40)	-0.73 (0.21)	-0.06 (0.16)	0.77 (0.20)	2.21 (0.40)
	C37	4.17 (0.10)	.592	1.70 (0.34)	-2.74 (0.57)	-2.74 (0.57)	-0.44 (0.17)	0.48 (0.14)	1.37 (0.25)
	C38	3.36 (0.11)	.601	1.73 (0.32)	-1.50 (0.38)	-0.47 (0.18)	0.18 (0.14)	1.24 (0.21)	2.62 (0.57)
	C39	3.51 (0.11)	.689	2.33 (0.35)	-1.49 (0.23)	-0.48 (0.15)	0.13 (0.12)	0.94 (0.14)	1.91 (0.29)
US37	C40	3.39 (0.11)	.623	1.77 (0.34)	-1.65 (0.41)	-0.41 (0.17)	0.32 (0.14)	1.16 (0.20)	2.18 (0.43)
	C41	3.34 (0.10)	.634	1.84 (0.23)	-1.90 (0.41)	-0.54 (0.19)	0.50 (0.15)	1.38 (0.24)	2.05 (0.38)
N52	C42	3.67 (0.10)	.668	2.52 (0.38)	-1.47 (0.23)	-0.76 (0.17)	0.08 (0.10)	0.80 (0.13)	1.81 (0.28)
US02	C43	4.04 (0.11)	.640	1.71 (0.28)	-2.33 (0.49)	-1.07 (0.24)	-0.22 (0.15)	0.56 (0.16)	1.52 (0.26)
	C44	4.55 (0.11)	.174	0.49 (0.24)	-7.90 (3.74)	-4.56 (2.20)	-2.55 (1.29)	-0.42 (1.00)	2.08 (1.05)
	C45	2.94 (0.13)	.471	1.37 (0.32)	-0.76 (0.34)	-0.03 (0.18)	0.70 (0.20)	1.51 (0.32)	2.47 (0.58)
	C46	3.50 (0.12)	.523	1.56 (0.31)	-1.43 (0.32)	-0.50 (0.21)	0.18 (0.15)	0.81 (0.19)	2.13 (0.42)
	C47	4.96 (0.10)	.252	0.55 (0.24)	-7.76 (3.36)	-4.50 (1.95)	-3.06 (1.40)	-1.74 (0.88)	0.63 (0.48)
	C48	4.02 (0.13)	.625	1.88 (0.25)	-1.58 (0.31)	-0.83 (0.19)	-0.28 (0.15)	0.35 (0.14)	1.24 (0.21)
	C49	4.30 (0.11)	.631	1.97 (0.35)	-2.02 (0.36)	-1.31 (0.30)	-0.35 (0.15)	0.10 (0.13)	1.12 (0.20)
	C50	3.68 (0.11)	.765	2.70 (0.39)	-1.27 (0.22)	-0.70 (0.15)	0.10 (0.10)	0.68 (0.12)	1.55 (0.20)
	C51	3.52 (0.12)	.670	1.87 (0.29)	-1.12 (0.23)	-0.48 (0.17)	0.02 (0.15)	0.95 (0.18)	1.94 (0.33)
	C52	3.01 (0.11)	.656	2.06 (0.37)	-0.91 (0.20)	-0.06 (0.13)	0.53 (0.12)	1.39 (0.21)	2.16 (0.39)
US23	C53	4.12 (0.11)	.702	2.23 (0.32)	-1.80 (0.33)	-1.13 (0.23)	-0.31 (0.14)	0.40 (0.12)	1.41 (0.21)
	C54	3.87 (0.13)	.424	0.88 (0.21)	-2.84 (1.00)	-1.35 (0.63)	-0.23 (0.29)	0.71 (0.31)	2.05 (0.63)
US11	C55	3.82 (0.12)	.729	2.35 (0.38)	-1.29 (0.24)	-0.64 (0.15)	0.02 (0.12)	0.61 (0.12)	1.37 (0.19)
	C56	3.64 (0.12)	.684	1.90 (0.27)	-1.61 (0.33)	-0.59 (0.15)	0.09 (0.15)	0.72 (0.16)	1.71 (0.26)
	C57	2.92 (0.12)	.539	1.47 (0.30)	-1.03 (0.30)	0.04 (0.17)	0.86 (0.20)	1.70 (0.28)	2.23 (0.40)
	C58	4.65 (0.11)	.516	1.32 (0.28)	-2.82 (0.69)	-2.38 (0.48)	-1.11 (0.32)	0.00 (0.18)	0.77 (0.24)
N03	C59	3.93 (0.10)	.564	1.38 (0.27)	-2.65 (0.63)	-1.45 (0.31)	-0.22 (0.19)	0.77 (0.21)	2.03 (0.41)
US08	C60	3.84 (0.12)	.665	2.25 (0.36)	-1.55 (0.29)	-0.61 (0.16)	-0.05 (0.12)	0.43 (0.13)	1.37 (0.19)
US45	C61	3.55 (0.12)	.610	1.74 (0.32)	-1.41 (0.29)	-0.45 (0.17)	0.16 (0.15)	0.87 (0.18)	1.98 (0.34)
US44	C62	3.68 (0.11)	.615	1.96 (0.37)	-1.84 (0.35)	-0.67 (0.19)	0.03 (0.13)	0.86 (0.17)	1.82 (0.32)
US28	C63	3.48 (0.11)	.743	2.62 (0.46)	-1.34 (0.20)	-0.49 (0.13)	0.28 (0.11)	0.94 (0.13)	1.69 (0.24)
	C64	4.47 (0.11)	.685	2.28 (0.43)	-1.95 (0.36)	-1.22 (0.20)	-0.54 (0.15)	0.05 (0.12)	0.87 (0.14)
	C65	4.33 (0.11)	.635	1.87 (0.48)	-2.03 (0.45)	-1.29 (0.28)	-0.58 (0.16)	0.18 (0.13)	1.24 (0.20)
	C66	4.76 (0.10)	.496	1.17 (0.32)	-3.26 (0.78)	-2.32 (0.53)	-1.42 (0.35)	-0.35 (0.21)	0.87 (0.27)
	C67	3.91 (0.12)	.565	1.52 (0.29)	-1.98 (0.36)	-1.00 (0.24)	-0.19 (0.17)	0.54 (0.17)	1.66 (0.31)
	C68	3.83 (0.14)	.663						

Note: SE is in brackets

In this instance, five items with relatively weak factor loadings were selected. These five items were C62, C09, C15, C14 and C43. Factor Four had only seven items in the original analysis, and C14 and C15 were retained primarily because of their good IRT characteristics and the need to maintain an appropriate mapping of the items to the Standards.

Table 18 Factor loadings and IRT parameters for 21 items selected from Form C

Item	Factor loadings					IRT parameters					
	1	2	3	4	5	*a*	b_1	b_2	b_3	b_4	b_5
C23	**.53**					2.12	-1.35	-.44	.28	1.23	2.36
C27	**.46**					2.89	-1.78	-.70	-.07	.49	1.37
C61	**.39**	.33				1.74	-1.41	-.45	.16	.87	1.98
C63	**.37**		.36			2.62	-1.34	-.49	.28	.94	1.69
C62	**.28**					1.96	-1.84	.67	.03	.86	1.82
C53		**.51**				2.23	-1.80	-1.13	-.31	.40	1.41
C60		**.47**				2.25	-1.55	-.61	-.05	.43	1.37
C29		**.44**				1.81	-2.64	-1.64	-.61	.32	1.27
C01		**.42**				1.95	-2.19	-1.11	-.12	.64	1.86
C59		**.35**				1.38	-2.65	-1.45	-.22	.77	2.03
C32			**.82**			1.74	-2.46	-1.19	-.15	.57	1.74
C40			**.45**			1.77	-1.65	-.41	.32	1.16	2.18
C42	.35		**.44**			2.52	-1.47	-.76	.08	.80	1.81
C31			**.39**			1.88	-1.49	-.59	.20	.98	1.99
C09			**.24**			1.97	-1.95	-.98	-.22	.74	1.54
C55				**.36**		2.35	-1.29	-.64	.02	.61	1.37
C15				**.29**		2.09	-1.92	-.82	.13	.90	2.17
C14				**.25**		1.44	-1.71	-.25	.52	1.29	2.61
C12				.36	**.46**	1.91	-2.20	-1.24	-.36	.60	1.71
C08					**.44**	1.78	-2.15	-1.17	-.23	.63	1.79
C43					**.27**	1.71	-2.33	-1.07	-.22	.56	1.52

On each of Forms A, B and C the Standard Error (SE) for the b_1 parameter is usually much larger than the SE of the other b_i parameters. The first response category was not used as frequently as the other response categories, and this results in a negative skew in the item responses. For example, in Form A, there were a total of 692 (1.0% of all responses) responses in Category 1 (the lowest rating category for the teacher), 1330 (7.7%) responses in Category 2, 1816 (14.8%) in Category 3, 1958 (21.8%) in Category 4, 1703 (19.0%) in Category 5, and 1379 (15.4%) in Category 6, the highest response category. This tendency to rate teachers towards the more favourable end of the scale can be seen in each of the Forms, with forty Form A items, forty-six Form B items and fifty-five Form C items with a negative skew statistic. It has been noted by several authors (Centra, 1973b; Hativa, 1996; Holmes, 1996; K. D. Peterson et al., 2000) that student ratings tend to use the higher end of the scale, with lowest category used relatively infrequently. Therefore, the relatively high values of the SE for Category 1 were not unexpected.

Questionnaire assembly for the November questionnaire

From the three Forms, a total of 65 items were selected for inclusion in the November questionnaire (Appendix Four). They consisted of 24 items from Form A, 20 items from Form B, and 21 items from Form C. The final global rating item (*My mathematics teacher compared with all other maths teachers I have had, is the best*) was added to these to give a total of 66 items for the second trial in November. The items from each form were randomly assigned to the November questionnaire.

Prior to the assembly of the November questionnaire, several of the items were amended to improve their wording. Four of these were the items that had been written to cover the mathematical content described in Standard III of the Standards. These were A15 (geometry), A45 (calculus), A35 (statistics), and A63 (algebra). The technology item (A12) was amended to show two examples (calculators and computers) to ensure that students understood that this item was not concerned with the curriculum subject, Technology. The other amendments were designed to make the language more accessible to high schools students, or to clarify the item and remove unnecessary detail. The wording used in the original form and the adapted wording for the November Form are shown in Table 19.

Table 19 Wording amendments made to items selected for Form November

Original wording		Amended wording	
A12	regards technology as an essential tool for teaching maths.	N07	regards technology (e.g., calculators and computers) as an essential tool for teaching maths.
A15	shows us how we can use geometry to solve problems in the real world.	N12	makes geometry interesting for me.
A35	shows us how we can use statistics to solve problems in the real world.	N51	makes statistics interesting for me.
A45	shows us how we can use calculus to solve problems in the real world.	N19	makes calculus interesting for me.
A57	conveys to the class the idea that maths relates to the real world.	N20	helps the class to understand that maths relates to the real world.
A63	shows us how we can use algebra to represent patterns and solve problems in the real world.	N65	makes algebra interesting for me.
B27	empowers students to think through and solve problems, either by themselves or together as a group.	N25	challenges students to think through and solve problems, either by themselves or together as a group.
B55	keeps my family informed about my progress in maths.	N64	keeps my family informed on a regular basis about my progress in maths.
B58	involves families, administrators and teachers in the school and community to help and support students to learn and continue in maths.	N29	involves our families and other teachers in the school to help and support us to learn and continue in maths.

	Original wording		Amended wording
C14	uses examples from other school subjects and the outside world to help us understand new ideas in maths.	N39	uses examples that help us to understand and learn new ideas.
C43	asks questions and uses skilful probing to help classroom discussion and thinking.	N05	skilfully asks questions to help classroom discussion and thinking.

Trial Two

Form November

The purpose of this second round was to re-trial the selected items with a view to reducing the number of items included in SEAT-M.

This trial took place in November (near the end of the NZ school year). Two of the three schools were single-sex, one for boys and the other for girls. Table 20 shows the roll, decile rating, ethnic makeup and gender balance for each of the schools. These schools were mid- to high decile schools. As in Trial One, the names of schools have been changed because the schools had agreed to participate anonymously. The students and teachers gave informed consent as required by the Human Participants Ethics Committee (Ref: 2000/090).

Table 20 School Descriptives for Trial Two

School	Roll N	Decile (SES)	Ethnicity (%)					Gender (%)	
			NZE	M	PI	A	O	F	M
Ngatai	750	6	56	4	10	22	8	50	50
Ross Boys'	1600	10	59	3	1	22	15	0	100
Teal Girls'	1700	10	78	4	0	18	0	100	0

All of the students participating in this trial were in Year 12, as Year 11 and 13 students were on leave for the annual School Certificate and University Bursary examinations.

Year 12 classes had completed all of their course requirements for Sixth Form Certificate and the teachers were prepared to make these classes available for data gathering. At this stage of the year, the students had experienced a complete academic year with their mathematics teacher, and were ideally placed to record their ratings about that teacher's performance. Table 21 shows the number of students, their ethnic affiliation, and gender.

Table 21 Participant Descriptives for Trial Two

School	N	Ethnicity (N)					Gender (N)	
		NZE	*M*	*PI*	*A*	*O*	*F*	*M*
Ngatai	53	28	1	3	14	6	21	32
Ross Boys'	151	89	6	10	30	14	0	151
Teal Girls'	125	74	1	3	38	8	125	0
Total	329	191	8	16	82	28	146	183
Percent		58	2	5	25	9	44	56

Four of the students (one at Ngatai, two at Ross Boys', and one at Teal Girls') did not provide information about their ethnicity. As with Trial One, New Zealand European and Asian students are over-represented, and Maori and Pacific Island students are under-represented.

Descriptive statistics

The number of responses, mean, standard deviation, skew and kurtosis statistics for each of the 66 items in November Form is shown in Table 22.

The mean rating on individual items ranged from a maximum of 4.36 (Item N07) to a minimum of 2.05 (Item N55), with a mean rating for all items on November Form of 3.58. Cronbach's alpha reliability for the whole test was $\alpha = .98$, indicating high internal consistency among the items on the Form. It is noted that none of the items had a mean rating in excess of 5, and that only six items had a mean rating greater than 4,

providing further evidence in support of the judicious way in which students assign their ratings to teachers.

Factor analysis

The 66 items of November Form were subjected to maximum likelihood factor analysis, with oblimin rotation. The Kaiser-Meyer-Olkin measure of sampling adequacy was a "marvelous" .97 (Kaiser, 1974, p. 35). Five interpretable factors were extracted, explaining 60.8% of the total variance. The goodness of fit statistic $\chi^2(1825) = 2734.91$, $p < .01$ indicates very good specification of the five factor model. The pattern matrix for these five factors is shown in Table 23.

Factor One deals with the teachers' Commitment to Students and Their Learning; Factor Two describes the teachers' Mathematical Pedagogy; Factor Three concerns Student Engagement with the Curriculum; Factor Four describes the links that the teacher makes with Family and Community; and, Factor Five describes how the teacher Relates Mathematics to the Real World.

Factor One, Commitment to Students and Their Learning, (eigenvalue = 33.89) accounts for 51.3% of the common variance; Factor Two, Mathematical Pedagogy, (eigenvalue = 1.90) accounts for 2.9% of the variance; Factor Three, Student Engagement with the Curriculum (eigenvalue = 1.56) accounts for 2.4% of the common shared variance; Factor Four, Family and Community (eigenvalue = 1.50) accounts for 2.3% of the common shared variance; and, Factor Five, Relates Mathematics to the Real World (eigenvalue = 1.27) accounts for 1.9% of the shared variance.

The absolute value of the correlations between factors range from .48 to .65. While the use of an oblique rotation allows for the factors to be correlated, these values are all quite high. These factors do not describe unique dimensions of exemplary mathematics teaching, but as has been noted above, given that the items were all derived from statements from the same or overlapping sections of the Standards as was clearly noted by the participants in the Focus Groups in Study 1, this may have been expected.

Table 22 Descriptive Statistics for Form November

	My mathematics teacher ...	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
N01	*makes maths come alive in the classroom.*	329	3.26	.07	1.34	.15	.13	-.85	.27
N02	*help us to communicate better in maths.*	328	3.63	.07	1.30	-.11	.14	-.73	.27
N03	*uses an appropriate range of formal and informal assessments to monitor individual and class progress.*	320	3.91	.07	1.29	-.29	.14	-.67	.27
N04	*provides time to develop problem solving skills that we can use both in maths and outside the classroom.*	327	3.78	.07	1.33	-.22	.14	-.67	.27
N05	*skilfully asks questions to help classroom discussion and thinking.*	327	3.74	.08	1.45	-.20	.14	-.96	.27
N06	*ensures that all students take courses that lead to increased mathematical knowledge.*	328	3.54	.08	1.45	.02	.14	-.90	.27
N07	*regards technology (e.g., calculators and computers) as an essential tool for teaching maths.*	328	4.36	.07	1.32	-.44	.14	-.68	.27
N08	*teaches us the fundamental processes of mathematical thinking – for example: exploration, interpretation, representation, modelling, and analysis.*	327	3.86	.07	1.31	-.15	.14	-.62	.27
N09	*shows us interesting and useful ways of solving problems.*	328	3.80	.08	1.48	-.23	.14	-.94	.27
N10	*understands the impact that home life, cultural background, community expectations and student attitudes can have on our learning.*	323	3.46	.09	1.56	.06	.14	-1.09	.27
N11	*enables us to develop confidence and self esteem in maths.*	328	3.51	.08	1.51	-.08	.14	-.98	.27
N12	*makes geometry interesting for me.*	327	3.14	.08	1.42	.14	.14	-.81	.27
N13	*creates a positive atmosphere in class where we feel part of a team of learners.*	327	3.68	.09	1.60	-.25	.14	-1.04	.27
N14	*adjusts the lesson if we experience difficulties in learning.*	328	3.88	.09	1.64	-.32	.14	-1.14	.27

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
N15	*helps us make the links between the different strands of maths and other aspects of our lives.*	324	3.63	.08	1.36	-.13	.14	-.62	.27
N16	*helps us construct an understanding of the language and processes of maths.*	326	3.82	.07	1.30	-.27	.14	-.64	.27
N17	*uses assessment results to provide extra help/extension to appropriate students.*	328	3.36	.09	1.57	.11	.14	-1.11	.27
N18	*teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.*	324	3.16	.07	1.30	.18	.14	-.66	.27
N19	*makes calculus interesting for me.*	308	3.08	.10	1.67	.20	.14	-1.20	.28
N20	*helps the class to understand that maths relates to the real world.*	327	3.82	.08	1.36	-.24	.14	-.72	.27
N21	*encourages us to seek more than one solution to problems.*	325	3.72	.08	1.41	-.12	.14	-.89	.27
N22	*holds my interest in class.*	326	3.29	.09	1.62	.03	.14	-1.22	.27
N23	*makes learning maths satisfying and stimulating.*	324	3.18	.08	1.50	.04	.14	-1.04	.27
N24	*provides time for us to reflect and talk about the maths we are learning.*	327	3.35	.09	1.57	.09	.14	-1.11	.27
N25	*challenges students to think through and solve problems, either by themselves or together as a group.*	326	4.07	.08	1.38	-.40	.14	-.58	.27
N26	*encourages us to try different techniques to solve problems.*	329	3.72	.08	1.38	-.10	.13	-.88	.27
N27	*stimulates our learning by varying the way we are taught to allow for the strengths and weaknesses of the people in the class.*	328	3.25	.08	1.44	.20	.14	-.95	.27
N28	*is committed to the learning of all the students in the class.*	326	4.05	.09	1.66	-.43	.14	-1.00	.27
N29	*involves our families and other teachers in the school to help and support us to learn and continue in maths.*	328	2.61	.08	1.37	.58	.14	-.50	.27
N30	*is able to explain something in different ways to help us understand.*	324	4.08	.09	1.54	-.41	.14	-.90	.27

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
N31	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.*	328	3.85	.08	1.48	-.33	.14	-.85	.27
N32	*encourages us to question and discuss the mathematical ideas and concepts we are taught.*	328	3.73	.09	1.55	-.18	.14	-.99	.27
N33	*consistently makes decisions about their teaching that will further our learning.*	322	3.44	.07	1.29	.11	.14	-.63	.27
N34	*explores ideas with us even if the answer is not known in advance.*	325	3.95	.07	1.32	-.23	.14	-.82	.27
N35	*integrates the goals of the curriculum and their knowledge of the students in the class.*	321	3.73	.07	1.26	-.08	.14	-.55	.27
N36	*sometimes involves us and our family in exploring career opportunities.*	324	2.46	.07	1.32	.65	.14	-.40	.27
N37	*provides the inspiration for student investigations.*	325	3.10	.08	1.36	.16	.14	-.72	.27
N38	*teaches us how to evaluate progress towards our goals.*	322	3.15	.08	1.40	.16	.14	-.80	.27
N39	*uses examples that help us to understand and learn new ideas.*	324	4.19	.08	1.41	-.47	.14	-.63	.27
N40	*uses a variety of methods to collect, organise, represent and summarise collections of data.*	325	3.70	.07	1.32	-.22	.14	-.53	.27
N41	*makes maths meaningful for me.*	325	3.40	.09	1.63	-.07	.14	-1.21	.27
N42	*uses interesting materials and resources that appeal to different people in the class.*	322	3.19	.08	1.45	.17	.14	-.89	.27
N43	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	322	3.60	.08	1.43	-.06	.14	-.82	.27
N44	*knows and caters for the problems we commonly encounter in learning new topics.*	325	3.97	.08	1.41	-.39	.14	-.65	.27
N45	*allows us to learn maths in different ways.*	327	3.61	.08	1.40	-.01	.14	-.85	.27
N46	*encourages us to place a high value on maths.*	324	4.02	.08	1.43	-.46	.14	-.58	.27

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
N47	*creates a welcoming environment in the classroom for family members and members of the community.*	319	3.15	.09	1.60	.26	.14	-.99	.27
N48	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.*	326	3.75	.09	1.68	-.23	.14	-1.16	.27
N49	*prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.*	325	3.67	.08	1.51	-.17	.14	-1.00	.27
N50	*helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	323	3.48	.08	1.42	-.00	.14	-.94	.27
N51	*makes statistics interesting for me.*	324	3.27	.08	1.48	-.04	.14	-1.00	.27
N52	*uses well defined goals to assess our work and learning.*	321	3.42	.08	1.34	-.09	.14	-.69	.27
N53	*works with other subject teachers to provide for students in the class.*	321	2.98	.08	1.39	.41	.14	-.57	.27
N54	*uses their knowledge about each of us to create problems that are interesting and worth solving.*	324	3.29	.08	1.47	.09	.14	-.97	.27
N55	*seeks information from my family about my strengths, interests, habits and home life.*	326	2.05	.07	1.27	1.33	.14	1.26	.27
N56	*helps us experience success in doing worthwhile maths.*	324	3.46	.08	1.41	.08	.14	-.75	.27
N57	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	323	3.34	.08	1.39	.06	.14	-.91	.27
N58	*applies concepts in realistic settings.*	322	3.91	.07	1.31	-.24	.14	-.66	.27
N59	*gets us to think about the nature and quality of our work.*	323	3.63	.07	1.29	-.01	.14	-.74	.27
N60	*tells us what the purpose of each lesson is.*	324	3.54	.09	1.58	-.07	.14	-1.14	.27
N61	*encourages us to test mathematical ideas and discover mathematical principles.*	321	3.58	.08	1.35	-.07	.14	-.78	.27

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
N62	*helps us apply our growing knowledge in both pure and applied settings.*	323	3.49	.07	1.31	.01	.14	-.61	.27
N63	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	325	3.67	.07	1.33	-.19	.14	-.55	.27
N64	*keeps my family informed on a regular basis about my progress in maths.*	325	2.20	.07	1.26	1.02	.14	.50	.27
N65	*makes algebra interesting for me.*	323	3.50	.09	1.67	-.07	.14	-1.17	.27
N66	*compared with all other maths teachers I have had, is the best.*	324	3.78	.10	1.87	-.20	.14	-1.44	.27
	Valid N (listwise)	223							

Table 23 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form November

		Factors					
Item	My mathematics teacher …	1	2	3	4	5	US
N14	*adjusts the lesson if we experience difficulties in learning.*	**.61**	.06	-.13	-.01	-.15	US08
N17	*uses assessment results to provide extra help/extension to appropriate students.*	**.58**	-.08	-.10	.27	.08	US11
N48	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.*	**.53**	.10	-.13	.10	-.11	US36
N11	*enables us to develop confidence and self esteem in maths.*	**.51**	.08	-.21	.17	-.02	US05
N28	*is committed to the learning of all the students in the class.*	**.51**	.18	-.07	.09	-.11	US20
N30	*is able to explain something in different ways to help us understand.*	**.45**	.06	-.12	.05	-.32	US22
N16	*helps us construct an understanding of the language and processes of maths.*	**.45**	.24	-.15	-.03	-.13	US10
N25	*challenges students to think through and solve problems, either by themselves or together as a group.*	**.43**	.12	-.05	.02	-.22	US18
N44	*knows and caters for the problems we commonly encounter in learning new topics.*	**.42**	.40	-.04	.01	-.08	US33
N10	*understands the impact that home life, cultural background, community expectations and student attitudes can have on our learning.*	**.40**	.17	.02	.28	.02	
N05	*skilfully asks questions to help classroom discussion and thinking.*	**.39**	.14	-.13	.08	-.18	US02
N26	*encourages us to try different techniques to solve problems.*	**.39**	.17	-.09	.04	-.25	US19
N21	*encourages us to seek more than one solution to problems.*	**.38**	.09	-.04	.04	-.37	US15
N24	*provides time for us to reflect and talk about the maths we are learning.*	**.37**	.10	-.19	.13	-.08	US17
N66	*compared with all other maths teachers I have had, is the best.*	**.37**	.08	-.37	-.13	-.19	US51
N31	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.*	**.36**	.24	-.31	-.05	-.08	US23
N03	*uses an appropriate range of formal and informal assessments to monitor individual and class progress.*	**.34**	.21	-.05	.15	.02	
N02	*help us to communicate better in maths.*	**.31**	.16	-.20	.09	-.17	
N27	*stimulates our learning by varying the way we are taught to allow for the strengths and weaknesses of the people in the class.*	**.29**	.11	-.19	.18	-.24	

Item	My mathematics teacher …	Factors					US
		1	2	3	4	5	
N32	*encourages us to question and discuss the mathematical ideas and concepts we are taught.*	**.28**	.20	-.16	.19	-.14	
N06	*ensures that all students take courses that lead to increased mathematical knowledge.*	**.25**	.18	-.20	.20	.03	
N54	*uses their knowledge about each of us to create problems that are interesting and worth solving.*	**.22**	.19	-.16	.21	-.11	
N62	*helps us apply our growing knowledge in both pure and applied settings.*	-.11	**.73**	-.06	.16	-.03	US47
N61	*encourages us to test mathematical ideas and discover mathematical principles.*	.03	**.62**	-.18	.05	.02	US46
N58	*applies concepts in realistic settings.*	.04	**.54**	-.02	-.09	-.24	US43
N63	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	.07	**.53**	-.20	.08	-.04	US48
N34	*explores ideas with us even if the answer is not known in advance.*	.30	**.47**	.01	.00	-.06	US25
N43	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	.09	**.43**	-.21	.02	-.13	US32
N60	*tells us what the purpose of each lesson is.*	.33	**.43**	-.02	.16	.05	US45
N40	*uses a variety of methods to collect, organise, represent and summarise collections of data.*	.04	**.42**	-.11	.12	-.19	US30
N39	*uses examples that help us to understand and learn new ideas.*	.21	**.42**	-.19	-.04	-.13	US29
N59	*gets us to think about the nature and quality of our work.*	.09	**.42**	.02	.11	-.13	US44
N46	*encourages us to place a high value on maths.*	.24	**.37**	.17	.24	.01	US34
N33	*consistently makes decisions about their teaching that will further our learning.*	.05	**.37**	.21	.05	-.07	US24
N35	*integrates the goals of the curriculum and their knowledge of the students in the class.*	.20	**.36**	.07	.22	-.08	US26
N08	*teaches us the fundamental processes of mathematical thinking – for example: exploration, interpretation, representation, modelling, and analysis.*	.16	**.33**	.18	.02	-.13	US03
N42	*uses interesting materials and resources that appeal to different people in the class.*	-.05	**.32**	.20	.26	-.15	US31
N04	*provides time to develop problem solving skills that we can use both in maths and outside the classroom.*	.18	.23	.10	.12	-.23	
N19	*makes calculus interesting for me.*	.02	-.01	**-.83**	.02	.04	US13
N65	*makes algebra interesting for me.*	-.09	.15	**-.82**	.02	.13	US50

Item	My mathematics teacher …	Factors					US
		1	2	3	4	5	
N41	*makes maths meaningful for me.*	-.06	.09	**-.74**	.01	-.21	
N23	*makes learning maths satisfying and stimulating.*	.17	-.05	**-.69**	.03	-.14	US16
N12	*makes geometry interesting for me.*	.12	-.05	**-.62**	.07	-.06	US06
N01	*makes maths come alive in the classroom.*	.19	-.05	**-.54**	.08	-.20	US01
N22	*holds my interest in class.*	.18	.00	**-.51**	.05	-.20	
N51	*makes statistics interesting for me.*	-.09	.15	**-.47**	.14	-.07	US39
N13	*creates a positive atmosphere in class where we feel part of a team of learners.*	.39	.05	**-.40**	.08	-.04	US07
N09	*shows us interesting and useful ways of solving problems.*	.26	.17	**-.34**	-.00	-.20	US04
N45	*allows us to learn maths in different ways.*	.15	.26	-.26	.10	-.15	
N56	*helps us experience success in doing worthwhile maths.*	.18	.17	-.26	.25	-.16	
N55	*seeks information from my family about my strengths, interests, habits and home life.*	-.06	-.06	-.01	**.76**	-.07	US41
N36	*sometimes involves us and our family in exploring career opportunities.*	.02	.03	-.05	**.62**	-.08	US27
N64	*keeps my family informed on a regular basis about my progress in maths.*	.02	-.04	.10	**.62**	-.04	US49
N29	*involves our families and other teachers in the school to help and support us to learn and continue in maths.*	.29	-.01	-.08	**.52**	.03	US21
N53	*works with other subject teachers to provide for students in the class.*	.03	.18	.04	**.52**	-.02	US40
N47	*creates a welcoming environment in the classroom for family members and members of the community.*	.04	.07	-.32	**.38**	.03	US35
N38	*teaches us how to evaluate progress towards our goals.*	.10	.12	-.11	**.34**	-.27	US28
N37	*provides the inspiration for student investigations.*	.05	.14	-.27	.28	-.23	
N52	*uses well defined goals to assess our work and learning.*	.14	.21	-.21	.24	-.12	
N07	*regards technology (e.g., calculators and computers) as an essential tool for teaching maths.*	.04	.13	.04	.17	-.06	
N20	*helps the class to understand that maths relates to the real world.*	.11	.00	-.02	-.05	**-.73**	US14
N49	*prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.*	-.03	.06	-.08	.12	**-.60**	US37
N50	*helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	-.10	.05	-.16	.15	**-.60**	US38

Item	My mathematics teacher …	Factors 1	2	3	4	5	US
N57	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	.23	.31	.00	.21	**-.51**	US42
N15	*helps us make the links between the different strands of maths and other aspects of our lives.*	.18	.06	-.00	.24	**-.43**	US09
N18	*teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.*	.13	.02	-.04	.23	**.40**	US12
	Factor correlations						
	Factor 1: Commitment to Students and their Learning	--					
	Factor 2: Mathematical Pedagogy	.56	--				
	Factor 3: Student Engagement with the Curriculum	-.61	-.65	--			
	Factor 4: Family and Community	.48	.58	-.57	--		
	Factor 5: Relates Mathematics to the Real World	-.49	-.63	.58	-.53	--	

Explains 60.8% of variance

Cronbach's alpha = .985

Item Response Theory

Table 24 shows the IRT parameters for the 66 items that were used for this trial, along with the mean ratings and point-biserial correlations.

The mean rating on individual items ranged from a high of 4.36 (Item N7) to a low of 2.20 (Item N64), with a mean of 3.51 for all items on the Form. Cronbach's alpha reliability for the whole test was α=.98, which indicates very high internal consistency among the items on the Form.

The *a* parameters for Form November were generally high, ranging from 0.47 (Item N7) to 2.66 (Item N56), with a mean *a* parameter for the 66 items of 1.83. The *b* parameters had the following ranges: $-9.18<b_1<-.48$; $-4.89<b_2<1.04$; $-2.28<b_3<1.86$; $.00<b_4<2.62$; $.93<b_5<3.47$. Item N7 (*My mathematics teacher regards technology (e.g., computers and calculators) as an essential tool for teaching maths*) accounted for the all but one (b_5) of the unusually low values for the lower limits of b_i.

Form Technology

Refinement of Technology Items

To ensure that the domain of excellent mathematics teaching was adequately covered, an item regarding technology was desirable. Items N07 and N57, plus twelve other items (Appendix Five) were written to cover a range of indicators of the possible use of technology in the classroom. They were formatted using the same questionnaire template and tested in two schools with a small sample of 118 students – 58 students from one school and 60 from the other. These schools had not been involved in the previous trials and when approached, agreed to participate. The Head of Department selected the classes and students in the same way as the previous schools. The survey was conducted early in the academic year, and all of the students were in Year 13. Table 25 shows the roll, decile rating, percentage of students by ethnicity and gender in each of these two schools.

Table 24 Classical Test Theory and Item Response Theory Item Statistics for 66 Form November Items

Scale	Item #	Item Mean	Item-Total correlation (r_{pbs})	a (SE)	b_1 (SE)	b_2 (SE)	b_3 (SE)	b_4 (SE)	b_5 (SE)
	1	3.26 (0.07)	.762	2.27 (0.23)	-1.93 (0.17)	-0.49 (0.08)	0.28 (0.09)	1.15 (0.11)	2.18 (0.21)
	2	3.63 (0.07)	.773	1.96 (0.20)	-2.44 (0.23)	-1.09 (0.12)	-0.10 (0.10)	0.87 (0.11)	2.10 (0.20)
	3	3.91 (0.07)	.564	1.21 (0.16)	-3.40 (0.46)	-1.84 (0.24)	-0.52 (0.11)	0.54 (0.15)	2.34 (0.31)
	4	3.78 (0.07)	.652	1.61 (0.17)	-2.55 (0.24)	-1.39 (0.16)	-0.24 (0.11)	0.69 (0.12)	2.05 (0.23)
	5	3.74 (0.08)	.722	1.92 (0.19)	-2.14 (0.20)	-0.95 (0.12)	-0.16 (0.10)	0.58 (0.10)	1.72 (0.17)
	6	3.54 (0.08)	.648	1.38 (0.16)	-2.37 (0.27)	-0.97 (0.15)	-0.04 (0.13)	1.04 (0.15)	2.02 (0.24)
	7	4.36 (0.07)	.337	0.47 (0.58)	-9.18 (2.82)	-4.89 (1.82)	-2.28 (2.55)	0.00 (0.34)	2.65 (2.80)
	8	3.86 (0.07)	.682	1.59 (0.18)	-2.81 (0.34)	-1.57 (0.18)	-.0.33 (0.12)	0.76 (0.13)	1.84 (0.20)
	9	3.80 (0.08)	.787	2.24 (0.21)	-2.00 (0.16)	-1.00 (0.11)	-0.20 (0.09)	0.50 (0.09)	1.49 (0.13)
	10	3.46 (0.09)	.669	1.42 (0.16)	-1.89 (0.23)	-0.68 (0.11)	0.19 (0.12)	0.93 (0.15)	1.95 (0.24)
	11	3.51 (0.08)	.800	2.34 (0.21)	-1.58 (0.13)	-0.77 (0.08)	0.07 (0.09)	0.75 (0.09)	1.68 (0.15)
	12	3.14 (0.08)	.664	1.60 (0.17)	-1.53 (0.18)	-0.59 (0.09)	0.49 (0.12)	1.38 (0.15)	2.48 (0.27)
	13	3.68 (0.09)	.794	2.32 (0.22)	-1.47 (0.14)	-0.78 (0.09)	-0.09 (0.09)	0.60 (0.08)	1.42 (0.13)
	14	3.88 (0.09)	.757	2.01 (0.20)	-1.82 (0.11)	-0.82 (0.11)	-0.29 (0.10)	0.33 (0.10)	1.24 (0.13)
	15	3.63 (0.08)	.702	1.72 (0.18)	-2.16 (0.24)	-1.19 (0.13)	-0.10 (0.11)	0.96 (0.12)	1.97 (0.22)
	16	3.82 (0.07)	.739	2.07 (0.21)	-2.50 (0.23)	-1.30 (0.14)	-0.32 (0.09)	0.62 (0.10)	1.88 (0.18)
	17	3.36 (0.09)	.644	1.36 (0.16)	-1.80 (0.24)	-0.61 (0.13)	0.28 (0.13)	1.01 (0.16)	2.13 (0.27)
	18	3.16 (0.07)	.676	1.45 (0.16)	-2.03 (0.23)	-0.57 (0.13)	0.46 (0.12)	1.63 (0.20)	2.86 (0.38)
	19	3.08 (0.09)	.700	1.74 (0.20)	-0.98 (0.13)	-0.32 (0.12)	0.42 (0.11)	1.02 (0.13)	1.96 (0.21)
	20	3.82 (0.08)	.629	1.40 (0.17)	-2.72 (0.32)	-1.37 (0.17)	-0.36 (0.13)	0.71 (0.13)	2.02 (0.25)
	21	3.72 (0.08)	.703	1.79 (0.19)	-2.38 (0.22)	-1.02 (0.13)	-0.17 (0.10)	0.74 (0.11)	1.77 (0.19)
	22	3.29 (0.09)	.784	2.28 (0.20)	-1.12 (0.13)	-0.37 (0.10)	0.24 (0.09)	0.83 (0.10)	1.79 (0.16)
	23	3.18 (0.08)	.786	2.63 (0.26)	-1.15 (0.11)	-0.40 (0.10)	0.30 (0.08)	1.02 (0.10)	1.98 (0.18)
	24	3.35 (0.09)	.693	1.73 (0.18)	-1.60 (0.16)	-0.51 (0.10)	0.19 (0.11)	0.95 (0.12)	1.84 (0.19)
	25	4.07 (0.08)	.695	1.58 (0.18)	-2.64 (0.30)	-1.69 (0.18)	-0.54 (0.09)	0.36 (0.11)	1.53 (0.18)
	26	3.72 (0.08)	.765	2.05 (0.20)	-2.35 (0.21)	-1.07 (0.12)	-0.10 (0.10)	0.68 (0.10)	1.73 (0.17)
	27	3.25 (0.08)	.781	2.55 (0.23)	-1.66 (0.13)	-0.42 (0.07)	0.30 (0.09)	0.98 (0.10)	1.90 (0.16)
	28	4.05 (0.09)	.782	2.08 (0.22)	-1.74 (0.18)	-1.02 (0.12)	-0.36 (0.10)	0.26 (0.09)	0.98 (0.11)
	29	2.61 (0.08)	.680	1.65 (0.20)	-0.95 (0.15)	0.11 (0.11)	0.98 (0.14)	1.91 (0.21)	2.83 (0.35)
	30	4.08 (0.09)	.799	2.37 (0.24)	-2.03 (0.19)	-1.12 (0.11)	-0.41 (0.06)	0.24 (0.09)	1.05 (0.11)
	31	3.85 (0.08)	.776	2.09 (0.21)	-1.99 (0.18)	-1.16 (0.12)	-0.26 (0.10)	0.44 (0.09)	1.52 (0.14)
	32	3.73 (0.09)	.765	2.15 (0.21)	-1.80 (0.15)	-0.95 (0.11)	-0.11 (0.09)	0.59 (0.09)	1.40 (0.14)
	33	3.44 (0.07)	.814	2.61 (0.25)	-2.03 (0.17)	-0.86 (0.08)	0.20 (0.08)	1.03 (0.10)	1.94 (0.16)
	34	3.95 (0.07)	.701	1.56 (0.19)	-3.10 (0.39)	-1.50 (0.17)	-0.36 (0.12)	0.51 (0.11)	1.86 (0.21)
	35	3.73 (0.07)	.740	2.14 (0.20)	-2.50 (0.27)	-1.25 (0.13)	-0.16 (0.09)	0.84 (0.11)	1.85 (0.17)
	36	2.46 (0.07)	.662	1.56 (0.18)	-0.76 (0.13)	0.33 (0.12)	1.25 (0.16)	2.17 (0.24)	3.25 (0.44)

Scale	Item #	Item Mean	Item-Total correlation (r_{pbs})	a (SE)	b_1 (SE)	b_2 (SE)	b_3 (SE)	b_4 (SE)	b_5 (SE)
	37	3.10 (0.08)	.767	2.37 (0.23)	-1.41 (0.13)	-0.39 (0.10)	0.48 (0.08)	1.31 (0.11)	2.17 (0.21)
	38	3.15 (0.08)	.749	2.01 (0.21)	-1.52 (0.15)	-0.43 (0.12)	0.48 (0.09)	1.28 (0.12)	2.21 (0.25)
	39	4.19 (0.08)	.756	1.94 (0.20)	-2.51 (0.25)	-1.47 (0.15)	-0.62 (0.09)	0.29 (0.10)	1.22 (0.13)
	40	3.70 (0.07)	.729	1.94 (0.19)	-2.18 (0.21)	-1.21 (0.13)	-0.24 (0.10)	0.82 (0.11)	1.93 (0.20)
	41	3.40 (0.09)	.803	2.54 (0.24)	-1.15 (0.12)	-0.50 (0.07)	0.13 (0.08)	0.71 (0.09)	1.62 (0.13)
	42	3.19 (0.08)	.689	1.82 (0.20)	-0.54 (0.18)	-0.46 (0.08)	0.40 (0.11)	1.20 (0.14)	2.21 (0.24)
	43	3.60 (0.08)	.718	1.93 (0.21)	-2.00 (0.19)	-0.91 (0.12)	0.00 (0.10)	0.90 (0.11)	1.75 (0.19)
	44	3.97 (0.08)	.784	2.19 (0.23)	-2.12 (0.19)	-1.32 (0.12)	-0.35 (0.10)	0.40 (0.09)	1.46 (0.13)
	45	3.61 (0.08)	.758	2.08 (0.20)	-2.15 (0.19)	-0.90 (0.12)	0.01 (0.09)	0.83 (0.10)	1.73 (0.17)
	46	4.02 (0.08)	.710	1.35 (0.16)	-2.50 (0.30)	-1.62 (0.21)	-0.60 (0.13)	0.38 (0.13)	1.73 (0.23)
	47	3.15 (0.09)	.627	1.43 (0.17)	-1.31 (0.19)	-0.41 (0.10)	0.47 (0.13)	1.28 (0.17)	2.02 (0.25)
	48	3.75 (0.09)	.776	2.12 (0.21)	-1.47 (0.15)	-0.77 (0.10)	-0.13 (0.09)	0.46 (0.09)	1.26 (0.12)
	49	3.67 (0.08)	.655	1.37 (0.17)	-2.18 (0.26)	-1.05 (0.15)	-0.14 (0.13)	0.78 (0.14)	2.00 (0.25)
	50	3.48 (0.08)	.660	1.56 (0.18)	-2.22 (0.24)	-0.76 (0.11)	0.03 (0.11)	1.04 (0.14)	2.22 (0.25)
	51	3.27 (0.08)	.643	1.40 (0.16)	-1.54 (0.20)	-0.64 (0.12)	0.20 (0.12)	1.26 (0.18)	2.64 (0.33)
	52	3.42 (0.07)	.739	1.99 (0.20)	-1.84 (0.18)	-0.89 (0.12)	0.18 (0.10)	1.09 (0.12)	2.25 (0.24)
	53	2.98 (0.08)	.548	1.14 (0.15)	-1.85 (0.28)	-0.38 (0.12)	0.90 (0.18)	1.92 (0.28)	3.09 (0.46)
	54	3.29 (0.08)	.697	1.95 (0.18)	-1.57 (0.17)	-0.50 (0.08)	0.30 (0.10)	1.07 (0.12)	2.06 (0.21)
	55	2.05 (0.07)	.569	1.36 (0.19)	-0.15 (0.14)	1.04 (0.16)	1.86 (0.24)	2.57 (0.35)	3.25 (0.49)
	56	3.46 (0.08)	.813	2.66 (0.24)	-1.75 (0.14)	-0.78 (0.08)	0.21 (0.08)	0.91 (0.09)	1.63 (0.14)
	57	3.34 (0.08)	.647	1.40 (0.17)	-2.12 (0.27)	-0.71 (0.11)	0.20 (0.12)	1.23 (0.17)	2.65 (0.34)
	58	3.91 (0.07)	.614	1.40 (0.16)	-2.97 (0.38)	-1.65 (0.19)	-0.36 (0.13)	0.61 (0.14)	2.05 (0.26)
	59	3.63 (0.07)	.610	1.46 (0.15)	-2.85 (0.34)	-1.26 (0.15)	-0.05 (0.13)	1.03 (0.15)	2.37 (0.28)
	60	3.54 (0.09)	.703	1.87 (0.19)	-1.61 (0.17)	-0.67 (0.10)	0.10 (0.09)	0.65 (0.11)	1.70 (0.17)
	61	3.58 (0.08)	.742	1.82 (0.20)	-2.24 (0.23)	-0.98 (0.13)	0.01 (0.10)	0.91 (0.12)	2.05 (0.23)
	62	3.49 (0.07)	.724	1.94 (0.20)	-2.12 (0.22)	-0.93 (0.12)	0.12 (0.10)	1.14 (0.12)	2.07 (0.21)
	63	3.67 (0.07)	.735	2.21 (0.20)	-2.24 (0.20)	-1.04 (0.11)	-0.23 (0.09)	0.87 (0.10)	1.83 (0.17)
	64	2.20 (0.07)	.558	1.35 (0.18)	-0.48 (0.11)	0.72 (0.14)	1.70 (0.22)	2.62 (0.34)	3.47 (0.50)
	65	3.50 (0.09)	.671	1.53 (0.19)	-1.44 (0.18)	-0.68 (0.12)	0.10 (0.11)	0.83 (0.12)	1.67 (0.20)
	66	3.78 (0.10)	.702	1.86 (0.20)	-1.30 (0.16)	-0.55 (0.09)	-0.08 (0.10)	0.37 (0.10)	0.93 (0.11)

Note: SE in brackets

Table 25 School Descriptives for Technology Trial

School	Roll (N)	Decile (SES)	Ethnicity (%)					Gender (%)	
			NZE	M	PI	A	O	F	M
Matangi	1100	3	35	17	27	13	8	48	52
Shortland	1200	10	76	3	0	12	9	47	53

The number of students, their ethnic affiliation and gender are shown in Table 26.

Table 26 Participant Descriptives for Technology Trial

School	N	Ethnicity (N)					Gender (N)	
		NZE	M	PI	A	O	F	M
Matangi	58	15	4	13	18	8	31	27
Shortland	60	24	1	0	19	15	20	40
Totals	118	39	5	13	37	23	51	67
Percent		32	4	11	31	19	42	57

As with the other trials, Maori and Pacific Island students were under-represented, while New Zealand European and Asian students predominated. One student from Shortland College was unclassified for ethnicity.

Descriptive statistics

The number of responses, mean, standard deviation, skew and kurtosis statistics for each of the 14 items in Form Technology is shown in Table 27.

The mean rating on individual items ranged from a maximum of 4.34 (Item T05) to a minimum of 2.14 (Item T11), with a mean rating for all items on Form Technology of 3.39. Cronbach's alpha reliability for the whole test was α=0.908, indicating high internal consistency among the items on the Form. For the Form Technology, 90.8% of

the observed score variance is due to variation in the performance of individuals, while 9.2% will be due to error. As with all the other trialled forms, none of the items had a mean rating in excess of 5.

Factor analysis

The 14 items of Form Technology were subjected to maximum likelihood factor analysis, with oblimin rotation. The Kaiser-Meyer-Olkin value was a "meritorious" .88, clearly exceeding the value of .6 recommended by Kaiser (1974). Three interpretable factors were extracted, explaining 65.1% of the total variance. The goodness of fit statistic $\chi^2(52) = 77.98$, $p < .05$ indicates good specification of the three factor model.

The pattern matrix for these three factors is shown in Table 28.

Factor One deals with the teachers' Use of Computer Technology for Teaching; Factor Two is concerned with the teachers' Use of Electronic Technology; and, Factor Three concerns the Use of Calculators for Teaching.

Factor One, Use of Computer Technology for Teaching, (eigenvalue = 6.45) accounts for 46.1% of the common variance; Factor Two, Use of Electronic Technology, (eigenvalue = 1.48) accounts for 10.58% of the variance; Factor Three, Use of Calculators for Teaching (eigenvalue = 1.18) accounts for 8.45% of the common shared variance.

The correlations between factors range from .32 to .55. The correlation between Factor One and Factor Two was .55, indicating that these two factors do not describe unique dimensions of exemplary mathematics teaching with regard to the use of technology. The other two factors are moderately correlated, and represent distinct but overlapping dimensions of the construct.

Table 27 Descriptive Statistics for Form Technology

	My mathematics teacher …	N	Mean		Standard Deviation	Skewness		Kurtosis	
			M	SE	SD	Statistic	SE	Statistic	SE
T01	*enjoys teaching maths using a computer.*	119	3.32	.12	1.35	-.03	.22	-.94	.44
T02	*is very confident when using a computer in our maths lessons.*	118	3.38	.14	1.54	.09	.22	-1.12	.44
T03	*enjoys the challenge of using a computer to solve problems.*	118	3.32	.14	1.48	.20	.22	-.89	.44
T04	*integrates the use of calculators and computers into their teaching of maths.*	117	4.03	.12	1.33	-.40	.22	-.43	.44
T05	*believes that calculators can help us to learn maths.*	119	4.34	.13	1.44	-.64	.22	-.41	.44
T06	*uses computer and calculator technology to enhance remedial instruction.*	112	3.69	.12	1.30	-.06	.23	-.65	.45
T07	*uses calculators and computers to motivate us.*	118	3.40	.12	1.30	.17	.22	-.66	.44
T08	*uses computers to help us work with each other.*	117	2.97	.14	1.55	.45	.22	-.85	.44
T09	*uses modern technology (e.g., computers, calculators, internet) to help us learn maths.*	118	3.74	.14	1.54	-.08	.22	-.91	.44
T10	*makes having computers available in maths fun.*	119	3.24	.14	1.57	.19	.22	-1.11	.44
T11	*uses e-mail and the internet to provide a better learning environment.*	118	2.14	.14	1.52	1.09	.22	-.09	.44
T12	*extends our understanding in maths by using challenging computer-based problems.*	119	2.77	.14	1.53	.45	.22	-.90	.44
T13	*regards technology (e.g., calculators and computers) as an essential tool for teaching maths.*	117	3.83	.12	1.32	-.16	.22	-.51	.44
T14	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	118	3.69	.15	1.60	-.26	.22	-.97	.44
	Valid N (listwise)	106							

Table 28 Summary of Items and Factor Loadings for Oblimin Three-Factor solution for Form Technology

	Item: My mathematics teacher …	Factors 1	2	3	Used in N/US
T01	*enjoys teaching maths using a computer.*	**.86**	-.01	-.13	
T02	*is very confident when using a computer in our maths lessons.*	**.83**	-.12	-.09	
T03	*enjoys the challenge of using a computer to solve problems.*	**.79**	.06	-.04	
T10	*makes having computers available in maths fun.*	**.57**	-.27	.07	
T04	*integrates the use of calculators and computers into their teaching of maths.*	**.55**	.09	.30	
T09	*uses modern technology (e.g., computers, calculators, internet) to help us learn maths.*	**.38**	-.14	.35	
T14	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	**.36**	-.17	.18	N57 US42
T11	*uses e-mail and the internet to provide a better learning environment.*	-.12	**-.99**	-.04	
T12	*extends our understanding in maths by using challenging computer-based problems.*	.17	**-.75**	-.00	
T08	*uses computers to help us work with each other.*	.36	**-.45**	.15	
T05	*believes that calculators can help us to learn maths.*	-.14	.07	**.72**	
T06	*uses computer and calculator technology to enhance remedial instruction.*	.27	-.03	**.58**	
T13	*regards technology (e.g., calculators and computers) as an essential tool for teaching maths.*	.05	-.25	**.53**	
T07	*uses calculators and computers to motivate us.*	.23	-.28	**.31**	
	Factor correlations	1	2	3	
	Factor 1	--			
	Factor 2	-.55	--		
	Factor 3	.39	-.33	--	

Explains 65.1% of variance

Cronbach's alpha = .908

Table 29 Classical Test Theory and Item Response Theory Item Statistics for 14 Form Technology Items

Item #	Item Mean	Item-total correlation r_{pbs}	*a* (SE)	b_1 (SE)	b_2 (SE)	b_3 (SE)	b_4 (SE)	b_5 (SE)
T01	3.32 (0.12)	.642	1.88 (0.26)	-1.96 (0.23)	-0.62 (0.20)	0.12 (0.15)	1.06 (0.22)	2.50 (0.43)
T02	3.38 (0.14)	.734	2.52 (0.38)	-1.55 (0.18)	-0.50 (0.14)	0.09 (0.14)	0.69 (0.15)	1.61 (0.21)
T03	3.32 (0.14)	.601	1.60 (0.27)	-1.93 (0.34)	-0.66 (0.21)	0.27 (0.18)	1.10 (0.27)	2.00 (0.38)
T04	4.03 (0.12)	.579	1.25 (0.27)	-3.17 (2.72)	-1.94 (0.38)	-0.87 (0.24)	0.42 (0.25)	1.85 (0.62)
T05	4.34 (0.13)	.273	0.42 (0.18)	-7.20 (3.75)	-5.03 (2.67)	-2.59 (1.23)	-0.32 (0.50)	2.48 (1.35)
T06	3.69 (0.12)	.611	1.34 (0.48)	-3.03 (1.35)	-1.58 (1.01)	-0.20 (0.23)	0.82 (0.39)	2.26 (0.73)
T07	3.40 (0.12)	.591	1.52 (0.55)	-2.55 (1.56)	-1.01 (0.21)	0.20 (0.19)	1.11 (0.37)	2.33 (0.73)
T08	2.97 (0.14)	.757	2.47 (0.32)	-1.15 (0.20)	-0.11 (0.16)	0.42 (0.14)	1.13 (0.18)	1.69 (0.24)
T09	3.74 (0.14)	.638	1.47 (0.70)	-2.20 (0.34)	-1.37 (0.23)	-0.23 (0.20)	0.69 (0.35)	1.44 (0.50)
T10	3.24 (0.14)	.778	2.25 (0.29)	-1.41 (0.18)	-0.38 (0.14)	0.16 (0.15)	0.83 (0.15)	1.65 (0.25)
T11	2.14 (0.14)	.567	1.69 (0.32)	0.12 (0.19)	0.73 (0.20)	1.15 (0.23)	1.79 (0.32)	2.62 (0.59)
T12	2.77 (0.14)	.687	1.91 (0.58)	-0.89 (0.21)	-0.04 (0.15)	0.50 (0.18)	1.39 (0.25)	2.22 (0.39)
T13	3.83 (0.12)	.572	1.12 (0.56)	-3.24 (2.41)	-2.05 (0.43)	-0.34 (0.25)	0.83 (0.47)	2.22 (1.01)
T14	3.69 (0.15)	.545	1.05 (0.25)	-2.09 (0.44)	-1.37 (0.32)	-0.37 (0.25)	0.73 (0.32)	2.03 (0.86)

Note: SE is in brackets

Item Response Theory

The *a* parameters for Form Technology were moderate to high, ranging from .42 (Item T5) to 2.14 (Item T11), with a mean *a* parameter for the 14 items of 1.61. The *b* parameters had the following ranges: $-7.20<b_1<.42$; $-5.03<b_2<.73$; $-2.59<b_3<1.15$; $-.32<b_4<1.79$; $1.44<b_5<2.62$. Item T11 accounted for the all of the extremes for the upper limits of $b_{i.}$, while Item T5 accounted for all but one (b_5) of the extremes for the lower limits of b_i. Unlike Forms A, B, and C, the SE for $b_{1.}$ was not generally higher than the SE for the other threshold parameters. Highest values of the SE were estimated for b_5 in 9 of the 14 items, with the other 5 items having their highest value of the SE occurring for b_1. The reluctance of students to use the endpoints may account for this, although these high SE values may also be attributable to the small N for this questionnaire.

Item selection from Study Two for SEAT-M

As with item selection from previous forms, items with poor IRT characteristics, or a weak loading on the interpretable factors were considered for elimination. At the same time, the deletion of items was moderated by the need to maintain an adequate mapping of the NBPTS Standards document. This latter consideration became more of an imperative than in the previous stages, as there was no opportunity to over specify the domain of highly accomplished teaching in the same way as the previous filtering process.

Table 30 shows the set of fifty items selected from Form November with their original factor loadings and IRT characteristics. The items in this table are arranged according to the factor analysis, with loadings above .30 only shown.

All of the selected items had a factor loading of at least .32, the slope index for all items was in excess of 1.14, with three-quarters of the items in excess of 1.50. This indicates that the items have strong factor loadings and discriminate well among teachers of differing proficiency on the traits described by the items.

Table 30 Factor loadings and IRT parameters for 50 items selected from Form November

Item	Factor loadings					IRT parameters					
	1	2	3	4	5	*a*	b_1	b_2	b_3	b_4	b_5
Commitment to students and learning											
N14	**.61**					2.01	-1.82	- .82	-.29	.33	1.24
N17	**.58**					1.36	-1.80	- .61	.28	1.01	2.13
N48	**.53**					2.12	-1.47	- .77	-.13	.46	1.26
N11	**.51**					2.34	-1.58	- .77	.07	.75	1.68
N28	**.51**					2.08	-1.74	-1.02	-.36	.26	.98
N30	**.45**					2.37	-2.03	-1.12	-.41	.24	1.05
N16	**.45**					2.07	-2.50	-1.30	-.32	.62	1.88
N25	**.43**					1.58	-2.64	-1.69	-.54	.36	1.53
N44	**.42**	.40				2.19	-2.12	-1.32	-.35	.40	1.46
N05	**.39**					1.92	-2.14	- .95	-.16	.58	1.72
N26	**.39**					2.05	-2.35	-1.07	-.10	.68	1.73
N21	**.38**				.37	1.79	-2.38	-1.02	-.17	.74	1.77
N24	**.37**					1.73	-1.60	- .51	.19	.95	1.84
N66	**.37**		.37			1.86	-1.30	- .55	-.08	.37	.93
N31	**.36**		.31			2.09	-1.99	-1.16	-.26	.44	1.52
Mathematical Pedagogy											
N62		**.73**				1.94	-2.12	- .93	.12	1.14	2.07
N61		**.62**				1.82	-2.24	- .98	.01	.91	2.05
N58		**.54**				1.40	-2.97	-1.65	-.36	.61	2.05
N63		**.53**				2.21	-2.24	-1.04	-.23	.87	1.83
N34		**.47**				1.56	-3.10	-1.50	-.36	.51	1.86
N43		**.43**				1.93	-2.00	- .91	.00	.90	1.75
N60	.33	**.43**				1.87	-1.61	- .67	.10	.65	1.70
N40		**.42**				1.94	-2.18	-1.21	-.24	.82	1.93
N39		**.42**				1.94	-2.51	-1.47	-.62	.29	1.22
N59		**.42**				1.46	-2.85	-1.26	-.05	1.03	2.37

Item	Factor loadings					IRT parameters					
	1	2	3	4	5	*a*	b_1	b_2	b_3	b_4	b_5
N46		**.37**				1.35	-2.50	-1.62	-.60	.38	1.73
N33		**.37**				2.61	-2.03	- .86	.20	1.03	1.94
N35		**.36**				2.14	-2.50	-1.25	-.16	.84	1.85
N08		**.33**				1.59	-2.81	-1.57	-.33	.76	1.84
N42		**.32**				1.82	- .54	- .46	.40	1.20	2.21
Student Engagement with the Curriculum											
N19			**.83**			1.74	- .98	- .32	.42	1.02	1.96
N65			**.82**			1.53	-1.44	- .68	.10	.83	1.67
N23			**.69**			2.63	-1.15	- .40	.30	1.02	1.98
N12			**.62**			1.60	-1.53	- .59	.49	1.38	2.48
N01			**.54**			2.27	-1.93	- .49	.28	1.15	2.18
N51			**.47**			1.40	-1.54	- .64	.20	1.26	2.64
N13	.39		**.40**			2.32	-1.47	- .78	-.09	.60	1.42
N09			**.34**			2.24	-2.00	-1.00	-.20	.50	1.49
Family and Community											
N55				**.76**		1.36	- .15	1.04	1.86	2.57	3.25
N36				**.62**		1.56	- .76	.33	1.25	2.17	3.25
N64				**.62**		1.35	- .48	.72	1.70	2.62	3.47
N29				**.52**		1.65	- .95	.11	.98	1.91	2.83
N53				**.52**		1.14	-1.85	- .38	.90	1.92	3.09
N47			.32	**.38**		1.43	-1.31	- .41	.47	1.28	2.02
N38				**.34**		2.01	-1.52	- .43	.48	1.28	2.21
Relates Mathematics to the Real World											
N20					**.74**	1.40	-2.72	-1.37	-.36	.71	2.02
N49					**.60**	1.37	-2.18	-1.05	-.14	.78	2.00
N50					**.60**	1.56	-2.22	- .76	.03	1.04	2.22
N57		.31			**.51**	1.40	-2.12	- .71	.20	1.23	2.65
N15					**.43**	1.72	-2.16	-1.19	-.10	.96	1.97
N18					**.40**	1.45	-2.03	- .57	.46	1.63	2.86

Wording and Content analysis of assembled items

The wording of one of the selected items was amended in the light of feedback from students during the administration of the November instrument. Item N30 was changed from *My mathematics teacher is able to explain something in different ways to help us understand* to *My mathematics teacher uses different ways of teaching to help us understand*. The wording of all other items remained exactly as trialled, apart from the deletion of the letter *s* from the word maths to suit the USA abbreviation for mathematics. The reading level of the chosen items was assessed using the Flesch-Kincaid Grade Level score at 8.6, which indicates that a student in Grade 9 or above should be able to read and understand the items in the instrument. As this instrument (SEAT-M) was intended for use with Grade 10-12 students in the USA (the equivalent of Years 11-13 in New Zealand), this represents a very acceptable reading level.

The content of the selected 50 items was mapped back on to the AYA/Mathematics Standards. The purpose of this was to ensure that the domain of exemplary teaching as defined by the Standards was adequately mapped by the items. This was done by matching the wording of each item with the wording of the Standards from which it was derived. Table 31 shows the items that were selected for SEAT-M, their path as items through the trials, the originating NBPTS Standard, and the paragraph in the Standard from which the text was drafted.

The number and proportion of SEAT-M items is representative of the extent of the Standards. However, there were no items covering Standard IX (Reflection and Growth), as this was the one specific Standard that was identified by the Study One focus groups as inappropriate for students to rate. One other Standard had drawn similar comment in the focus groups (Standard III, Knowledge of Mathematics), but items were drafted to cover the subject strands without asking for an assessment specifically about the teacher's knowledge of the strands. Initially, the students were asked to evaluate whether they thought the teacher had shown them how to use the different strands to solve problems in the real world, but this was later changed to whether the teacher had made the respective strands interesting for them. By adopting these approaches, the manner in which the content of mathematics was taught could be assessed.

Table 31 Final set of items for SEAT-M, their development history, and origins in the Standards

Text	Item No			Originating Standard											NBPTS #
My mathematics teacher …	US	Nov	A/B/C	1	2	3	4	5	6	7	8	9	10	11	
This standard mostly covers >>				CS	CS	MC	MP	MP	MP CS	M	MP	NA	FC	FC	
Makes maths come alive in the classroom.	***U01***	***N01***	***B47***												***106***
Skilfully asks questions to help classroom discussion and thinking.	***U02***	***N05***	***C43 amend***					✓							***172***
Teaches us the fundamental processes of mathematical thinking – for example: exploration, interpretation, representation, modelling, and analysis.	***U03***	***N08***	***A50***			✓									***127, 228***
Shows us interesting and useful ways of solving problems.	***U04***	***N09***	***A30***							✓					***191***
Enables us to develop confidence and self esteem in maths.	***U05***	***N11***	***B51***						✓						***179***
Makes geometry interesting for me.	***U06***	***N12***	***A15 amend***			✓									***132***
Creates a positive atmosphere in class where we feel part of a team of learners.	***U07***	***N13***	***B22***						✓						***181***
Adjusts the lesson if we experience difficulties in learning.	***U08***	***N14***	***C60***					✓							***165***
Helps us make the links between the different strands of maths and other aspects of our lives.	***U09***	***N15***	***A05***			✓									***122, 127, 128***
Helps us construct an understanding of the language and processes of maths.	***U10***	***N16***	***A52***		✓										***116***
Uses assessment results to provide extra help/extension to appropriate students.	***U11***	***N17***	***C55***								✓				***199***

Text	Item No			Originating Standard											NBPTS #
My mathematics teacher …	**US**	**Nov**	**A/B/C**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
This standard mostly covers >>				CS	CS	MC	MP	MP	MP CS	M	MP	NA	FC	FC	
Teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.	***U12***	***N18***	***A41***			✓									***124***
Makes calculus interesting for me.	***U13***	***N19***	***A45 amend***			✓									***135***
Helps the class to understand that maths relates to the real world.	***U14***	***N20***	***A57***			✓									***122***
Encourages us to seek more than one solution to problems.	***U15***	***N21***	***A44***							✓					***191***
Makes learning maths satisfying and stimulating.	***U16***	***N23***	***B48***					✓							***163, 164***
Provides time for us to reflect and talk about the maths we are learning.	***U17***	***N24***	***A53***					✓	✓	✓	✓				***167, 183, 192, 194, 200***
Challenges students to think through and solve problems, either by themselves or together as a group.	***U18***	***N25***	***B27***						✓	✓					***182, 192, 194,***
Encourages us to try different techniques to solve problems.	***U19***	***N26***	***A31***			✓									***124***
Is committed to the learning of all the students in the class.	***U20***	***N28***	***B40***	✓					✓						***104, 106, 107, 108, 180***
Involves our families and other teachers in the school to help and support us to learn and continue in maths.	***U21***	***N29***	***B58 amend***	✓									✓		***111, 112, 202, 210***
Is able to explain something in different ways to help us understand.	***U22***	***N30***	***B45***		✓										***116***

Text	Item No			Originating Standard											NBPTS #
My mathematics teacher …	**US**	**Nov**	**A/B/C**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
This standard mostly covers >>				CS	CS	MC	MP	MP	MP CS	M	MP	NA	FC	FC	
Sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.	***U23***	***N31***	***C53***				✓								***156***
Consistently makes decisions about their teaching that will further our learning.	***U24***	***N33***	***C15***				✓								***118, 156***
Explores ideas with us even if the answer is not known in advance.	***U25***	***N34***	***C08***					✓							***166***
Integrates the goals of the curriculum and their knowledge of the students in the class.	***U26***	***N35***	***C09***		✓										***116***
Sometimes involves us and our family in exploring career opportunities.	***U27***	***N36***	***B30***										✓		***213***
Teaches us how to evaluate progress towards our goals.	***U28***	***N38***	***C63***								✓				***200***
Uses examples that help us to understand and learn new ideas.	***U29***	***N39***	***C64***		✓										***117***
Uses a variety of methods to collect, organise, represent and summarise collections of data.	***U30***	***N40***	***A56***			✓									***134***
Uses interesting materials and resources that appeal to different people in the class.	***U31***	***N42***	***B18***				✓								***158***
Teaches us about the fundamental role of proof in establishing the truth of mathematical statements.	***U32***	***N43***	***A54***			✓									***127***
Knows and caters for the problems we commonly encounter in learning new topics.	***U33***	***N44***	***C29***				✓								***156***
Encourages us to place a high value on maths.	***U34***	***N46***	***C32***					✓	✓						***162, 179***
Creates a welcoming environment in the classroom for family members and members of the community.	***U35***	***N47***	***B39***										✓		***213, 214***

Text	Item No			Originating Standard											NBPTS #
My mathematics teacher …	**US**	**Nov**	**A/B/C**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
This standard mostly covers >>				CS	CS	MC	MP	MP	MP CS	M	MP	NA	FC	FC	
Take extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.	***U36***	***N48***	***B54***	✓											***107, 111***
Prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.	***U37***	***N49***	***C40***		✓										***120***
Helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.	***U38***	***N50***	***A21***			✓									***123, 124***
Makes statistics interesting for me.	***U39***	***N51***	***A35 amend***			✓									***134***
Works with other subject teachers to provide for students in the class.	***U40***	***N53***	***B37***											✓	***223***
seeks information from my family about my strengths, interests, habits and home life.	***U41***	***N55***	***B44***										✓		***211***
Teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.	***U42***	***N57***	***A49***			✓									***127***
Applies concepts in realistic settings.	***U43***	***N58***	***C12***			✓									***152***
Gets us to think about the nature and quality of our work.	***U44***	***N59***	***C62***							✓					***187***
Tells us what the purpose of each lesson is.	***U45***	***N60***	***C61***				✓								***156***
Encourages us to test mathematical ideas and discover mathematical principles.	***U46***	***N61***	***A34***			✓									***134***
Helps us apply our growing knowledge in both pure and applied settings.	***U47***	***N62***	***A17***							✓					***190, 193***

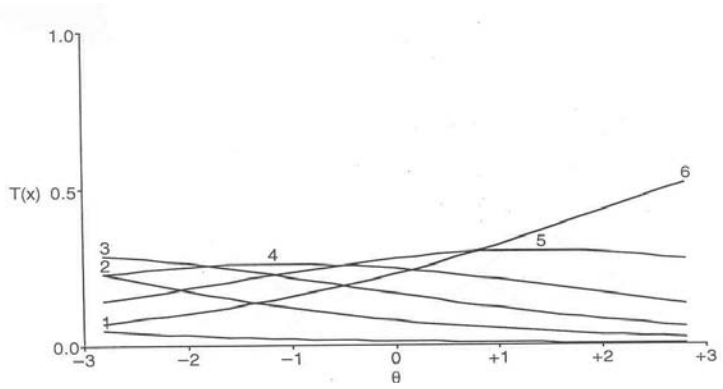
Text	Item No			Originating Standard											NBPTS #
My mathematics teacher …	**US**	**Nov**	**A/B/C**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
This standard mostly covers >>				CS	CS	MC	MP	MP	MP CS	M	MP	NA	FC	FC	
Develops our ability to think and reason mathematically, and have a mathematical point of view.	***U48***	***N63***	***A55***							✓					***190***
Keeps my family informed on a regular basis about my progress in maths.	***U49***	***N64***	***B55***										✓		***211, 212***
Makes algebra interesting for me.	***U50***	***N65***	***A63 amend***			✓									***131***
Compared with all other maths teachers I have had, is the best.	***U51***	***N66***													
Help us to communicate better in maths.	***X***	***N02***	***A47***				✓								***156***
Uses an appropriate range of formal and informal assessments to monitor individual and class progress.	***X***	***N03***	***C59***								✓				***197,198***
Provides time to develop problem solving skills that we can use both in maths and outside the classroom.	***X***	***N04***	***A27***							✓					***192***
Ensures that all students take courses that lead to increased mathematical knowledge.	***X***	***N06***	***C31***	✓											***108***
Regards technology (e.g. calculators and computers) as an essential tool for teaching maths.	***X***	***N07***	***A12 amend***		✓	✓	✓								***159 (117, 125)***
Understands the impact that home life, cultural background, community expectations and student attitudes can have on our learning.	***X***	***N10***	***B20***		✓										***115***
Holds my interest in class.	***X***	***N22***	***B24***	✓											***110***
Stimulates our learning by varying the way we are taught to allow for the strengths and weaknesses of the people in the class.	***X***	***N27***	***B34***				✓	✓							***157, 163***

Text	Item No			Originating Standard											NBPTS #
My mathematics teacher …	**US**	**Nov**	**A/B/C**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
This standard mostly covers >>				CS	CS	MC	MP	MP	MP CS	M	MP	NA	FC	FC	
Encourages us to question and discuss the mathematical ideas and concepts we are taught.	***X***	***N32***	***A61***						✓						***181***
Provides the inspiration for student investigations.	***X***	***N37***	***C23***					✓							***174***
Makes maths meaningful for me.	***X***	***N41***	***C01***	✓											***107***
Allows us to learn maths in different ways.	***X***	***N45***	***B38***		✓										***116***
uses well defined goals to assess our work and learning.	***X***	***N52***	***C42***								✓				***197***
Uses their knowledge about each of us to create problems that are interesting and worth solving.	***X***	***N54***	***B29***		✓										***120***
Helps us experience success in doing worthwhile maths.	***X***	***N56***	***C27***	✓					✓						***111, 182***
Number of items (for USA qnnaire only)				*4*	*6*	*15*	*4*	*6*	*6*	*7*	*3*	*0*	*5*	*1*	***Σ=57***
Number of items (for November qnnaire)				*8*	*10*	*16*	*7*	*8*	*8*	*8*	*5*	*0*	*5*	*1*	***Σ=76***
Percentage of N (USA)				*7.0*	*10.5*	*26.3*	*7.0*	*10.5*	*10.5*	*12.3*	*5.3*	*0*	*8.8*	*1.8*	***Σ=100.0***
Percentage of N (November)				*10.5*	*13.2*	*21.1*	*9.2*	*10.5*	*10.5*	*10.5*	*6.6*	*0*	*6.6*	*1.3*	***Σ=100.0***
Percentage of paragraphs in Standards				*5.9*	*5.9*	*32.1*	*4.8*	*15.5*	*4.8*	*8.3*	*3.6*	*3.6*	*9.5*	*5.9*	***Σ=99.9***
Percentage of words in Standards				*6.8*	*8.6*	*24.7*	*8.1*	*11.8*	*4.6*	*12.1*	*4.5*	*4.9*	*7.9*	*5.9*	***Σ=99.9***

Note: NBPTS # indicates the paragraph in the Standards document that is covered by this item.

At this stage, all Standards were adequately mapped, except that a single item represented the use of technology in teaching (Unit 159). FormNovember contained two items relating to the impact and use of technology in mathematics teaching (Items N07 and N57). One of these items (Item N07: *My mathematics teacher regards technology (e.g., calculators and computers*) as an essential tool for teaching mathematics) produced the highest mean rating (4.36) of all items on FormNovember, but the point bi-serial correlation (.34) was the smallest of all items on Form November. This latter statistic was supported by the Trace lines (Figure 5), which indicated that the responses reflected random error. That is, in terms of the measured proficiency, the responses to this item were randomly spread across the six response points such that the *a* parameter was exceptionally small (0.47, with a standard error of 0.58) and the *b* parameters were spread between –9.18 and 2.65 (with standard errors as high as 2.82). Clearly, this item was not acceptable without further evidence of its suitability. The other item (Item N57: *My mathematics teacher teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths*) had moderate IRT and factor loading characteristics but was included in the Form Technology for further trialling.

Figure 5 Item characteristic curves for an item with random responses, Item N07



Discussion

From a pool of 470 statements, 191 drafted items were field tested with Year 11 to Year 13 students in New Zealand schools in two iterations, and a further fourteen items in a third trial. The application of CTT and IRT test assembly methods using polytomous item responses was central to the process of item selection for a single 51-item

instrument to assess accomplished mathematics teaching. At the same time, the selected items were mapped back on to the domain of accomplished mathematics teaching as articulated in the amended NBPTS AYA Mathematics Standards.

The 51 items selected for SEAT-M (Appendix Six) represent five strong interpretable factors that map the domain of accomplished mathematics teaching, and have good IRT characteristics. The mean loading on the respective factors is .50, and the factor analysis of the November instrument suggests that, when viewed from a student perspective, the Standards could be structured around five dimensions: Commitment to Students and Their Learning; Student Engagement with the Curriculum (Knowledge of Mathematics); Mathematical Pedagogy; Relating Mathematics to the Real World; and, Families and Community. This does not mean that these are the only dimensions that should be written into any proposed Standards for New Zealand, as some of the Standards were not included in the field trials and assessed by the students. This was a deliberate action, as the consensus of the focus groups was that students should not be asked about a teacher's contribution to their professional community, the teacher's reflection and growth outreach, and the breadth and depth of a teacher's mathematical content knowledge.

Using IRT selection criteria, the items discriminate well among teachers with respect to accomplished teaching, with a mean a parameter of 1.81. The SE for the b_1 and b_5 parameters was generally higher than for the other threshold parameters. This is a consequence of the limited use of these end-points on the Likert-type scale. The first response option (Strongly Disagree) was used relatively infrequently, with only 9.7% of all responses in this category, and the sixth response category (Strongly Agree) was used 12.6% of the time. If all response categories were used equally, then the expected values would be 16.7% of the time. Nine items over all Forms had over 33% (twice the expected frequency) of all responses for the item in the first category option category. Similarly, another nine items over all Forms had over 33% of all responses to the item in the sixth category. In essence, the end points were used rarely, with the exception of 18 items. The preponderance of items that are negatively skewed lends support to the minimal use of the first category, and high SE for the b_1 parameter.

The tension between the two methodologies (CTT and IRT) to obtain an optimal set of items is a compromise centred around the need to assemble a test that fully represents the construct of interest, while reducing the potential for "bloated specifics" to inflate the factor analysis. Fletcher (1998, p. 102) noted that manual test assembly does not "always allow for the best test to be assembled", and it is difficult to know whether the items selected are the best combination. Further confirmatory analyses are required to establish the optimal content representation.

This is well illustrated in the case of the items relating to Technology. The two items that had the highest mean teacher ratings (T05 and T11) also had the least desirable IRT characteristics. These two items illustrate the difficulty of choosing items based on either type of analysis, and indeed the tension that can occur between CTT and IRT in item selection.

The descriptive statistics for the items confirm the typical finding in the literature that SETs are negatively skewed, and the high discrimination indices demonstrated that students can be discerning judges of their teachers. Clearly, the moderate proportion (12.6%) of responses in the highest score category (Strongly Agree), and the fact that on a six-point scale only one item (A09) had a mean rating in excess of 5, indicates that secondary students do not capriciously award high ratings to their teachers.

Chapter Five: Study Three

The previous chapter described the development of items for the Students Evaluating Accomplished Teaching – Mathematics (SEAT-M) instrument, and psychometric characteristics of the items and assembled questionnaire. SEAT-M is designed to assess highly accomplished mathematics teaching from a student perspective, using the NBPTS AYA Mathematics Standards to define the construct of accomplished mathematics teaching. This chapter will describe the use of SEAT-M to evaluate 58 mathematics teachers from the USA, 32 of whom are NBCTs. The other 26 teachers were colleagues, who were not National Board Certified. The questionnaire responses of 1611 students were analysed to determine whether the students could reliably distinguish the NBCTs from their colleagues, and which factors and items best discriminated between the two groups.

Procedure

Using the names of all 521 AYA Mathematics NBCTs on the NBPTS website, an Internet search was started to locate the email address of as many NBCTs as possible and invite them to participate. E-mail addresses were obtained for 304 (58.3%), and expressions of interest were received from 53 (17.4% response rate). To maximise the sampling, the NBCTs were also invited to 'snowball' the invitation to other NBCTs. After further correspondence, a total of 26 (8.6% of those originally located and invited) from this group agreed to participate. In November 2001 a further 253 teachers received AYA Mathematics NBCT certification, and a similar approach was made to 131 (51.8%) of them. Replies were received from 35 (26.7% response rate), with 20 (15.2% of those approached) expressing initial interest. From this 2001 group, an additional seven NBCTs agreed to participate, giving a total of 33 Board certified teachers whose classes completed the questionnaire. On checking the status of each NBCT against the official listing, one of the 'snowball' recruits was found to hold the EA/Mathematics certificate, which covers students in the age bracket 11-15. The data

received from her class (and that of her colleague) were deleted from the results and analyses. This left a total of 32 NBCTs in the study.

The invitation to participate outlined the requirements of the study, including the need to invite a non-Board certified colleague to participate. The NBCT was asked to approach any colleague they thought would agree to participate as part of a comparison group. The non-Board certified teacher did not have to "match" the NBCT in terms of experience, qualification, gender or other attributes (although this information was sought and is presented below). Therefore, the two groups are convenience comparative groups and not matched. Seven of the NBCTs were unable to engage a non-NBCT from their school to participate, but one NBCT was successful in engaging two colleagues as participants, giving a total of 26 non-NBCTs in the comparison group. In total, the classes of 58 teachers (32 AYA/Mathematics NBCTs and 26 non-NBCTs) participated.

Each participating teacher received a package that contained participant information sheets and consent forms for the school principal, participating teachers, and for the parent/guardian of students aged 16 years or under at the time of administration. One school district required participant information sheets and consent forms for all students, and these were also provided. The information sheets and the consent forms covered all ethical issues relating to participation in the project, as approved by the University of Auckland Human Subjects Ethics Committee regarding informed consent (AUHSEC Reference 2001/Q/016). The package also contained sufficient questionnaires for the students, an administration manual and report sheet, a sheet to capture teacher demographic details and a pre-paid return envelope. In appreciation for their time and cooperation, all participating teachers also received sets of writing cards depicting scenes from the University of Auckland.

The SEAT-M was administered to students in a Grade 9-12 mathematics class of each participating teacher. The teacher chose which class completed the questionnaire, so the students comprise a convenience sample. An administration manual was supplied to ensure uniform administration of the instrument, and a teacher other than the class teacher was asked to conduct the administration (most commonly, the two teachers exchanged classes).

One school district approved the use of the instrument subject to the deletion the last three words of one item. Item U41 "My mathematics teacher seeks information from my family about my strengths, interest, habits and home life" was ruled to be intrusive because it asked about a student's home life. Therefore, for these students, the item read "My mathematics teacher seeks information from my family about my strengths, interest, and habits". The responses of the students at this school to this item were treated as if they had responded to the complete item.

The teachers completed an administration report to indicate any deviations from the standardised procedure laid out. Seven teachers (12.1%) indicated that they self-administered the questionnaire to their own class, instead of using another teacher for the administration. As part of the administration process, students were given the opportunity to opt out, and 88 students in NBCT classes (8.2% of all students in those classes) and 60 students in non-NBCT classes (8.7%) chose to do so. Therefore, the returned questionnaires represent the opinions of just over 91% of the students in the selected classes. This is a very high rate of return. All teachers reported that they were able to complete the administration process (instructions, distribution of the papers, completion and collection) in the time available for that class.

The wording of all items in the questionnaire had been carefully checked during the New Zealand field trials, but several questions arose, possibly due to differences in the way that mathematics teaching is organised in the two countries, and in terminology and language. Administering teachers were asked to note any student questions that arose during administration. A total of five students (0.003% of all respondents) queried the four items that specifically cover the curriculum strands and their relevance to their class. In this case, it was a consequence of the way in which the curriculum is delivered in the two countries – in the USA, mathematics classes are organised strictly by content (for example, Algebra 1 or Geometry), whereas the New Zealand mathematics curriculum is taught as a single subject. One other student asked for a clarification of the term "pure" mathematics, while another asked a procedural question relating to how to indicate that they had changed their mind regarding one of their ratings. In addition, teachers were asked for their comments and suggestions for improvement. One commented on the time consuming nature of issuing and collecting consent forms (a school district requirement in this case), two found that the instrument was longer than

they thought necessary, two referred to the redundant nature of the curriculum strand items, while three commented that they found the task easy and the administration booklet "really good".

Subjects

Data were collected from the classes of 33 National Board Certified Teachers (NBCTs) and 27 non-certified teachers who participated in this study. As noted above, one of the NBCTs held a certificate for Early Adolescent Mathematics, and was removed from the analysis along with her non-NBCT colleague. Demographic data were obtained from all of the remaining 32 NBCTs and 26 non-NBCTs (Table 32).

Table 32 Descriptive statistics for National Board Certified Teachers (NBCT) and non-National Board Certified Teachers (non-NBCT)

Variable	NBCT		Non-NBCT	
	N (32)	%	N (26)	%
Female	22	68.8%	14	53.8%
Caucasian	30	93.8%	24	92.3%
Masters degree	23	71.9%	11	42.3%
Bachelors degree	6	18.8%	15	57.7%
Doctorate	1	3.1%	0	0%
Basic scale teachers	23	71.9%	23	88.5%
Lead teachers	9	28.1%	4	15.4%
Years of service (mean)	16.9		13.6	
Years of teacher training (mean)	3.0		3.5	
Mean class size	30.6		24.3	

The NBCTs were predominantly female (68.8%), while the proportion of females to males in the comparison group was more evenly distributed There was no significant difference between the two groups in terms of length of teaching service, $F(1,54)=3.630$, $p=.062$, ns. The NBCTs had taught for an average of 16.9 years (SD = 5.2 years) while the non-NBCTs had been teaching for an average of 13.6 years (SD=8.0). There was no significant difference in years of teacher training between the

two groups, $F(1,54) = 0.887$, p=.351, ns. The NBCT teachers had an average of 3.0 years (SD=1.6) of teacher training and the comparison group had an average of 3.5 years (SD=1.7). There was a higher proportion of masters degrees (and one doctorate) among the NBCTs than among the non-NBCTs. Three-quarters of the NBCTs held postgraduate degrees, while fewer than half of the non-NBCTs held a masters degree. Twenty-three of the NBCTs (69.7%) held basic scale teacher positions in their schools, while 84.6% of the comparison group held similar positions. Nine of the NBCTs (27.3%) held positions of responsibility (Lead teacher, coordinator or Chair of the department) compared with four (15.4%) of the comparison group. The teachers in both groups were almost exclusively Caucasian (93.9% of the NBCTs and 88.9% of the non-NBCTs). There was one Hispanic teacher in each group (3% of the NBCTs and 3.8% of the non-NBCTs), one Asian in the comparison group, while one teacher in each group did not provide ethnicity data.

One of the National Board certified teachers was in the original batch of teachers who received their certification in 1997, another one in 1998, fourteen in 1999, nine in 2000 and seven in 2001. Board certification in this subject speciality has been available since 1997 when 47 teachers first gained the AYA/Mathematics certificate, with a further 75 in 1998, 194 in 1999, 205 in 2000 and 253 in 2001. A total of 1430 teachers currently hold NBPTS AYA Mathematics certificates.

The teachers were spread geographically through the USA. Eight NBCTs were from Florida (six non-NBCT colleagues), four from California (three colleagues) and Illinois (three colleagues), three from Ohio (three colleagues) and North Carolina (two colleagues), two from each of Massachusetts (three colleagues) and Mississippi (two colleagues), and one from each of Iowa (no colleague), Maryland (one colleague), Minnesota (one colleague), Oklahoma (no colleague), Washington (one colleague) and Wisconsin (one colleague).

Student participants.

Responses were received from a total of 1611 students in 58 classes. Where students were aged 16 or under, a parent consent form was required. A total of 979 students completed the SEAT-M in the 32 classes of NBCT teachers and 632 students in the 26

non-NBCT classes. In total, 1611 students completed the questionnaire, with a mean of 27.8 students per class (SD=14.2). Class size over the two groups was not significantly different, $F(1,56)=2.893$, $p=.094$, ns.

Instrument/Materials

The 51-item questionnaire developed specifically for this research, Students Evaluating Accomplished Teaching – Mathematics (SEAT-M), was used. It was detailed in the previous chapter. The SEAT-M has a five-factor structure: Commitment to Students and Their Learning; Mathematical Pedagogy; Student Engagement with the Curriculum; Family and Community; and, Relates Mathematics to the Real World. The instrument had high internal consistency when trialled with students in New Zealand, and the items had strong IRT discrimination and location parameters.

Results

Data processing and analysis

The instrument was a scannable document, and processed using the Remark Office OMR v5 software (Principia Products Inc, 2000). The data were stored in Excel files as well as SPSS files for analysis. All teachers who participated received the aggregated results from their class, with some notes on the appropriate interpretation of the results.

In the rare instance where a student had marked three consecutive responses categories for an item, the mean was taken but if the three responses were not consecutive the response for that item was entered as a blank. Where two adjacent responses were made, a random number was generated. If the random number was even, the higher of the two responses was recorded. If the random number was odd, the lower of the two responses was recorded. Non-adjacent responses were entered as a blank. Two students who had completed fewer than 40% of the items were deleted from the analysis.

The number of student responses, mean, standard deviation, skew and kurtosis statistics for the National Board Certified Teachers (NBCT) and non-National Board Certified Teachers (non-NBCT) on each of the fifty-one items are shown in Table 33.

Table 33 Item Statistics for National Board Certified Teachers (NBCT) and non-National Board Certified Teachers (non-NBCT)

Item	NBCT					Non-NBCT				
	N	M	SD	Skew	Kurtosis	N	M	SD	Skew	Kurtosis
U01	977	4.18	1.34	-.56	-.33	631	3.78	1.59	-.21	-1.05
U02	977	4.62	1.24	-.84	.17	631	4.20	1.47	-.48	-.72
U03	975	4.75	1.25	-.89	.20	626	4.27	1.44	*-.58*	*-.52*
U04	979	4.73	1.20	-.97	.50	631	4.25	1.51	-.63	-.63
U05	976	4.20	1.45	-.56	-.50	631	4.04	1.58	-.47	-.86
U06	734	2.70	2.06	-.17	-.98	471	3.38	1.69	-.01	-1.22
U07	971	4.51	1.41	-.75	-.27	626	4.17	1.59	-.55	-.83
U08	977	4.43	1.47	-.77	-.30	628	4.31	1.65	-.69	-.73
U09	977	4.17	1.39	-.52	-.49	632	4.09	1.44	-.46	-.66
U10	976	4.62	1.18	-.79	.32	632	4.16	1.42	-.57	-.51
U11	967	4.21	1.45	-.57	-.46	619	4.12	1.51	-.49	-.76
U12	971	3.98	1.44	-.35	-.69	622	3.72	1.40	-.21	-.83
U13	761	2.97	2.17	-.36	-1.06	436	3.36	1.76	.04	-1.29
U14	965	4.26	1.50	-.58	-.48	626	4.22	1.47	-.55	-.62
U15	975	4.63	1.33	-.82	-.03	632	4.23	1.47	-.62	-.56
U16	977	4.07	1.39	-.48	-.46	628	3.68	1.66	-.21	-1.13

Item	NBCT					Non-NBCT				
	N	M	SD	Skew	Kurtosis	N	M	SD	Skew	Kurtosis
U17	977	4.22	1.50	-.59	-.59	626	3.97	1.59	-.36	-.99
U18	976	5.16	1.15	-1.63	2.68	632	4.56	1.40	-.75	-.32
U19	976	4.76	1.26	-.96	.39	632	4.30	1.38	-.57	-.49
U20	976	4.89	1.32	-1.18	.70	629	4.56	1.54	-.85	-.40
U21	972	3.32	1.54	.13	-1.02	630	3.19	1.55	.13	-1.00
U22	977	4.32	1.36	-.61	-.37	631	4.02	1.58	-.47	-.87
U23	976	4.46	1.35	-.76	-.09	631	4.15	1.63	-.57	-.84
U24	975	4.41	1.29	-.67	-.10	628	4.06	1.52	-.50	-.74
U25	976	4.74	1.26	-.99	.52	625	4.30	1.49	-.66	-.51
U26	973	4.52	1.28	-.78	.20	621	4.09	1.45	-.47	-.67
U27	969	2.75	1.49	.49	-.77	625	2.86	1.54	.45	-.86
U28	972	3.73	1.41	-.20	-.69	625	3.65	1.56	-.14	-1.08
U29	976	4.77	1.23	-.91	.33	627	4.41	1.48	-.69	-.56
U30	969	4.37	1.40	-.56	-.43	624	4.00	1.51	-.39	-.86
U31	972	3.97	1.50	-.35	-.76	626	3.61	1.57	-.14	-1.01
U32	954	4.29	1.51	-.64	-.36	615	3.93	1.53	-.40	-.83
U33	971	4.52	1.39	-.78	-.09	626	4.17	1.54	-.52	-.78
U34	972	4.87	1.31	-1.21	1.05	624	4.40	1.50	-.70	-.52

Item	NBCT					Non-NBCT				
	N	M	SD	Skew	Kurtosis	N	M	SD	Skew	Kurtosis
U35	956	3.96	1.70	-.36	-1.03	617	3.84	1.71	-.27	-1.20
U36	972	4.43	1.50	-.74	-.39	628	4.20	1.64	-.57	-.86
U37	973	4.13	1.47	-.49	-.61	627	4.06	1.53	-.46	-.84
U38	966	4.07	1.51	-.72	-.72	626	3.86	1.56	-.30	-.99
U39	805	3.11	2.06	-.25	-1.06	513	3.48	1.69	-.09	-1.23
U40	970	3.45	1.55	-.01	-1.03	625	3.38	1.64	.09	-.17
U41	967	2.66	1.60	.63	-.77	623	2.69	1.61	.57	-.84
U42	975	3.91	1.46	-.32	-.75	628	3.77	1.57	-.21	-1.05
U43	973	4.33	1.43	-.70	-.22	627	4.09	1.49	-.50	-.66
U44	966	4.44	1.43	-.73	-.19	612	3.97	1.55	-.44	-.88
U45	975	4.45	1.43	-.82	-.08	623	4.22	1.56	-.59	-.73
U46	967	4.39	1.41	-.70	-.28	614	3.93	1.54	-.38	-.86
U47	971	4.27	1.33	-.54	-.33	623	3.88	1.54	-.36	-.90
U48	975	4.57	1.27	-.83	.27	625	4.08	1.57	-.49	-.84
U49	972	3.35	1.73	.10	-1.27	626	3.26	1.67	.13	-1.17
U50	751	3.11	2.18	-.49	-.70	504	3.68	1.80	-.23	-1.31
U51	979	4.42	1.68	-.78	-.68	632	3.87	1.96	-.34	-1.42

The mean rating for NBCTs ranged from a low of 2.66 to a high of 5.16. As with the New Zealand trials, the students were very parsimonious in awarding high ratings to their teachers. Only one item had a mean rating in excess of 5 for the NBCTs (Item U18: *My mathematics teacher challenges students to think through and solve problems, either by themselves or together as a group*). The mean rating for the non-NBCTs ranged from a low of 2.69 to a high of 4.56 (Items U18: *My mathematics teacher challenges students to think through and solve problems, either by themselves or together as a group* and U20: *My mathematics teacher is committed to the learning of all the students in the class*). Item U18 stood out as an item on which all teachers (both NBCTs and non-NBCTs) garnered high mean ratings. Non-NBCTs had a higher mean rating on six items (U06: *"makes geometry interesting for me"*; U13: *"makes calculus interesting for me"*; U27: *"sometimes involves us and our family in exploring career opportunities"*; U39 *"makes statistics interesting for me"*; U41: *"seeks information from my family about my strengths, interests, habits and home life"* and, U50: *"makes algebra interesting for me"*) all of which load on either the Family and Community factor or represent the four curriculum strands in the Student Engagement with the Curriculum factor, while NBCTs had higher mean ratings on all other items. The six items with the lowest mean ratings for NBCTs were U39 (mean rating 3.11), U50 (3.11), U13 (2.97), U27 (2.75), U06 (2.70) and U41 (2.66). For non-NBCTs, the lowest mean ratings were U40 (mean rating 3.38), U13 (3.36), U49 (3.26), U21 (3.19), U27 (2.86) and U41 (2.69).

There are several notable features about the data in Table 33. Firstly, the rate of non-response to the curriculum strand items (Items U6 *geometry*, U13 *calculus*, U39 *statistics* and U50 *algebra*) which had 1205 (74.8% of the 1611 students), 1197 (74.3%), 1318 (81.8%) and 1255 (77.9%) respondents respectively. Of the remaining 47 items, the lowest response rate was 98.0% for Item U44 (1578 respondents). Mathematics classes in the US are organised by curriculum strands, so low response rate for these four curriculum items can be expected, with students responding only to the appropriate curriculum strand. The Administration Manual noted that if students felt that a specific item did not apply, then they should not respond to that item. Given that instruction, these totals represent a higher rate of response than should have occurred. From the class descriptions, 205 students were studying geometry (Item U6), 628

studying calculus or pre-calculus (Item U13), 107 statistics students (Item U39) and 354 algebra students (Item U50). In addition, there were a further 284 of the students who were in classes that studied an integrated course, similar to the New Zealand organisation of the mathematics curriculum (thirty-three students could not be classified by subject). This would give an expected maximum of approximately 489 responses (31.0% of the classified students) for the geometry item, 912 (57.8%) responses to the calculus item, 391 (24.8%) responses to the statistics item, and 638 (40.4%) to the algebra item respectively. Responses to these four items need to be treated with some caution, as it is not possible to determine whether the teacher in question had previously taught the students in those subjects.

Four items had a positive skew for both of the NBCT and non-NBCT groups – these four items all related to the way in which the teacher involved of the student's family in making decisions about the students and the way in which they are taught. (U21 *"involves our families and other teachers in the school to help and support us to learn and continue in math"*, U27 *"sometimes involves us and our family in exploring career opportunities"*, U41 *"seeks information from my family about my strengths, interests, habits and home life"*, and U49 *"keeps my family informed on a regular basis about my progress in math"*). For non-NBCTs, two other items (U13 *"makes calculus interesting for me"*, and U40 *"works with other subject teachers to provide for students in the class"*) also had a positive skew, indicating that students tended to rate teachers towards the less favourable end of the scale for these characteristics. Item U40 along with the four previously mentioned items all load on Factor 4, Family and Community, while U13 loads on the Factor 3, Student Engagement with the Curriculum. Cronbach's alpha reliability coefficient for the instrument was .98, and for each of the factors was .93 (Factor One), .96 (Factor Two), .91 (Factor Three), .89 (Factor Four), and .89 (Factor Five).

A factor analysis of the 1611 cases indicated five interpretable factors. While these factors were not identical to those obtained in the New Zealand trial, there is considerable similarity in the pattern matrix (Table 34). The table shows the factor loadings, as well as indicating the number of the factor each item loaded on in the New Zealand trial (NZ).

Table 34 Summary of Items and Factor Loadings for Oblimin Five-Factor solution for Form SEAT-M

		Factors					
Item	My mathematics teacher …	1	2	3	4	5	NZ
U20	*is committed to the learning of all the students in the class.*	**.54**	.22	.02	.12	.05	1
U08	*adjusts the lesson if we experience difficulties in learning.*	**.48**	.06	.16	.08	.16	1
U36	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.*	**.46**	.23	.01	.27	.05	1
U05	*enables us to develop confidence and self esteem in maths.*	**.42**	.07	.35	.10	.13	1
U11	*uses assessment results to provide extra help/extension to appropriate students.*	**.41**	.04	.02	.26	.15	1
U07	*creates a positive atmosphere in class where we feel part of a team of learners.*	**.40**	.06	.30	.10	.08	3
U17	*provides time for us to reflect and talk about the maths we are learning.*	**.30**	.09	.17	.14	.11	1
U46	*encourages us to test mathematical ideas and discover mathematical principles.*	.01	**.64**	.15	.14	.00	2
U32	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	.01	**.52**	.12	.11	.10	2
U48	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	.10	**.50**	.24	.11	.02	2
U44	*gets us to think about the nature and quality of our work.*	.06	**.49**	.11	.17	.11	2
U19	*encourages us to try different techniques to solve problems.*	.18	**.45**	.14	.04	.08	1
U34	*encourages us to place a high value on maths.*	.14	**.45**	.05	.02	.16	2
U18	*challenges students to think through and solve problems, either by themselves or together as a group.*	.37	**.44**	.05	.10	.04	1
U45	*tells us what the purpose of each lesson is.*	.08	**.44**	.11	.14	.12	2
U47	*helps us apply our growing knowledge in both pure and applied settings.*	.00	**.43**	.20	.17	.18	2
U25	*explores ideas with us even if the answer is not known in advance.*	.25	**.42**	.13	.03	.05	2
U43	*applies concepts in realistic settings.*	.08	**.41**	.04	.13	.40	2
U15	*encourages us to seek more than one solution to problems.*	.11	**.39**	.16	.07	.20	1
U33	*knows and caters for the problems we commonly encounter in learning new topics.*	.34	**.35**	.09	.12	.09	1

Item	My mathematics teacher …	Factors					NZ
		1	2	3	4	5	
U29	*uses examples that help us to understand and learn new ideas.*	.32	**.35**	.10	.07	.13	2
U26	*integrates the goals of the curriculum and their knowledge of the students in the class.*	.27	**.33**	.15	.11	.12	2
U30	*uses a variety of methods to collect, organise, represent and summarise collections of data.*	.15	**.33**	.13	.18	.14	2
U24	*consistently makes decisions about their teaching that will further our learning.*	.27	**.28**	.19	.12	.15	2
U13	*makes calculus interesting for me.*	.09	.05	**.70**	.04	.09	3
U06	*makes geometry interesting for me.*	.00	.07	**.62**	.10	.09	3
U16	*makes learning maths satisfying and stimulating.*	.13	.05	**.61**	.09	.10	3
U50	*makes algebra interesting for me.*	.00	.17	**.55**	.23	.06	3
U01	*makes maths come alive in the classroom.*	.15	.12	**.55**	.01	.09	3
U39	*makes statistics interesting for me.*	.10	.11	**.42**	.35	.06	3
U04	*shows us interesting and useful ways of solving problems.*	.24	.19	**.40**	.08	.16	3
U02	*skilfully asks questions to help classroom discussion and thinking.*	.30	.21	**.33**	.06	.11	1
U23	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.*	.31	.18	**.33**	.03	.13	1
U03	*teaches us the fundamental processes of mathematical thinking – for example: exploration, interpretation, representation, modelling, and analysis.*	.18	.30	**.32**	.12	.12	2
U51	*compared with all other maths teachers I have had, is the best.*	.22	.23	**.31**	.19	.05	1
U10	*helps us construct an understanding of the language and processes of maths.*	.26	.23	**.29**	.07	.22	1
U22	*is able to explain something in different ways to help us understand.*	.24	.18	**.24**	.18	.12	1
U41	*seeks information from my family about my strengths, interests, habits and home life.*	.10	.01	.06	**.83**	.07	4
U27	*sometimes involves us and our family in exploring career opportunities.*	.01	.11	.07	**.77**	.08	4
U49	*keeps my family informed on a regular basis about my progress in maths.*	.03	.07	.05	**.61**	.04	4
U21	*involves our families and other teachers in the school to help and support us to learn and continue in maths.*	.17	.10	.02	**.60**	.17	4
U35	*creates a welcoming environment in the classroom for family members and*	.26	.06	.07	**.55**	.06	4

Item	My mathematics teacher …	Factors 1	2	3	4	5	NZ
	members of the community.						
U40	*works with other subject teachers to provide for students in the class.*	.01	.16	.07	**.53**	.06	4
U28	*teaches us how to evaluate progress towards our goals.*	.12	.12	.12	**.49**	.09	4
U31	*uses interesting materials and resources that appeal to different people in the class.*	.07	.21	.21	**.34**	.11	2
U14	*helps the class to understand that maths relates to the real world.*	.00	.00	.05	.02	**.81**	5
U09	*helps us make the links between the different strands of maths and other aspects of our lives.*	.12	.13	.09	.05	**.81**	5
U37	*prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.*	.02	.25	.10	.26	**.50**	5
U12	*teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.*	.03	.08	.17	.09	**.50**	5
U42	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	.15	.34	.01	.28	**.38**	5
U38	*helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	.08	.34	.01	.24	**.37**	5
	Factor correlations						
	Factor 1: Commitment to Students and their Learning	--					
	Factor 2: Mathematical Pedagogy	-.53	--				
	Factor 3: Student Engagement with the Curriculum	-.58	-.64	--			
	Factor 4: Family and Community	.33	-.51	-.55	--		
	Factor 5: Relates Mathematics to the Real World	-.51	-.62	.67	-.56	--	

Explains 62.7% of variance
Cronbach's alpha = .98

Multivariate analysis of variance (MANOVA)

In the development of the questionnaire, a five-factor structure was interpretable. The five factors were: Commitment to Students and Their Learning; Mathematical Pedagogy; Student Engagement with the Curriculum; Family and Community; and Relates Mathematics to the Real World

A one-way between groups multivariate analysis of variance (MANOVA) was used to ascertain whether there was a significant difference between the NBCTs and the non-NBCTs across the five factor scores describing accomplished mathematics teaching. There was a statistically significant difference between the NBCTs and non-NBCTs on the combined dependent variables (Wilks' Lambda = .921, Mult F (5, 1605) =27.516, $p < .001$). The effect size measured by the partial eta squared was .079, which is a moderate effect. As there were overall differences, the univariate F values were computed to determine the contribution of each factor to the overall difference. (Table 35).

The use of multiple univariate ANOVAs on the dependent variables to explore a significant multivariate effect has been criticised because of the problem of inflated alpha levels which a single multivariate analysis of variance "protects" against (Bray & Maxwell, 1982, p. 341). The use of the Bonferroni procedure is one alternative to overcome inflated alpha levels when employing separate univariate ANOVAs on the individual independent variables (R. J. Harris, 1975). In this case, the alpha criterion for significance (.05) is divided by the number of independent variables (viz., 5) to obtain a modified criterion (alpha = .01). Factors 1, 2, 3, and 5 all reach statistical significance with this revised alpha level.

Further, because of the influence of sample size, statistical significance on its own is no longer regarded as an acceptable index of the effect that is being measured (Jacob Cohen, 1994; L. Cohen, Manion, & Morrison, 2000; Dancey & Reidy, 2002; Thompson, 1994; Thompson & Snyder, 1997; Vacha-Haase, 2001). In addition to testing for statistical significance, an effect size (Cohen's d) was computed for each factor using the factor scores for each student to determine the actual magnitude of the effect being observed. There is a moderate effect size for Factors 1, 2 and 3, and a small

effect for Factors 4 and 5 (J Cohen, 1977). All effects are positive, indicating that students rate NBCTs more favourably than non-NBCTs.

Table 35 One-way analyses of variance for effects of National Board Certification status on five SEAT-M factors

	df	MS	*F*	p	d
Factor 1 Commitment to Student Learning					
Between groups	1	10836.96	41.73	<.001	.32
Within groups	1609	259.72			
Factor 2 Mathematical Pedagogy					
Between groups	1	17718.10	66.43	<.001	.41
Within groups	1609	266.72			
Factor 3 Student Engagement with the Curriculum					
Between groups	1	3438.82	38.85	<.001	.31
Within groups	1609	88.52			
Factor 4 Family and Community					
Between groups	1	148.23	2.05	.153	.07
Within groups	1609	72.40			
Factor 5 Relates Mathematics to the Real World					
Between groups	1	372.51	7.59	.006	.14
Within groups	1609	49.11			

The only factor that did not achieve statistical significance was Factor Four, which describes the teacher's involvement with Family and Community. On each of the other factors, the mean factor scores of the NBCTs were higher than those of the non-NBCTs.

Discriminant Function Analysis

In this study, discriminant analysis is used to assist in addressing whether statistically significant differences exist between the mean ratings obtained by teachers on each of the items for the two a priori defined groups (NBCTs and non-NBCTs), and to determine which of the items account most for these observed differences between the two groups. To assess the extent to which students are able to correctly identify and

classify NBCTs and non-NBCTs, discriminant function analysis was conducted on the dataset, first on the 50 items, then on the five factors.

The Wilk's Lambda (.847) is statistically significant indicating that the two groups had different means across the variables, and that the discriminant function is statistically significant in its ability to predict group membership (chi-square 137.01, df = 51, p <.001). The cases are then classified using the discriminant function to determine the probability that they belong to the NBCT category or the non-NBCT category. In the first instance, only full-data cases (that is, those cases with responses to all 51 items) are considered and classified. The results of this classification are presented in Table 36.

Table 36 Classification Analysis for NBCT status (51 items, 852 full-data cases)

			Predicted group membership		
			NBCT	Non-NBCT	Total
Original group membership	NBCT	N	386	155	541
		%	71.3	28.7	100.0
	Non-NBCT	N	114	197	311
		%	36.7	63.3	100.0

The discriminant function, based on actual responses to all 51 items, is able to correctly classify teachers in over two-thirds of all cases (68.4%). Indeed, slightly more than seven out of ten NBCTs are correctly classified, while over six out of ten non-NBCTs are correctly classified. Press's Q statistic was computed as 168.49, which clearly exceeds the chi-square cut score of 6.635 for an alpha value of .01. The null hypothesis that the hit-ratio of correct classifications using this discriminant function is no better than chance is rejected.

The structure matrix (Table 37) is ordered to show the pooled within-groups correlations between discriminating variables and standardised canonical discriminant function. The item that contributes the most to the discriminant function is Item U18 *"challenges students to think through and solve problems, either by themselves or*

together as a group" with a coefficient of .635, followed by Items U34, U10, U44, U48, U03, U19, and U04 all of which had a coefficient greater than .40. Two items make a negative contribution to the discriminant function - Item U41 *"seeks information from my family about my strengths, interest, habits and home life"* with a coefficient of -.012 and Item U27 *"sometimes involves us and our family in exploring career opportunities"* with a coefficient of -.051. All seven items in the factor Family and Community occurred in the 13 items that made the least contribution to the discriminant function. The items that contribute most to the discriminant function are loaded on Factors One and Two, Commitment to Students and Their Learning and, Mathematical Pedagogy.

Table 37 Structure Matrix of pooled within-groups correlations between 51 SEAT-M items and the standardised canonical discriminant function

Item	Factor	*My mathematics teacher …*	r
U18	1	*challenges students to think through and solve problems, either by themselves or together as a group*	.635
U34	2	*encourages us to place a high value on math*	.530
U10	1	*helps us construct an understanding of the language and processes of math*	.467
U44	2	*gets us to think about the nature and quality of our work*	.456
U48	2	*develops our ability to think and reason mathematically, and have a mathematical point of view*	.441
U03	2	*teaches us the fundamental processes of mathematical thinking – for example: exploration, interpretation, representation, modelling, and analysis*	.428
U19	1	*encourages us to try different techniques to solve problems*	.424
U04	3	*shows us interesting and useful ways of solving problems*	.402
U46	2	*encourages us to test mathematical ideas and discover mathematical principles*	.395
U26	2	*integrates the goals of the curriculum and their knowledge of the students in the class*	.373
U33	1	*knows and caters for the problems we commonly encounter*	.365

Item	Factor	*My mathematics teacher …*	r
		in learning new topics	
U25	2	*explores ideas with us even if the answer is not known in advance*	.359
U47	2	*helps us apply our growing knowledge in both pure and applied settings*	.354
U29	2	*uses examples that help us to understand and learn new ideas*	.354
U51	1	*compared with all other math teachers I have had, is the best*	.348
U24	2	*consistently makes decisions about their teaching that will further our learning*	.339
U15	1	*encourages us to seek more than one solution to problems*	.333
U32	2	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements*	.333
U02	1	*skillfully asks questions to help classroom discussion and thinking*	.330
U01	3	*makes math come alive in the classroom*	.330
U07	3	*creates a positive atmosphere in class where we feel part of a team of learners*	.326
U30	2	*uses a variety of methods to collect, organize, represent and summarize collections of data*	.326
U13	3	*makes calculus interesting for me*	.312
U50	3	*makes algebra interesting for me*	.308
U16	3	*makes learning math satisfying and stimulating*	.300
U31	2	*uses interesting materials and resources that appeal to different people in the class*	.278
U23	1	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile math*	.273
U20	1	*is committed to the learning of all the students in the class*	.250
U45	2	*tells us what the purpose of each lesson is*	.250
U06	3	*makes geometry interesting for me*	.244
U12	5	*teaches us that math is a "science of patterns" with the power to describe significant patterns from the real world*	.242

Item	Factor	*My mathematics teacher …*	r
U22	1	*uses different ways of teaching to help us understand*	.239
U36	1	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in math*	.225
U17	1	*provides time for us to reflect and talk about the math we are learning*	.225
U43	2	*Applies concepts in realistic settings*	.224
U38	5	*helps us to realize that math is continuously evolving and growing to make sense of the world – its order, chaos, stability and change*	.207
U05	1	*Enables us to develop confidence and self esteem in math*	.203
U42	5	*teaches us about the way that math contributes to technological changes in society, and the way that technology has changed math*	.200
U35	4	*creates a welcoming environment in the classroom for family members and members of the community*	.199
U11	1	*uses assessment results to provide extra help/extension to appropriate students*	.197
U28	4	*Teaches us how to evaluate progress towards our goals*	.182
U37	5	*prepares us for adult life by helping us to see how important math will be to our careers and to everyday life*	.150
U21	4	*involves our families and other teachers in the school to help and support us to learn and continue in math*	.147
U08	1	*Adjusts the lesson if we experience difficulties in learning*	.146
U09	5	*helps us make the links between the different strands of math and other aspects of our lives*	.142
U39	3	*makes statistics interesting for me*	.132
U40	4	*works with other subject teachers to provide for students in the class*	.096
U14	5	*helps the class to understand that math relates to the real world*	.091
U49	4	*keeps my family informed on a regular basis about my progress in math*	.085

Item	Factor	*My mathematics teacher …*	r
U27	4	*sometimes involves us and our family in exploring career opportunities*	-.051
U41	4	*seeks information from my family about my strengths, interests, habits and home life*	-.012

The computed eigenvalues, canonical correlation, Wilks' Lambda and chi-square for a discriminant analysis involving the five factors are shown in Table 38, and Box's M statistic is shown in Table 39.

Table 38 Eigenvalues, Canonical correlation, Wilks' Lambda and chi-square for discriminant analysis of SEAT-M five factors

Function	Eigenvalue	Percent of variance explained	Canonical correlation	Wilks' Lambda	Chi square	df	Sig
1	.092	100.0	.290	.916	140.74	5	.000

Table 39 Box's M statistics for five SEAT-M factors

Box's M	*F* approx	Df	Sig
134.76	8.952	15, 7283181	.000

The Box M statistic reaches statistical significance and indicates that the pooled covariance matrix is not appropriate and separate covariance matrices should be used. The high value for Wilks' Lambda forecasts a moderate separation of the groups using the discriminant function. The use of factor scores to discriminate between the a priori groups is not as powerful in achieving separation of the two groups, with a little over 60% of the cases correctly classified.

The Press's Q statistic is calculated as 96.8, which clearly exceeds the table score of 6.635 for chi-square at the .01 alpha level, indicating that the hypothesis that this classification occurred by chance is rejected.

Classification of the groups using the factor scores instead of the item scores is shown in Table 40. The use of factor scores produces a similar separation of the groups to that obtained using the item scores, but is not as efficient or as accurate.

Table 40 Classification summary for NBCT status (5 factors)

			Predicted group membership		
			NBCT	Non-NBCT	Total
Original group membership	NBCT	N	639	340	979
		%	65.3	34.7	100.0
	Non-NBCT	N	268	364	632
		%	42.4	57.6	100.0

Note. 62.3% of original grouped cases correctly classified

Table 41 Structure Matrix of pooled within-groups correlations between 5 SEAT-M factors and the standardised canonical discriminant function.

Factor	r
Factor 2 – Mathematical Pedagogy	.672
Factor 1 – Commitment to Students and Their Learning	.532
Factor 3 – Student Engagement with the Curriculum	.514
Factor 5 – Relates Mathematics to the Real World	.227
Factor 4 – Family and Community	.118

Factor 2, Mathematical Pedagogy, contributes most to the discriminant function (Table 41), while Factor 4 contributes the least. This latter result is not surprising, given the very low position of all of the seven Family and Community items in the item structure matrix above.

Discussion

The results of this investigation indicate that students can and do differentiate between highly accomplished teachers and their colleagues. As such, this lends strength to the

extensive research that shows student evaluations of teaching performance are reliable and valid indicators of teaching effectiveness, and in this case, of highly accomplished teaching. There is a statistically significant difference on the ratings received by NBCTs and non-NBCTs on the SEAT-M instrument. On four of the underlying factors, Commitment to Students and Their Learning, Mathematical Pedagogy, Engagement with the Curriculum, and, Relates Mathematics to the Real World, there are statistically significant differences in favour of the NBCTs, but there is no significant difference on the fifth factor, Family and Community. The effect size is moderate for the first three of these factors, and very small on the other two factors. All effects are positive, which indicates that collectively the ratings favour the NBCTs. At the item level, on all but nine of the items, NBCTs receive significantly better ratings than their non-NBCT counterparts. The SEAT-M discriminant function correctly categorises the NBCTs over 70% of the time, and the non-NBCTs approximately 60% of the time.

These results are encouraging, positive and moderate in effect. Discriminant analysis shows that the SEAT-M correctly classifies the NBCTs seven times out of ten, and non-NBCTs approximately six times out of ten. There is a good separation of the two groups on the diagonal in the Classification Tables, with approximately one-third of the teachers mis-classified according to their status. In an evaluation of the psychometric qualities of the National Board's assessment, Jaeger (1998) calculated that the probability of a false positive was .09, while the probability of a false negative was .20. In all, the probability of a candidate being incorrectly categorised was .13. To achieve this precision, the National Board utilised eleven assessment measures of a candidate's proficiency, each requiring trained teacher/scorers (not necessarily NBCTs themselves) and a complex specification of a performance standard. In the present study, a single assessment of the teacher candidate has been taken, with a false negative probability of approximately .28 and a false positive of .37. The probability of a false negative compares very favourably with Jaeger's outcomes, but the false positive figures are relatively inflated. Successive administrations of the SEAT-M using different classes can improve the reliability of the assessment and hence reduce the likelihood of a false positive or false negative.

The implications of incorrect classifications can be considerable. For schools and the education system, misclassified NBCTs (i.e., teachers whose true scores would be less

than the classification standard, a false positive) create a false impression of what constitutes accomplished teaching, and this does not enhance the professional standing of teaching with any of the stakeholders. In addition, the salary premium that many US school districts pay to NBCTs is misdirected. Likewise, teachers who are awarded a failing score when their true score exceeds the cut score (a false negative) represent a lost opportunity to the profession. These teachers may feel professionally devastated, although they do have the right to 'bank' their passing scores, and re-submit the exercises that were judged to not meet the passing standard. The cost of incorrect classifications can be considerable. The National Board charges a fee of $2000 for each teacher candidate, but the actual cost of certification is greater than this. The cost is subsidised through sponsorship and corporate support. In addition, there are the additional costs to the many schools districts, which offer increased salary enhancements, plus the intangible costs of the time invested by teacher candidates. Incorrect classification represents a poor allocation of these resources.

The structure matrix shows the correlation between each item and the discriminant function. The items that emerge with the strongest correlations tell an encouraging story about accomplished mathematics teaching. There is a complex interaction between a set of characteristics that describe 'good' teaching, with a strong disposition to engage students with mathematics, and to place a high value on that. The teacher challenges and motivates their students to think and reason mathematically, and to adopt a problem solving orientation. These teachers also place a high value on mathematics, and on the nature and quality of the work that their students do. There is a clear expectation that the students will cognitively engage with the mathematics and problems that are presented. Accomplished mathematics teachers build a relationship, not between themselves and the students, but between the students and mathematics. All of this suggests that accomplished mathematics teachers foster and develop an inquiring mathematical mind in their students fed by challenging material and rich tasks, and that the 'drill and routine' approach to learning mathematics is not the mark of our best teachers.

These characteristics go to the heart of the search for the elusive "good" teacher. Extensive research over many decades has failed to find stable correlates of accomplished teaching in terms of years of experience, mathematical qualifications and

knowledge, and attitudes to mathematics. Instead, what emerges from this research is a strong orientation on the part of the teacher to facilitate learning, rather than a set of specific teaching skills – it is not how they teach mathematics, but rather how they assist students to develop their own mathematical skills and knowledge and become mathematical learners. The students are treated as active participants in their own learning, rather than as objects to be taught. The 'style' attributes that Scriven (1988b, p. 135) decries are also relegated to the back seat, and are less important than engaging the students' minds. Accomplished teachers do not score highly just because of the way they ask questions (U02), sequence lessons (U23), or use advance organisers (U45).

At the same time as we see the world of 'good' teaching through student eyes, we also have a clear picture of what they say it is not. Involving the family (and other members of the community) in a student's learning does not mark out accomplished mathematics teachers. There are several ways of viewing this outcome. First, both NBCTs and non-NBCTs really are comparable in this respect, and it is not possible to distinguish between the two groups on this basis. Indeed, neither of the groups performs very well, as teachers in both groups are consistently rated as poor performers in this respect. A second possible explanation is that the students are clearly telling the teachers that they do not want them to become involved with their family and home life. This would be consistent with the peer-oriented socialisation of teenagers, rather than a family focus. Of interest is that this dimension of teaching fared equally poorly in the New Zealand calibration exercises.

In economic terms, the use of student evaluation questionnaires such as the SEAT-M would provide the National Board with a cost-effective, complementary evaluation tool for the Board's assessment profile. At minimum, it could be used to assist teachers screen themselves on their possible chances of attaining NBCT status, at very low cost. The information and feedback obtained could be of much formative value even if teachers then decide not to pursue further assessment of their proficiencies.

Two issues arise from the nature of the samples of teachers involved in the study. First, it could be argued that there is the possibility of a halo effect with regard to the NBCTs. These teachers often attract a lot of favourable publicity on attaining NBCT status, and their students would not be immune from this information. The students may therefore

believe that they have a good teacher, and accordingly assign better ratings to them. This could be especially so if they were in the class that the teacher used for their certification portfolio, with the result that the ratings were inflated. On the other hand, the students in this study were not making ratings that compared two teachers against each other – rather the ratings were comparing the teachers with the Standards, and it was against the Standards that the NBCTs rated well and therefore stood out.

Second, the sample of non-NBCT teachers may not have been typical of all the teachers who are not National Board certified. A number of the structural variables favoured the non-NBCTs. As a group, the non-NBCTs were experienced teachers (M =13.6 years), with a slightly longer period of teacher preparation than the NBCTs, and taught smaller classes than their NBCT counterparts. Two possible explanations may account for the moderate effect observed in this study. First, the non-NBCT teachers were a self-selected group who may have had sufficient confidence in their own ability such that they were prepared to act as a comparison/contrast for a NBCT in this research. To illustrate this point, seven of the NBCTs were not able to recruit a non-Board Certified teacher from their department. Anecdotally, one of the NBCTs noted that the other Mathematics teachers in their school felt that they would be judged against the standard set by the NBCT, and this was threatening to them. The NBCTs had already gone through a process that carefully scrutinised their teaching, but the non-NBCTs may not have had such rigorous exposure. For the non-NBCTs, willingness to participate may have been an act of faith in their own ability. Second, it is not possible to determine whether these non-NBCT teachers were actually highly accomplished teachers, even though they had not submitted to the certification process. A study involving two groups of National Board candidates, those who pass and those who fail, would overcome both of these difficulties.

Chapter Six: Conclusion and Discussion

This research reports on three linked studies concerning a student evaluation of teaching performance instrument (SEAT-M) that has been specifically designed to assess highly accomplished mathematics teaching, and its use to identify those characteristics that, in the eyes of their students, distinguish the highly accomplished mathematics teacher.

In the USA and in New Zealand, there have been frequent calls to encourage high calibre candidates to enter the teaching profession, and retain them in the classroom. The National Board of Professional Teaching Standards (NBPTS) in the USA was formed in 1987 to answer that call by articulating demanding standards for highly accomplished classroom teaching in a broad range of subjects and levels in schools. These high and rigorous standards describe the knowledge, skills and dispositions that distinguish exemplary teachers from their colleagues. A complex assessment programme is required to certify those teachers who meet the Standards, and to fail those who do not. This assessment consists of four portfolio entries (three entries document their teaching supported by video evidence from the classroom, and one accounts for the candidate's contribution as a member of the professional community) plus six 30-minute exercises completed at an assessment centre. In setting out to add the student voice to this assessment regime, it was noticeable that the literature was silent on the use of SETs to identify the best mathematics teachers in high schools, and to determine what the students see as the hallmarks of highly accomplished teaching. A new assessment tool was needed, and the development and validation of this instrument was central to this study.

The knowledge, skills and dispositions described in the NBPTS AYA Mathematics Standards have been subject to considerable consultation, debate, scrutiny and research in the USA, and have stood up well to this intense investigation. The Standards provide

a robust model of accomplished teaching. The eleven Standards are structured around five core propositions that are common to all certificates – teachers are committed to students and their learning; teachers know the subjects they teach and how to teach those subjects to students; teachers are responsible for managing and monitoring student learning; teachers think systematically about their practice and learn from experience; and, teachers are members of learning communities.

While these Standards have received general acceptance in the USA and been used to provide suitable forms of recognition for high performing teachers, that does not mean that the Standards can be applied in other countries where there may be different conceptions of what constitutes an outstanding teacher. Therefore, the first study was to establish the appropriateness of the Standards in describing what exemplary New Zealand teachers know and can do. Two focus groups of highly recommended teachers examined the Standards and approved them for use in New Zealand with some modifications. In essence, they found that there were more similarities than differences between the two countries when it comes to professional teaching Standards. Almost 80% of the content of the Standards can be directly applied without the need for modification or deletion, with almost all of the remaining Standards suitable provided minor modifications were made.

The main areas of difference centred on the details of the mathematics curriculum, the role of the teacher vis-à-vis interaction with the families of their students, and, the impact of the differing assessment regimes in the respective countries. In addition, the focus groups commented on the absence of classroom management and discipline, and inserted a small statement on classroom management and disciplined inquiry in the section of the Standards entitled The Art of Teaching. The teachers also noted the repetitive nature of the Standards, and the idealistic, almost unattainable requirements for certification. At the same time, the participants were asked to indicate any issues they could foresee that might arise when using these Standards as the foundation document for a SET. While it was agreed that students were in a unique position to observe and report on a teacher's classroom work, some reservations were expressed about converting all of the Standards into items for a SET instrument, especially in connection with the teacher's mathematical knowledge, and the teacher's role in the professional community and with the student's family. These findings are in keeping

with the comments of Worrell and Kuterbach (2001) that SETs are not an acceptable way to assess teachers on all aspects of their teaching.

Once minor modifications were made to the documented Standards in line with the findings of the focus groups, the Standards were systematically converted into statements that would form the basis of three trial student questionnaires. A total of 470 statements were drafted, and crafted into 191 items that would go to trial. This reduction from 470 statements to 191 items did not reduce the effective coverage of the Standards, as it had earlier been noted that the Standards were very repetitive. Three different questionnaire forms were assembled from the 191 items, with a general evaluation item added at the end of each questionnaire form. The three questionnaire forms were administered to 452 students in 16 classrooms, such that each class had a random allocation of each form. The questionnaire data were analysed in three ways – traditional CTT analytical tools including factor analysis, IRT analysis for item characteristics, and a mapping of the items against the Standards to retain domain specification. Through several iterations of this process, 65 items with the best estimates of the underlying trait, the best factor loadings and the best content/construct coverage were retained for further trial. At this stage, the wording of the items was also revised to address possible weaknesses in the item revealed through the trial and trial data. A second trial of these items was conducted in New Zealand with 329 students in 22 classes, and a smaller trial of 14 technology items with 118 students in 4 classes. The resulting analysis of this data further reduced the item set to 50 items for inclusion in the Students Evaluating Accomplished Teaching - Mathematics (SEAT-M) instrument. The SEAT-M has a strong five factor structure, and the individual items had good estimates of the accomplished teaching trait.

The third and final study in the research utilised the SEAT-M in the classrooms of 32 National Board Certified Teachers (NBCTs) and 26 colleagues who were not Board certified. This provided an ideal platform for testing the ability of the instrument to discriminate between the two groups of teachers, and answer the question of whether students can discern accomplished teachers from their colleagues. At the same time, those teacher characteristics that most differentiate between the two groups could be examined to better understand what makes for a good teacher, as reported by their students. The rating responses of 1611 students were available for analysis.

Through these three studies, several strong messages emerge. The first is that teachers in the USA and New Zealand view accomplished mathematics teaching in much the same way. There are more similarities than differences. Teaching in each country is a demanding profession, and to achieve at the highest level requires a very special set of skills, knowledge and dispositions. The NBPTS AYA Mathematics Standards detail these, and provide a benchmark for those who aspire to the highest standards of professional practice. The context of teaching in each country may differ in a variety of ways, but the focus groups report that the practice is more alike than different. Therefore, the Standards offer a comprehensive and appropriate gauge for teachers to use as a measure of the best standards of professional practice. For the New Zealand high school mathematics teacher, the NBPTS AYA Mathematics Standards present a picture of what it would be like to be a highly accomplished practitioner in the field, provided some modifications were made to take specific local conditions into account. These changes were minor.

The description of the Standards as very demanding is encouraging, as the good New Zealand teachers in the focus groups believed that for most New Zealand teachers (including themselves), the AYA Mathematics Standards present an almost unachievable, though not impossible, goal. Only the very best teachers would offer themselves for certification, and for them personally, certification seemed very difficult to attain given the nature and quality of the Standards. What is encouraging is that advanced certification of this sort would not be the outcome of course completion and minimal demonstrated competency. While such a high standard would act as a barrier for entry to the profession, there needs to be another goal for teachers to strive for as they improve and advance their professional practice.

New Zealand has a set of professional teaching standards, which are premised on the competencies that are required for continued registration as a teacher. These standards (administered through the Teachers Council) do not attempt to portray the additional knowledge, skills and dispositions that the highly skilled and competent teachers embody. Instead, teachers undergo annual appraisal and attestation processes to show that they have maintained the requirements of the standards for continued registration, and met the annual goals that they have set for themselves. If they wish to show that

they have these special skills and knowledge, they usually do so by seeking promotion to middle and senior management positions, thereby reducing their time in the classroom and the potential they have to turn their students on to mathematics.

The role of the teacher, especially vis-à-vis the involvement of the family and community to enhance student opportunity to learn, clearly differs between the two countries. In the USA, this is regarded as a central tenet of successful and effective teaching. On the other hand, the New Zealand teachers did not see this as a critical or achievable role. The reasons for this may be two-fold. First, many schools place restrictions on the interaction of teachers with families. This is usually done to ensure that a single, consistent communication pathway is maintained between the schools and the family. Teachers are generally limited in their communication with families to the usual report to parents and the parent-teacher interviews that follow the reports. Thus, New Zealand teachers find that it is difficult to elicit additional information that may assist in building a better teaching programme for an individual. Second, limiting the amount of communication between teachers and families is a simple matter of time management. Teachers have a heavy workload related to teaching and associated activities (for instance, preparation, assessment, reporting), and there is limited time available to undertake the additional activities required to establish good relations with families.

While there are marked differences between the espoused theory and theories in use (Argyris & Schon, 1974) regarding family outreach, it is interesting to note that irrespective of the country, the students do not rate their teachers very highly on any of the items relating to this aspect of teaching. The Family and Community factor was the only factor that did not distinguish between the two groups of teachers (NBCTs and non-NBCTs), and the items on this factor were amongst those with the lowest mean rating, as well as providing the least weight to the discriminant function. This could be an outcome of an aspect of teaching which is not well implemented, as the teachers in the focus groups noted, but could also be an artefact of the social distance that teenagers tend to build between themselves and their families. “Keep away from my home” could be the message that teenagers want to convey to their teachers. In both countries, it is not possible to distinguish outstanding teachers on the basis of this factor.

The NBPTS assessments for certification involve a complex set of exercises that require considerable resources to score. The Board has invested heavily in researching new forms of performance assessment especially for this purpose, as well as psychometric analyses of the outcomes to ensure that the assessments measured up to the highest standards of assessment practice. The focus of these evaluations is to have them analysed and scored by the teacher's peers, people who are deemed to have a complete understanding of the behaviours that are being reported by the candidate. In all of this the Board has overlooked those who are arguably in the best position to evaluate the teachers – the students who share the classroom with the teacher day in and day out. The myths that students are capricious, and that they are likely to award their teachers high grades are not supported by this research. On the contrary, it was rare for the mean rating of teachers on any SEAT-M item to exceed five on a six-point scale – one item (A09) in the New Zealand trials, and one item (U18) for the NBCTs (but not the non-NBCTs) fitted into this category. High ratings were not awarded lightly, even though there was a tendency for the students in these studies to give positive ratings, as has been frequently observed in the literature (Bendig, 1952b; Centra & Linn, 1973; Hativa, 1996; Holmes, 1996; K. D. Peterson et al., 2000; Tagomori & Bishop, 1995). From a total of 378 computed mean ratings, the proportion of ratings in excess of five represents approximately half of one percent.

What is more, students can distinguish between those teachers who are high performers and their colleagues. The results from using SEAT-M indicate that students are able to discern the difference between accomplished and other mathematics teachers. On four of the five factors underlying the SEAT-M, significant differences in favour of the NBCT were found. The effect sizes were moderate on three of these factors, and small on the fourth. Students observe and report that NBCTs are noticeably better than their non-NBCT colleagues.

The items that contribute most to this discrimination have a focus on cognitive engagement with the content of the mathematics curriculum, and developing a mathematical way of thinking and reasoning. There is an emphasis on problem solving, and using mathematical language and processes to achieve this. It is what they get the students to do in the class that emerges as the strong component of the NBCT's repertoire, rather than what the teacher specifically does. Students must be actively

involved in their learning, with a focus on multiple paths to problem solving. As mathematical thinkers and problem solvers, the students are also encouraged to go beyond the successful solution of the problem to include the interpretation and analysis of the solution. All the while, students are encouraged to greatly value mathematics and the work that they do in mathematics, and always check the quality of their work to strive for the very best standards.

The teacher and the students work together to co-construct mathematics. To give meaning to the fundamental processes and the language of mathematics, the teacher and students engage in a practical interactive relationship that builds a learning programme to enable the students to receive meaningful and purposeful instruction. This relationship also involves the parents and community, which this research indicates is one of the neglected elements of teaching. This model does not allow for just the transmission model of teaching, whereby there is an information transfer and exchange. Students are involved in the communication process as active participants, and not as receptors of the knowledge and skills that the teacher has. The students are not passive learners. Thus the classroom of a highly accomplished teacher utilises active learning techniques, which include social activity – the problem solving together that the students rate so highly. This construction then is not just an up-down relationship, but sideways as well – the students learn from each other.

Implications

For high school mathematics teachers, the main message is to give high priority to the cognitive aspects of teaching mathematics rather than focus on developing personal links with the students. Students identify highly accomplished teachers by the way that they focus on developing thinking minds that rise to the mathematical challenges presented by the teacher. Teacher educators and professional developers could also emphasise this message. Knowing how to engage students is a key teaching task that should be communicated to teachers and trainee teachers.

As costs increase to assess each candidate's portfolio, there is the potential to develop and include SETs such as the SEAT-M. As a form of triangulation on the existing elements of the NBPTS assessment regime, SEAT-M offers the National Board an

economic alternative to more expensive, people intensive assessments. Alternatives in themselves do not make sense unless they can meet rigorous psychometric standards as well. A cheaper alternative just for the sake of an alternative (*faute de mieux*) is not a sensible option if the Board wishes to maintain a credible certification process. The Board has any number of critics who are quick to attack any move that they see as diluting what they see as the one essential measure of a teacher's worth – improved student learning as measured on standardised tests. Although SETs have been thoroughly researched, any attempt to introduce them as a component of the teacher evaluation portfolio may be met with scepticism. These doubts would most likely be predicated on the ability of the teacher to exercise undue influence on the students when completing the SEAT-M. Such a concern would have considerable cachet, so the instrument in its current form would have its best use as a screening and/or formative tool for those who are considering presenting for Board certification. In this way, the stakes involved in the assessment are reduced and the teacher and students can use the SEAT-M to best advantage. The teacher receives immediate feedback on how they perform, and can act to remedy any faults that are identified. They also receive a good indication of the possibility of success as a candidate.

Another contribution of this study is that this student evaluation instrument was developed directly to a set of standards. Too often SETs are developed by factor analysing many items, discovering a factor pattern, and then declaring that these are the multiple dimensions that needed to be included. While reference is often made to other SET instruments with similar dimensions, this study worked from a well argued and well critiqued set of propositions about accomplished teaching. The methodology to create the items was rigorous in its cross referencing back to the NBPTS Standards, and thus there is some confidence that the items are related to a specific model of excellence in teaching.

Would Jaime Escalante become a NBCT? In his story, Escalante defined four key characteristics – knowing the subject; knowing how to teach it; respecting the students; and, ganas de triunfar (desire on the part of the students). Knowing the subject corresponds to the factor describing Engagement with the Curriculum; knowing how to teach it is parallel to Mathematical Pedagogy; and respecting students is akin to Commitment to Students and Their Learning. Each of these factors clearly

differentiated between the NBCTs and non-NBCTs. In the latter case, although Escalante appears to describe a student characteristic, it is not a case of students arriving primed with ganas. Rather, it is the motivation and hunger for learning that a good teacher develops in the students that makes those students love learning mathematics. There is a clear message in the responses of the students in this study that the NBCTs provided them with challenging learning experiences and rich tasks. Escalante believed that teaching should be "peppered with lively examples, ingenious demonstrations of math at work and linkages between math principles and their real-world applications" (Escalante & Dirmann, 1990, p. 410). Through these examples, he challenged and motivated the students, all the time developing that 'desire to succeed' in his students. In return, it is clear that Jaime's students had tremendous admiration for their teacher, a respect that would be translated into high ratings on all items, including those that address involvement with their family. Escalante established firm links with their families and with the business community, and used these link to the benefit his students. Whether measured by the National Board assessment regime of portfolios and assessment centre exercises, or via the use of SEAT-M, there is no doubt that Jaime Escalante would gain certification as a National Board Certified Teacher.

Future research

To overcome the inherent difficulty arising from the unknown quality of the non-NBCT comparison group, future research would benefit from a study that utilised a comparison group comprised of teachers who were failed candidates for the Board certificate. The comparison group would then consist of teachers who believed that they were highly accomplished, and had been through the preparation process. There is also a distinct psychometric advantage in that it would then be possible to obtain the scores of the two groups, and refine the instrument so that it provides the most information about the cut scores that distinguish NBCTs from non-NBCTs. Once the NBPTS cut score is known, it is possible to develop a test information curve that maximises items at and above the proposed cut score, and then use the instrument to sort the potential candidates. A second benefit of this approach is that it would enable finer calibration and greater precision, with a consequent reduction in the rate of false positives and negatives.

This research has also established that student evaluations can distinguish and identify highly accomplished teachers of mathematics. NBPTS has over 30 different certificates available for highly accomplished teaching. For each of the Standards that define these certificates, different versions of the SEAT instrument could be developed to use in the identification of teachers who meet the Standards. As noted in the discussion, the best use of the SEAT suite would be for teachers considering candidacy for the NBPTS certification process, to assist them in their decision of when or whether to commence the process.

Adapting all of the NBPTS Standards will also enable studies to be conducted into the use to SETs with students of a younger age. The literature on SETs with primary and middle school-age students is very sparse. The NBPTS Standards provide a well-articulated framework for developing an instrument suitable for use with students at these levels. In addition, a series of studies involving successful and unsuccessful candidates would provide the external criterion to determine the validity of using SETs with younger students.

References

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A., & Hess, C. W. (1990). Satisfaction with processes of collecting student opinions about instruction: The student perspective. *Journal of Educational Psychology, 82*, 201-206.

Abrami, P. C. (1977). *The generalizability of student ratings of instruction* (No. ERIC Document Reproduction Service ED139832). New York, NY.

Abrami, P. C. (2001a). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research, 109*, 59-87.

Abrami, P. C. (2001b). Improving judgments about teaching effectiveness: How to lie without statistics. *New Directions for Institutional Research, 109*, 97-102.

Abrami, P. C., & d'Apollonia, S. (1999). Current concerns are past concerns. *American Psychologist, 54*, 519-520.

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*, 219-231.

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 321-367). New York: Agathon Press.

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research, 52*, 446-464.

Aiken, L. R. (1996). *Rating scales and checklists: Evaluating behavior, personality, and attitudes*. New York: John Wiley & Sons.

Aiken, L. R. (1997). *Questionnaires and inventories : surveying opinions and assessing personality*. New York: J. Wiley.

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching & Learning*, 25-31.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153-166.

Aleamoni, L. M., & Graham, M. H. (1974). The relationship between CEQ ratings and instructor's rank, class size and course level. *Journal of Educational Measurement, 11*, 189-202.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science, 9*, 67-84.

Aleamoni, L. M., & Thomas, G. S. (1980). Differential relationships of students, instructor, and course characteristics to general and specific items on a course evaluation questionnaire. *Teaching of Psychology, 7*, 233-235.

Aleamoni, L. M., & Yimer, M. (1973). An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. *Journal of Educational Psychology, 64*, 274-277.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality & Social Psychology, 64*, 431-441.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2002). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18).

Amrein-Beardsley, A. L., & Berliner, D. C. (2003). Re-analysis of NAEP Math and Reading Scores in States with and without High-Stakes Tests: Response to Rosenshine. *Education Policy Analysis Archives, 11*(25).

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Argulewicz, E. N., & O'Keeffe, T. O. (1978). An investigation of signed vs. anonymously completed ratings of high school student teachers. *Educational Research Quarterly, 3*(3), 39-44.

Argyris, C., & Schon, D. A. (1974). *Theory in practice: increasing professional effectiveness*. San Francisco: Jossey-Bass Publishers.

Aubrecht, J. D. (1984). Better faculty evaluation systems. In P. Seldin (Ed.), *Changing Practices in Faculty Evaluation*: Jossey Bass.

Aubrecht, J. D., Hanna, G. S., & Hoyt, D. P. (1986). A comparison of high school student ratings of teaching effectiveness with teacher self-ratings: Factor analytic and multitrait-multimethod analyses. *Educational and Psychological Measurement, 46*, 2223-2231.

Australian Senate Employment Education and Training Reference Committee. (1998). *A class act: Inquiry into the status of the teaching profession*. Canberra: Senate Printing Unit.

Ausubel, D. P. (1978). In defense of advance organizers: A reply to the critics. *Review of Educational Research, 48*, 251-257.

Ausubel, D. P. (1980). Schemata, cognitive structure, and advance organizers: A reply to Anderson, Spiro, and Anderson. *American Educational Research Journal, 17*, 400-404.

Avi-Itzhak, T., & Kremer, L. (1985). An investigation into the relationship between university faculty attitudes toward student rating and organizational and background factors. *Educational Research Quarterly, 10*(2), 31-38.

Ayala, C., & Martin, C. (1997). The development of children's learning conception. *Infancia y Aprendizaje*, No 77.

Babad, E., Avni-Babad, D., & Rosenthal, R. (2003). Teachers' brief nonverbal behaviors in defined instructional situations can predict students' evaluations. *Journal of Educational Psychology, 95*, 553-562.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* (Vol. 129). New York: Marcel Dekker, Inc.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*, 239-251.

Ballou, D., & Podgursky, M. (1998a). The case against teacher certification. *Public Interest, 132*, 17-29.

Ballou, D., & Podgursky, M. (1998b, 10 June). Some unanswered questions concerning national board certification of teachers. *Education Week on the Web, 17,* 39-40.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, N.J.: Prentice Hall.

Bandura, A. (1986). *Social foundations of thought and action : a social cognitive theory*. Englewood Cliffs, N.J.: Prentice-Hall.
Bangert-Drowns, R. L. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213-238.
Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*, 308-314.
Batten, M. (1989). Teacher and pupil perspectives on the positive aspects of classroom experience. *Scottish Educational Review, 21*, 48-57.
Batten, M. (1994). Students perceptions of effective teaching. *Score, 2*(6), 4-5.
Beecher, D. E. (1949). *The evaluation of teaching: Backgrounds and concepts*. Syracuse: Syracuse University Press.
Bendig, A. W. (1952a). A statistical report on a revision of the Miami Instructor Rating Sheet. *Journal of Educational Psychology, 43*, 423-429.
Bendig, A. W. (1952b). The use of student-rating scales in the evaluation of instructors in introductory psychology. *Journal of Educational Psychology, 43*, 167-175.
Bentley, R. R., & Starry, A. R. (1970). Purdue Teacher Evaluation Scale (PTES). Princeton, NJ: Educational Testing Service.
Benz, C., & Blatt, S. J. (1995). Factors underlying effective college teaching: What students tell us. *Mid-Western Educational Researcher, 8*, 27-31.
Bergstrom, B. A., & Lunz, M. E. (1992). Confidence in pass/fail decisions for computer adaptive and paper and pencil examinations. *Evaluation & the Health Professions, 15*, 453-464.
Berk, R. A. (1979). The construction of rating instruments for faculty evaluation. *Journal of Higher Education, 50*, 650-669.
Berliner, D. C. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*(August/September), 5-13.
Berliner, D. C. (1987). Knowledge is power: A talk to teachers about a revolution in the teaching profession. In D. C. Berliner & B. V. Rosenshine (Eds.), *Talks to Teachers: A Festschrift for N L Gage* (pp. 3-33). New York: Random House.
Berry, B., & Ginsberg, R. (1989). Legitimizing subjectivity: Meritorious performance and the professionalization of teacher and principal evaluation. *Journal of Personnel Evaluation in Education, 2*, 123-140.
Biggs, J. B. (1987). *Student approaches to learning and studying. Research Monograph*. Hawthorn, Australia: Australian Council for Educational Research Ltd.
Biggs, J. B. (1996). Stages of Expatriate Involvement in Educational Development: Colonialism, Irrelevance, or What? *Educational Research Journal, 11*, 157-164.
Birenbaum, M., & Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students' knowledge states in multiplication and division with exponents. *Applied Measurement in Education, 6*, 255-268.
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley Publishing.
Blackburn, R. T., & Clark, M. J. (1975). An assessment of faculty performance: Some correlates between administrator, colleague, student and self-ratings. *Sociology of Education, 48*, 242-256.
Blanchard, B. E. (1967). *Illinois Ratings of Teacher Effectiveness Manual. Grades 9-12*. Beverly Hills, CA: Western Psychological Services.
Blank, D. L. (1985). Socratics versus sophists on payment for teaching. *Classical Antiquity, 4*, 1-24.

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review, 18*(4), 48-54.
Bohen, D. B. (2001). Strengthening teaching through national certification. *Educational Leadership, 58*(8), 50-53.
Bond, L., Smith, T., Baker, W. K., & Hattie, J. A. (2000a). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: Center for Educational Research and Evaluation, The University of North Carolina at Greensboro.
Bond, L., Smith, T., Baker, W. K., & Hattie, J. A. (2000b). *A distinction that matters: Why national teacher certification makes a difference*. Washington DC: National Board for Professional Teaching Standards.
Bonner, S. F. (1977). *Education in ancient Rome: From the elder Cato to the younger Pliny*. London: Methuen & Co Ltd.
Boshier, R., & Onn, C. M. (2000). Discursive constructions of web learning and education. *Journal of Distance Education, 15*(2), 1-16.
Boulton-Lewis, G. M., Smith, D. J. H., McCrindle, A. R., Burnett, P. C., & Campbell, J. (2001). Secondary teachers' conceptions of teaching and learning. *Learning & Instruction, 11*, 35-51.
Braskamp, L. A., Brandenburg, D. C., & Ory, J. C. (1984). *Evaluating teaching effectiveness: A practical guide*. Beverly Hills, CA: Sage Publications.
Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*, Retrieved 16 October 2004 from http://epaa.asu.edu/epaa/v2012n2001/.
Braunstein, D. N., Klein, G. A., & Pachla, M. (1973). Feedback expectancy shifts in student ratings of college faculty. *Journal of Applied Psychology, 58*, 254-258.
Bray, J. H., & Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVAs. *Review of Educational Research, 52*, 340-367.
Brooks, P. (1990). *Teacher leniency/strictness and students' grades* (No. ERIC Document Reproduction Service ED322140).
Brown, D. L. (1976). Faculty ratings and student grades: A university-wide multiple regression analysis. *Journal of Educational Psychology, 68*, 573-578.
Brown, G. T. L. (2004). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports, 94*, 1015-1024.
Brown, S., & McIntyre, D. (1993). *Making sense of teaching*. Buckingham: Open University Press.
Buck, S., & Tiene, D. (1989). The impact of physical attractiveness, gender, and teaching philosophy on teacher evaluations. *Journal of Educational Research, 82*, 172-177.
Bullock, C. (2004). *Instructor & Course Evaluation System (ICES): Myths & Misperceptions*. Champaign, IL: Center for Teaching Excellence. University of Illinois at Urbana-Champaign.
Burdsal, C. A., & Bardo, J. W. (1986). Measuring student's perceptions of teaching: Dimensions of evaluation. *Educational & Psychological Measurement, 46*, 63-79.
Burroughs, R. (2001). Composing standards and composing teachers: The problem of National Board Certification. *Journal of Teacher Education, 52*, 223-232.
Burroughs, R., Schwartz, T. A., & Hendricks-Lee, M. (2000). Communities of practice and discourse communities: Negotiating boundaries in NBPTS certification. *Teachers College Record, 102*, 344-374.
Burton, D. (1975). *Student ratings - an information source for decision making.* Paper presented at the 15th Annual Forum of the Association for Institutional Research, St Louis, MO.

Bushweller, K. (1998). Other voices: listening to what fellow teachers, parents, and students have to say in teacher evaluation. *The American School Board Journal, 185*(9), 24-27.
Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational & Behavioral Statistics, 24*, 323-341.
Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st century. Report of the Task Force on Teaching as a profession.* Washington, DC: Carnegie Forum on Education and the Economy.
Carney, T. F. (1969). Content analysis for high school students. *Quarterly of the Manitoba History Teachers' Association, 1*, 3-17.
Carney, T. F. (1972). *Content analysis: A technique for systematic inference from communications*. London: B T Batsford Ltd.
Cartwright, D. P. (1953). Analysis of qualitative material. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 421-470). New York: Holt, Rinehart, and Winston.
Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research* (IDEA Paper No. No 20): Center for Faculty Evaluation and Development, Kansas State University.
Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84*, 563-572.
Cashin, W. E., & Downey, R. G. (1995). Disciplinary differences in what is taught and in students' perceptions of what they learn and of how they are taught. *New Directions for Teaching and Learning, 64*, 81-92.
Cattell, R. B. (1973). *Personality and mood by questionnaire: A handbook of interpretive theory, psychometrics and practical procedures*. San Francisco: Jossey-Bass Publishers.
Centra, J. A. (1973a). The effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology, 65*, 395-401.
Centra, J. A. (1973b). Self-ratings of college teachers: A comparison with student ratings. *Journal of Educational Measurement, 10*, 287-296.
Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal, 14*, 17-24.
Centra, J. A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education, 18*, 379-389.
Centra, J. A., & Linn, R. L. (1973). *Student points of view in ratings of college instruction* (No. Report Number: RB-73-60). Lawrenceville, NJ: ETS, Princeton University.
Centre for Educational Research and Innovation. (1994). *Quality in teaching*. Paris: Centre for Educational Research and Innovation, Organisation for Economic Co-operation and Development (OECD).
Centre for Professional Development. (2004, 14 January). Student surveys - myths and research. Retrieved 1 May, 2004, from www2.auckland.ac.nz/cpd/evaluations/mythsresearch.html
Chang, T.-S. (2000a, April). *An application of regression models with student ratings in determining course effectiveness.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
Chang, T.-S. (2000b, April). *Student ratings: What are teacher college students telling us about them?* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
Chang, T.-S. (2002, August). *Results of student ratings: Does faculty attitude matter?* Paper presented at the Annual Meeting of the American Psychological Association, Chicago, IL.

Chapman, D. W., & Kelly, E. F. (1981). A comparison of the dimensions used by Iranian and American students in rating instruction. *International Review of Education, 27*, 41-60.
Chow, P., & Winzer, M. A. (1992). Reliability and validity of a scale measuring attitudes toward mainstreaming. *Educational & Psychological Measurement, 52*, 223-228.
Clark, D. L. (1987). High school seniors react to their teachers and their schools. *Phi Delta Kappan*(March), 503-509.
Clehouse, R. E. (2000). *A self-report by National Board-certified teachers of their perceptions of the impact of the National Board certification process upon them and their students.* Unpublished doctoral thesis, Northern Illinois University, IL.
Cliff, N. (1959). Adverbs as multipliers. *Psychological Review, 66*, 27-44.
Cochran, H. K. (1997, November). *The development and psychometric analysis of the Scale of Teachers' Attitudes Toward Inclusion (STATIC).* Paper presented at the 26th Annual Meeting of the Mid-South Educational Research Association, Memphis, TN.
Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Revised ed.). New York: Academic Press.
Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.
Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education. Fifth edition* (Fifth ed.). London: RoutledgeFalmer.
Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309.
Cohen, P. A. (1982). Validity of student ratings in psychology courses: A research synthesis. *Teaching of Psychology, 9*, 78-82.
Cohen, P. A. (1986, April). *An updated and expanded meta-analysis of multisection student rating validity studies.* Paper presented at the 70th Annual Meeting of the American Educational Research Association, San Francisco.
Cohen, P. A. (1987, April). *A critical analysis and reanalysis of the multisection validity meta-analysis.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.
Cohen, P. A., & Herr, G. (1979). A procedure for diagnostic instructional feedback: The Formative Assessment of College Teaching (FACT) model. *Educational Technology, 19*, 18-23.
Cohen, P. A., & Herr, G. (1982). Using an interactive feedback procedure to improve college teaching. *Teaching of Psychology, 9*, 138-140.
Cohen, S. R. (1997). The mismeasure of the college professor. *College Student Journal, 31*, 293-300.
Cole, L. (1940). *The background for college teaching*. New York: Farrar and Rinehart.
Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity* (No. OE-38001). Washington, DC: US Department of Health, Education & Welfare. Office of Education.
Cook, L. L., & Eignor, D. R. (1991). An NCME Instructional Module: IRT equating methods. *Educational Measurement: Issues & Practice., 10*(3), 37-45.
Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin Company.
Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist--Revised. *Psychological Assessment, 9*, 3-14.
Corey, D. D. (2002). *The Greek sophists: Teachers of virtue.* Unpublished doctoral thesis, Louisiana State University. LA
Costa, A. L. (1988). Foreword. In S. J. Stanley & W. J. Popham (Eds.), *Teacher Evaluation: Six Prescriptions for Success*. Alexandria, VA: Association for Supervision and Curriculum.

Covert, J. L. (2000). *A narrative analysis of national board- and non-national board-certified teachers' belief systems.* Unpublished doctoral thesis, The Ohio State University, Columbus, OH.
Cravens, T. F. (1996, March). *Students' perceptions of the characteristics of teaching excellence.* Paper presented at the Annual Meeting of the National Social Science Conference, Reno, NV.
Crocker, L. (1997). Assessing the content representativeness of performance assessment exercises. *Applied Measurement in Education, 10*, 83-95.
Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy*. Urbana, IL: University of Illinois Press.
Croshaw, B. J. (1999). *A study of National Board teacher's perceptions of the impact of effective administrative behaviors on successful completion of National Board certification.* Unpublished EdD, Idaho State University, Pocatello, ID.
Crozier, G. (1999). Is it a case of 'We know when we're not wanted'? The aprents' persepctive on parent-teacher roles and relationships. *Educational Research, 41*, 315-328.
Cruickshank, D. R. (2000). What makes teachers good? *Mid-Western Educational Researcher, 13*, 2-6.
Crumbley, D. L. (1995). The dysfunctional atmosphere of higher education: Games professors play. *Accounting Perspectives, 1*, 27-33.
Dancey, C. P., & Reidy, J. (2002). *Statistics without maths for psychology: Using SPSS for Windows.* (Second ed.). Harlow, England: Prentice Hall.
Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Princeton, NJ: Educational Testing Service.
d'Apollonia, S., & Abrami, P. C. (1996). *Variables moderating the validity of student ratings of instruction: A meta-analysis.* Paper presented at the 77th Annual Meeting of the American Educational Research Association, New York, NY.
Darling-Hammond, L., & Wise, A. E. (1992). Teacher professionalism. In M. C. Alkin (Ed.), *Encyclopaedia of educational research* (6th ed., Vol. 4, pp. 1359-1366). New York: Macmillan.
Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*, 285-328.
De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education, 5*, 17-34.
Department of Education and Science Welsh Office. (1985). *Better schools*. London: HMSO.
Dommeyer, C. J., Baum, P., Chapman, K. S., & Hanna, R. W. (2002). Attitudes of business faculty towards two methods of collecting teaching evaluations: Paper vs online. *Assessment & Evaluation in Higher Education, 27*, 455-463.
Dommeyer, C. J., Baum, P., & Hanna, R. W. (2002). College students' attitudes toward methods of collecting teaching evaluations: In-class versus on-line. *Journal of Education for Business, 78*, 11-15.
Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education, 53*, 51-62.
Dowling, W. C. (2000, 4 December). All lose when education is sold. *The Daily Targum*.
Doyle, K. O., & Crichton, L. I. (1978). Student, peer and self evaluations of college instructors. *Journal of Educational Psychology, 70*, 815-826.
Doyle, K. O., & Whitely, S. E. (1974). Student ratings as criteria for effective teaching. *American Educational Research Journal, 11*, 259-274.

Drews, D. R., Burroughs, W. J., & Nokovich, D. (1987). Teacher self-ratings as a validity criterion for student evaluations. *Teaching of Psychology, 14*, 23-25.
Dyer, K. M. (2001). The power of 360-degree feedback. *Educational Leadership* (February), 35-38.
Edwards, W. (1977). How to use multiattribute utility measurements for social decision-making. *IEEE transactions on systems, man, and cybernetics., SMC-7*, 326-340.
Edwards, W., & Newman, J. R. (1982). *Multiattribute evaluation*. Beverly Hills, CA: Sage Publications.
Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*, 483-501.
Eklund-Myrskog, G. (1998). Students' Conceptions of Learning in Different Educational Contexts. *Higher Education, 35*, 299-316.
Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education, 11*, 37-46.
Escalante, J., & Dirmann, J. (1990). The Jaime Escalante Math Program. *Journal of Negro Education, 59*, 407-423.
Etzioni, A. (Ed.). (1969). *The semi-professions and their organization. Teachers, nurses, social workers.* New York: The Free Press.
Ezzy, D. (2002). *Qualitative research: Practice and innovation*. London: Routledge.
Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.
Felder, R. M. (1992). What do they know, anyway? *Chemical Engineering Education, 26*, 134-135.
Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*, 199-242.
Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10*, 149-172.
Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education, 24*, 139-213.
Feldman, K. A. (1992). College students' views of male and female college teachers: Part I--Evidence from the social laboratory and experiments. *Research in Higher Education, 33*, 317-375.
Feldman, K. A. (1993). College students' views of male and female college teachers: Part II--Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*, 151-211.
Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368-395). New York: Agathon Press.
Fennema, E., & Sherman, J. A. (1976). Fennema-Sherman Mathematics Attitudes Scales: Instruments designed to measure attitudes toward the learning of mathematics by females and males. *Journal for Research in Mathematics Education, 7*, 324-326.
Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston IL: Row, Paterson and Co.
Finn, C. E., Kanstoroom, M., & Petrilli, M. J. (1999). *The quest for better teachers: data wish list*. Washington, DC: Thomas B Fordham Foundation.
Finn, C. E., & Wilcox, D. D. (1999, 9 August). Board games: Failure of National Board for Professional Teaching Standards to accomplish objective of improving quality of teaching in the US. Business backs a losing education strategy. *National Review*.

Finn, C. E., & Wilcox, D. D. (2000, 13 January). Teachers should be graded on how well their students are learning. *The Los Angeles Times*.
Fitz-Gibbon, C. T. (1985). A-level results in comprehensive schools: The COMBSE project, Year 1. *Oxford Review of Education, 11*, 43-58.
Fitz-Gibbon, C. T. (1996). *Monitoring education: Indicators, quality and effectiveness*. London: Cassell.
Fitz-Gibbon, C. T. (1997). *From value added indicators to evidence based education; The task for the next decade* (Seminar Series No. 65). Jolimont, Victoria: Incorporated Association of Registered Teachers of Victoria.
Fitz-Gibbon, C. T., & Tymms, P. B. (2002). Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Education Policy Analysis Archives, 10*(6).
Flanders, N. A. (1973). Basic teaching skills derived from a model of speaking and listening. *Journal of Teacher Education, 24*, 24-37.
Flanders, N. A. (1974). The changing base of performance-based teaching. *Phi Delta Kappan, 55*, 312-315.
Fletcher, R. B. (1998). *Test assembly for polytomous items: An application using the Physical Self Description Questionnaire.* Unpublished doctoral thesis, The University of North Carolina at Greensboro, Greensboro, NC.
Fordham Foundation. (1999). The teachers we need and how to get more of them. Retrieved 11 June, 2001, from http://www.edexcellence.net/library/teacher.html
Fox, R., Peck, R. F., Blattstein, A., & Blattstein, D. (1983). Student evaluation of teacher as a measure of teacher behavior and teacher impact on students. *Journal of Educational Research, 77*, 16-21.
Frankhouser, W. M. (1984). The effects of different oral directions as to disposition of results on student ratings of college instruction. *Research in Higher Education, 20*, 367-374.
Franklin, J., & Theall, M. (1989, March). *Who Reads Ratings: Knowledge, Attitude, and Practice of Users of Student Ratings of Instruction.* Paper presented at the 70th Annual Meeting of the American Educational Research Association, San Francisco.
Franklin, J., & Theall, M. (1990). Communicating student ratings to decision makers: Design for good practice. *New Directions for Teaching & Learning*, 75-93.
Franklin, J., Theall, M., & Ludlow, L. H. (1991, April). *Grade inflation and student ratings: A closer look.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
Fraser, B. J. (1986). *Classroom environment*. London: Croom Helm.
Frazier, J. W. (1999). *A description of the process of National Board Certification from the perspectives of a group of North Carolina candidates: A case study.* Unpublished doctoral thesis, University of South Carolina, ,SC.
Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. *Science, 182*, 83-85.
Frey, P. W. (1976). Validity of student instructional ratings: Does timing matter? *Journal of Higher Education, 47*, 327-336.
Frey, P. W., Leonard, D. W., & Beatty, W. M. (1975). Student ratings of instruction. *American Educational Research Journal, 12*, 435-447.
Friedlander, J. (1978). Student perceptions on the effectiveness of midterm feedback to modify college instruction. *Journal of Educational Research, 71*, 140-143.
Gage, N. L. (1961). The appraisal of college teaching. *Journal of Higher Education, 32*, 17-22.

Gillmore, G. M., & Brandenburg, D. C. (1974). *Would the proportion of students taking a class as a requirement affect student rating of the course?* (No. RR-347). Urbana Office of Instructional Resources: Illinois University.
Glas, C. A. W., & Beguin, A. A. (1996). *Appropriateness of IRT observed score equating. Research Report 96-04* (Research Report No. Research Report 96-04). Enschede, Netherlands: Twente University, Faculty of Educational Science and Technology.
Goddard, P. R. (2003). *Personal communication*. Auckland, NZ.
Goldberg, G., & Callahan, J. (1991). Objectivity of student evaluations of instructors. *Journal of Education for Business, 66*, 377-378.
Goldhaber, D. D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Seattle, WA: Center on Reinventing Public Education, Daniel J. Evans School of Public Affairs, University of Washington.
Goldhaber, D. D., Perry, D., & Anthony, E. (2003). *NBPTS certification: Who applies and what factors are associated with success?*
Goode, W. J. (1969). The theoretical limits of professionalization. In A. Etzioni (Ed.), *The semi-professions and their organization* (pp. 266-313). New York: The Free Press.
Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217.
Greenwood, G. E., Bridges, C. M., Ware, W. B., & McLean, J. E. (1973). Student Evaluation of College Teaching Behaviors Instrument: A factor analysis. *Journal of Higher Education, 44*, 596-604.
Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality - analysis of relevant factors based on empirical evaluation research. *Assessment & Evaluation in Higher Education, 28*, 229-238.
Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26*, 534-552.
Gumpel, T., Wilson, M., & Shalev, R. (1998). An item response theory analysis of the Conners Teacher's Rating Scale. *Journal of Learning Disabilities, 31*, 525-532.
Guthrie, E. R. (1954). *The evaluation of teaching. A progress report*. Seattle, WA: The University of Washington.
Haak, R. A., Kleiber, D., & Peck, R. F. (1972). *Student Evaluation of Teacher Instrument, II. Manual*. Austin, TX: Research and Development Center for Teacher Education, University of Texas.
Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), *Educational measurement* (3rd ed.): Macmillan.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications Inc.
Hamermesh, D. S., & Parker, A. M. (2003). *Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity* (No. NBER Working Paper No. w9853). Cambridge, MA: National Bureau of Economic Research.
Hargreaves, A. (1997). The four ages of professionalism and professional learning. *Unicorn, 23*, 86-114.
Hargreaves, A. (2000). Four ages of professionalism and professional learning. *Teachers and Teaching: History and Practice, 6*, 151-182.
Hargreaves, A., & Goodson, I. F. (1996). Teachers' professional lives: Aspirations and actualities. In I. F. Goodson & A. Hargreaves (Eds.), *Teachers' professional lives.* (pp. 1-27). London: Falmer Press.
Harris, R. J. (1975). *A primer of multivariate statistics*. New York: Academic Press.

Harris, S. M., & Halpin, G. (2002). Development and validation of the Factors Influencing Pursuit of Higher Education Questionnaire. *Educational and Psychological Measurement, 62*, 79-96.
Haskell, R. E. (1997). Academic freedom, tenure, and the student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives, 5*(6).
Hativa, N. (1996). University instructors' ratings profiles: Stability over time, and disciplinary differences. *Research in Higher Education, 37*, 341-365.
Hattie, J. A. (1996). *Validating the specification of a complex content domain.* Paper presented at the Annual Conference of the American Educational Research Association, New York, NY.
Hattie, J. A. (1999). Influences on student learning. *The International Principal, 5*(3).
Hattie, J. A., & Clinton, J. (in press). Identifying accomplished teachers: a validation study. In L. C. Ingvarson (Ed.), *Assessing teachers for professional certification: The National Board for Professional Teaching Standards*. Greenwich, Connecticut: JAI Press Inc.
Haynes, D. D. (1995). One teachers' experience with National Board assessment. *Educational Leadership, 52*(6), 58-60.
Helsby, G. (1995). Teachers' construction of professionalism in England in the 1990s. *Journal of Education for Teaching, 21*, 317-332.
Hildebrand, M. (1972). How to recommend promotion of a mediocre teacher without actually lying. *Journal of Higher Education, 43*, 44-62.
Holly, K. A. (1971). *Structure-of-intellect factor abilities and a self-concept measure in mathematics relative to performance in high school modern algebra.* Unpublished doctoral thesis, University of Southern California.
Holmes, T. A. M. (1996). *Student evaluations of teachers, student ratings of teacher behaviors, and their relationship to student achievement: A cross-ethnic comparison.* Unpublished doctoral thesis, Andrews University, Berrien Springs, MI.
Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley Publishing.
Holsti, O. R., Loomba, J. K., & North, R. C. (1968). Content analysis. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. Two, pp. 596-692). Reading, MA: Addison-Wesley Publishing Company.
Howard, G., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology, 77*, 187-196.
Hoyle, E. (1974). Professionality, professionalism and control in teaching. *London Educational Review, 3*, 13-19.
Hoyle, E., & John, P. D. (1995). *Professional knowledge and professional practice.* London: Cassell.
Hunter, M. (1985). *Prescription for improved instruction.* El Segundo, CA.: TIP Publications.
Hunter, M. (1993). *Enhancement of teaching through coaching, supervision, and evaluation.* Kalamazoo, MI: Center for Research on Educational Accountability and Teacher Evaluation (CREATE).
Iovacchini, L. C. (1998). *National Board for Professional Teaching Standards: What teachers are learning.* Unpublished doctoral thesis, University of South Carolina.
Irving, S. E. (1996). *Using student evaluations for teacher development in nine urban secondary schools.* Unpublished masters thesis, Massey University, Palmerston North, New Zealand.
Jacobs, L. C. (1987). *University faculty and students' opinions of student ratings.* (Indiana Studies in Higher Education No. 55). Bloomington, IN: Indiana University, Bureau of Evaluative Studies and Testing.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*, 461-475.
Jaeger, R. M. (1995). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues & Practice, 14*(4), 16-20.
Jaeger, R. M. (1998). Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' assessments: A methodological accounting. *Journal of Personnel Evaluation in Education, 12*, 189-210.
Jensen, J. W. (1998, April). *Teacher candidates' conceptions of teaching and learning: A review.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
Johannessen, T. A., Gronhaug, K., Risholm, N. G., & Mikalsen, O. (1997). What is important to students? Exploring dimensions in their evaluations of teachers. *Scandinavian Journal of Educational Research, 41*, 165-177.
Johnson, B. L. (1997). An organizational analysis of multiple perspectives of effective teaching: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 69-87.
Johnson, G. D., Stanfield, C., Haywick, D., Flynn, R., Pardue, H., Woodruff, J., et al. (2003). *Final report of the Student Evaluation of Teaching Committee*. Mobile. AL: University of South Alabama.
Johnson, H. C. (1980). Theoretical development of advance organizers. *International Journal of Mathematical Education in Science and Technology, 11*, 511-515.
Jones, B. D., & Egley, R. J. (2004). Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives, 12*(39), Retrieved 16 October 2004 from http://epaa.asu.edu/epaa/v2012n2039/..
Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31-36.
Kaplan, A. (1943). Content analysis and the theory of signs. *Philosophy of Science, 10*, 230-247.
Karger, T. (1987, 28 August). Focus groups are for focusing, and for little else. *Marketing News,* 52-55.
Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education, 40*, 69-97.
Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology, 80*, 342-344.
Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355.
King, M. B. (1994). Locking ourselves in: National standards for the teaching profession. *Teaching and Teacher Education, 10*, 95-108.
Kitzinger, J. (1994). The methodology of focus groups: The importance of interaction between research participants. *Sociology of Health & Illness, 16*, 130-121.
Klann, W. E., & Hoff, E. (1976). The use of judgment analysis in analyzing student evaluation of teachers. *MATYC Journal, 10*, 137-139.
Kolitch, E., & Dean, A. V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about "good" teaching. *Studies in Higher Education, 24*, 27-42.
Kratz, H. E. (1896). Characteristics of the best teachers as recognized by children. *Pedagogical Summary, 3*, 413-418.

Krehbiel, T. C., & McClure, R. H. (1997). Using student disconfirmation as a measure of classroom effectiveness. *Journal of Education for Business, 72*, 224-229.
Krueger, R. A. (1988). *Focus groups: A practical guide for applied research*. Newbury Park, CA: Sage Publications.
Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 109*(Spring), 9-25.
Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. *Review of Research in Education, 3*, 210-240.
Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement., 19*, 317-322.
Leeds, M., Stull, W., & Westbrook, J. (1998). Do changes in classroom techniques matter? Teaching strategies and their effects on teaching evaluations. *Journal of Education for Business, 74*, 75-78.
Leef, G. C. (2003). *National Board certification: Is North Carolina getting its money's worth?* (Policy report). Raleigh , NC: North Carolina Education Alliance.
Leventhal, L., Abrami, P. C., & Perry, R. P. (1976). Do teacher rating forms reveal as much about students as about teachers? *Journal of Educational Psychology, 68*, 441-445.
Lewis, K. G. (2001a). Using midsemester student feedback and responding to it. *New Directions for Teaching and Learning, 87*(Fall), 33-44.
Lewis, K. G. (Ed.). (2001b). *Techniques and strategies for interpreting student evaluations.* San Francisco: Jossey-Bass.
Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching.* (Eighth ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.
Linsky, A. S., & Straus, M. A. (1975). Student evaluations, research productivity, and eminence of college faculty. *Journal of Higher Education, 46*, 89-102.
Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.
Locke, T. (2001). Curriculum, assessment and the erosion of professionalism. *New Zealand Journal of Educational Studies, 36*, 5-23.
Lortie, D. C. (1969). The balance of control and autonomy in elementary school teaching. In A. Etzioni (Ed.), *The semi-professions and their organization* (pp. 1-53). New York: The Free Press.
Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education, 10*, 203-226.
Ludlow, L. H. (2001). Teacher Test Accountability: From Alabama to Massachusetts. *Education Policy Analysis Archives, 9*(6).
Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and non-reached items: Practical data analysis implications. *Educational & Psychological Measurement, 59*, 615-630.
Luiten, J. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *American Educational Research Journal, 17*, 211-218.
MacBeath, J. (1999). *Schools must speak for themselves: The case for school self-evaluation*. London: Routledge.
Manatt, R. P. (2000). Feedback at 360 degrees. *School Administrator, 57*(9), 10.
Marlin, J. W. (1987). Student perception of end-of-course evaluations. *Journal of Higher Education, 58*, 704-716.
Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal, 17*, 219-237.

Marsh, H. W. (1981). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. *Australian Journal of Education, 25*, 177-193.
Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*, 77-95.
Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74*, 264-279.
Marsh, H. W. (1983). Multitrait-multimethod analysis: Distinguishing between items and traits. *Educational & Psychological Measurement, 43*, 351-358.
Marsh, H. W. (1984). Student evaluations of university teaching: Dimensionality, reliability, validity, potential biases, utility. *Journal of Educational Psychology, 76*, 707-754.
Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388.
Marsh, H. W. (1994). Student evaluation of teaching. In T. Husen & T. N. Postlewaite (Eds.), *International encyclopaedia of education: Research and studies.* (2nd ed., Vol. 10, pp. 6227-6235). Oxford: Pergamon.
Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness. *Journal of Higher Education, 64*, 1-18.
Marsh, H. W., Fleiner, H., & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology, 67*, 833-839.
Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21*, 341-366.
Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*, 9-18.
Marsh, H. W., & Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology, 72*, 468-475.
Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71*, 149-160.
Marsh, H. W., & Roche, L. A. (1992). The use of student evaluations of university teaching in different settings: The applicability paradigm. *Australian Journal of Education, 36*, 278-300.
Marsh, H. W., & Roche, L. A. (1999). Reply upon SET research. *American Psychologist, 54*, 517-520.
Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92*, 202-228.
Marsh, H. W., & Ware, J. E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology, 74*, 126-134.
Massey University. (1993). New system to SET direction for evaluating teaching performance. *Massey Alumnus,* 1-2.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.
Masters, J. R. (1977). *Improving instruction through student observation of teaching methods*. Harrisburg Bureau of Information Systems: Pennsylvania State Dept. of Education.

Masters, J. R. (1979). High school student ratings of teachers and teaching methods. *Journal of Educational Research, 72*, 219-224.
Masters, J. R., & Weaver, W. G. (1977, April). *The development of a student observation of teachers instrument for use in high schools.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.
Mateo, M. A., & Fernandez, J. (1996). Incidence of class size on the evaluation of university teaching quality. *Educational & Psychological Measurement, 56*, 771-778.
McBean, E. A. (1991). Analyses of teaching and course questionnaires: A case study. *Engineering Education, 81*, 439-441.
McBean, E. A., & Lennox, W. C. (1987). Measurement of quality of teaching and courses by a single question versus a weighted set. *European Journal of Engineering Education, 12*, 329-335.
McCabe, N. (1995). Twelve high school 11th grade students examine their best teachers. *Peabody Journal of Education, 70*, 117-126.
McCall, W. A., & Krause, G. R. (1959). Measurement of teacher merit. *Journal of Educational Research, 53*, 73-75.
McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education, 21*, 150-158.
McConnell, J. W. (1978, March). *Relations between teacher attitudes and teacher behavior in ninth-grade algebra classes.* Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Ontario.
McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe: Bulletin of the AAUP, 65*, 384-397.
McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.
McKeachie, W. J., Lin, Y.-G., Daugherty, M., Moffett, M. M., Neigler, C., et al. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology, 50*, 168-174.
McKeachie, W. J., Lin, Y.-G., & Mann, W. (1971). Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal, 8*, 435-445.
McLennan, R. (1992). *The OD focus group: A versatile tool for change* (Vol. 5/92). Wellington, NZ: Graduate School of Business and Government Management, Victoria University of Wellington.
McMillan, J. J., & Cheney, G. (1996). The student as consumer: The implications and limitations of a metaphor. *Communication Education, 45*, 1-15.
Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-based teacher evaluation of teacher performance: An empirical approach*. New York: Longman Inc.
Menges, R. J. (1990, April). *A profile of recent research on feedback. Revised.* Paper presented at the Paper presented at the Annual Conference of the American Educational Research Association, Boston, MA.
Mertler, C. A. (1999). Teacher perceptions of students as stakeholders in teacher evaluation. *American Secondary Education, 27*(3), 17-30.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Old Tappan. NJ: MacMillan.
Miklich, D. R. (1969). An experimental validations study of the Purdue Rating Scale for Instruction. *Educational & Psychological Measurement, 29*, 963-967.
Millea, M., & Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal, 36*, 582-590.
Miller, M. T. (1971). Instructor attitudes toward, and their use of, student ratings of teachers. *Journal of Educational Psychology, 62*, 235-239.

Ministry of Education. (1995). *Draft national guidelines for performance management in schools*. Wellington, NZ: Ministry of Education.
Ministry of Education. (1997a). *Performance management in kura kaupapa Maori : consultation draft : appraisal of teacher performance.* Wellington, N.Z.: Ministry of Education,.
Ministry of Education. (1997b). *Performance management systems : a series of guidelines on performance management systems.* Wellington, N.Z.: Ministry of Education,.
Ministry of Education. (1998a). *Interim professional standards for primary, secondary and area school principals*. Wellington, NZ: Ministry of Education.
Ministry of Education. (1998b). *Principal performance management : a resource for boards of trustees & principals.* Wellington, N.Z.: School Labour Market Development Unit, Ministry of Education.
Ministry of Education. (1998c). *Teacher performance management : primary school teachers, primary school deputy/assistant principals : a resource for boards of trustees, principals & teachers.* Wellington, N.Z.: School Labour Market Development Unit, Ministry of Education.
Ministry of Education. (1999a). *Professional standards: Criteria for quality teaching. Area school teachers and unit holder*. Wellington, NZ: Ministry of Education.
Ministry of Education. (1999b). Professional standards: Criteria for quality teaching. Secondary school teachers and unit holders. Wellington, New Zealand: Author.
Ministry of Education. (1999c). Teacher performance management: A resource for boards of trustees, principals and teachers. Secondary school teachers; area school teachers; unit holders. Wellington, New Zealand: Author.
Ministry of Education. (2000a). Attestation for teacher salary progression. Circular 2000/02. Wellington, New Zealand: Author.
Ministry of Education. (2000b). *Education statistics news sheet: March 2000 school statistics. Volume 10 No. 2*. Wellington, New Zealand: Ministry of Education.
Mislevy, R. J., & Bock, R. D. (1990). BILOG 3. Mooresville, IN: Scientific Software.
Moore, A., Masterson, J. T., Christophel, D. M., & Shea, K. A. (1996). College teacher immediacy and student ratings of instruction. *Communication Education, 45*, 29-39.
Moore, D. P. (1999). *The National Board for Professional Teaching Standards (NBPTS) assessment: Learning style and other factors that lead to success.* Unpublished doctoral thesis, University of Cincinnati, Cincinnati.
Morgan, D. L. (1988). *Focus groups as qualitative research* (Vol. 16). Newbury Park, CA: Sage Publications.
Moss, P. A., & Schutz, A. (1999). Risking frankness in educational assessment. *Phi Delta Kappan, 80*, 680-687.
Munz, D. C., & Munz, H. E. (1997). Student mood and teaching evaluations. *Journal of Social Behavior & Personality, 12*, 233-242.
Murphy, R. (1990). Proletarianization or bureaucratization: The fall of the professional? In R. Torstendahl & M. Burrage (Eds.), *The formation of professions: knowledge, state and strategy*. London: Sage.
Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*, 138-149.
Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82*, 250-261.
Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630-635.

Nasser, N., & Abouchedid, K. (2000, April). *Educational Research in the Levantine: Revisited.* Paper presented at the "Towards the Global University II: Redefining Excellence in the Third Millennium" International Conference & Exhibition, Cape Town, South Africa.

National Board for Professional Teaching Standards. (1996). *Adolescent and Young Adulthood/Mathematics: Standards for National Board certification*. Washington, DC: National Board for Professional Teaching Standards.

National Board for Professional Teaching Standards. (2003a, 31 July). General Information About National Board Certification. Retrieved 22 January, 2004, from http://www.nbpts.org/standards/nbcert.cfm

National Board for Professional Teaching Standards. (2003b, 30 May). General Information About the NBPTS Standards. Retrieved 25 January, 2004, from http://www.nbpts.org/standards/stds.cfm#stand_dev_a

National Board for Professional Teaching Standards. (2003c, 4 June). Guide to National Board certification. Retrieved 23 January, 2004, from http://www.nbpts.org/candidates/index.cfm

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform: A report to the Nation and Secretary of Education, United States Department of Education*. Washington, DC: National Commission on Excellence in Education.

National Governors' Association. (1986). *Time for results: The Governors' 1991 report on education*. Washington, DC: Author.

Neustel, S. B. (2001). *A psychometric investigation of NBPTS assessments: A comparative analysis of information functions.* Unpublished doctoral thesis, The University of North Carolina at Greensboro, Greensboro.

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill Book Company.

Ogden, D. H. (1994, November). *Characteristics of good/effective teachers: Gender differences in student descriptors.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Nashville, TN.

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology, 26*, 382-399.

Ontario College of Teachers. (1999). *Standards of practice for the teaching profession*. Ontario, ON: Ontario College of Teachers,.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter Publishers.

Ory, J. C., Braskamp, L. A., & Pieper, D. M. (1980). Congruency of student evaluation information collected by three methods. *Journal of Educational Psychology, 72*, 181-185.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework. *New Directions for Institutional Research, 109*(Spring), 27-44.

Page, C. F. (1974). *Student evaluation of teaching: The American experience.* London: Society for Research in Higher Education.

Paisley, W. J. (1969). Studying 'style' as a deviation from encoding norms. In G. Gerbner, O. R. Holsti, K. Krippendorff, W. J. Paisley & P. J. Stone (Eds.), *The analysis of communication content* (pp. 133-146): Wiley.

Papalewis, R. (1990, April). *Interpretation of student data: Contextual and sociocultural variables.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Pearlman, M. A. (2002, January). *The architecture of the NBPTS certification assessments.* Paper presented at the NBPTS Invitational Research Conference, Chicago, IL.

Pearson, P. D., & Garavaglia, D. R. (1997). *Improving the information value of performance items in large scale assessments*. Palo Atlo, CA: NAEP Validity Studies, American Institute for Research.

Pehkonen, E. (1992). *Problem fields in mathematics teaching. Part 3: Views of Finnish seventh-grades about mathematics teaching*. Helsinki, Finland: Dept of Teacher Education, Helsinki University.

Penny, A. R. (2003a). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*, 399-411.

Penny, A. R. (2003b, July). *Practices and strategies for consultative support with student feedback: A meta-analytic review.* Paper presented at the Evidence-Based Policies and Indicator Systems. Fourth international, inter-disciplinary, biennial conference: What kind of evidence does government need?, London.

Perry, R. P. (1985). Instructor expressiveness: Implications for improved teaching. In J. G. Donald & A. M. Sullivan (Eds.), *Using research to improve teaching* (No 23, September ed.).

Perry, R. P., Abrami, P. C., & Leventhal, L. (1979). Educational seduction: The effect of instructor expressiveness and lecture content on student ratings and achievement. *Journal of Educational Psychology, 71*, 107-116.

Perry, R. P., & Smart, J. C. (1997). Introduction. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 1-10). New York: Agathon Press.

Peterson, D., Micceri, T., & Smith, B. O. (1985). Measurement of teacher performance: A study in instrument development. *Teaching & Teacher Education, 1*, 63-77.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed ed.). Thousand Oaks, CA: Corwin Press, Inc.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*, 135-153.

Petrosky, A. R. (1994). Schizophrenia, the National Board for Professional Teaching Standards' policies, and me. *English Journal, 83*(7), 33-42.

Phelan, P., Davidson, A. L., & Cao, H. T. (1992). Speaking up: Students' perspectives on school. *Phi Delta Kappan, 73*, 695-696,698-704.

Phillips, J. R., & Kanstoroom, M. (1999). Title II: Does professional development work? In M. Kanstoroom & C. E. Finn (Eds.), *New Directions: Federal education policy in the twenty-first century* (pp. 61-77). Washington DC: The Thomas B Fordham Foundation.

Plake, B. S., Hambleton, R. K., & Jaegar, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment and some field tests. *Educational and Psychological Measurement, 57*, 400-411.

Podgursky, M. (2001a). Defrocking the National Board: Will the imprimatur of "board certification" professionalize teaching? *Education Matters*(Summer), 79-82.

Podgursky, M. (2001b, 11 April). Should states subsidize National Certification? *Education Week*.

Pohlmann, J. T. (1975). A multivariate analysis of selected class characteristic and student ratings of instruction. *Multivariate Behavioral Research, 10*, 81-91.

Pratt, D. D., Kelly, M., & Wong, W. S. S. (1999). Chinese conceptions of "effective teaching" in Hong Kong: Towards culturally sensitive evaluation of teaching. *International Journal of Lifelong Education, 18*, 241-258.

Prave, R. S., & Baril, G. L. (1993). Instructor ratings: Controlling for bias from initial student interest. *Journal of Education for Business, 68*, 362-366.

Principia Products Inc. (1997). Remark Office OMR 4.0 (Version 4.0 for Windows). West Chester, PA: Principia Products Inc,.
Principia Products Inc. (2000). Remark Office OMR 5.0 (Version 5.0 for Windows). Paoli, PA: Principia Products Inc,.
Prybylo, D. (1998). Beyond a positivistic approach to teacher evaluation. *Journal of School Leadership, 8*, 558-583.
Purdie, N., Pillay, H., & Boulton-Lewis, G. M. (2000). Investigating Cross-Cultural Variation in Conceptions of Learning and the Use of Self-Regulated Strategies. *Education Journal, 28*, 65-84.
Rayder, N. F. (1968). College student ratings of instructors. *Journal of Experimental Education, 37*, 76-81.
Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision, 13*, 519-527.
Roberts, J. S., & Laughlin, J. E. (1996). *The Graded Unfolding Model: A unidimensional item response model for unfolding graded responses* (No. ETS-RR-96-16). Princeton, NJ.: Educational Testing Service.
Rodabaugh, R., & Kravitz, D. (1994). Effects of procedural fairness on student judgments of professors. *Journal on Excellence in College Teaching, 5*(2), 67-83.
Rodin, M., & Rodin, B. (1972). Student evaluation of teachers. *Science, 177*, 1164-1166.
Rosenholtz, S. J. (1984). Myth 7: Since almost everyone can recall at least one great teacher, the characteristics of great teachers are easy to identify. In G. C. Hall (Ed.), *Teacher competence* (pp. 25-26). Bloomington, IN: Phi Delta Kappa, Center on Evaluation, Development and Research.
Rosenshine, B. (2003). High-Stakes Testing: Another Analysis. *Education Policy Analysis Archives, 11*(24).
Rosenthal, J. S. (1995). Active learning strategies in advanced mathematics classes. *Studies in Higher Education, 20*, 223-228.
Rosenthal, R. (1973). The Pygmalion effect lives. *Pyschology Today, 7*(4), 56-60.
Rosenthal, R. (1997, August). *Interpersonal expectancy effects: A forty year perspective.* Paper presented at the Teachers of Psychology in the Secondary Schools section of the American Psychological Association Convention, Chicago, IL.
Rotberg, I. C., Futrell, M. H., & Lieberman, J. M. (1998). National Board certification: Increasing participation and assessing impacts. *Phi Delta Kappan, 79*, 462-466.
Rowe, K. (2003). *The importance of Teacher Quality as a key determinant of students' experiences and outcomes of schooling.* Paper presented at the ACER Research Conference. Building Teacher Quality: What does the research tell us?, Melbourne, Vic.
Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books.
Sailor, P., Worthen, B. R., & Shin, E.-H. (1997). Class level as a possible mediator of the relationship between grades and student ratings of teaching. *Assessment & Evaluation in Higher Education, 22*, 261-269.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 34*(4, Pt 2, Whole No 17).
Sanders, W. L. (1998). Value-Added Assessment. *School Administrator, 11*(55), 24-27.
Sanders, W. L. (2000). Value-added assessment from student achievement data: opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*, 329-339.
Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247-256.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Santeusanio, R. (1998). Improving performance with 360-degree feedback. *Educational Leadership*(February), 30-32.

Sawa, R. (1995). *To cultivate or to weed: An assessment of teacher evaluation policies and practices in rural Saskatchewan school divisions* (No. Report #95-04). Regina, Saskatchewan: SSTA Research Centre.

Scherr, F. C., & Scherr, S. S. (1990). Bias in student evaluations of teacher effectiveness. *Journal of Education for Business, 65*, 356-358.

Schmuck, R. A., & Schmuck, P. A. (1989). *Adolescents' attitudes toward school and teachers: From 1963 to 1989* (No. ERIC Document Reproduction Service ED316381).

Schon, D. (1987). *Educating the reflective practitioner: toward a new design for teaching and learning in the professions*. San Franscisco, CA: Jossey-Bass.

Schuckman, H. (1990). Students' perceptions of faculty and graduate students as classroom teachers. *Teaching of Psychology, 17*, 162-165.

Schunk, D. H. (1996, April). *Attributions and the development of self-regulatory competence.* Paper presented at the Annual Conference of the American Educational Research Association., New York, NY.

Sclan, E. M. (1994). *Performance evaluation for experienced teachers: An overview of state policies.* (Trends and Issues Paper, No. 10). Washington, DC: ERIC Clearinghouse on Teaching and Teacher Education.

Scott, C. S. (1975). Some remarks on "Student ratings: Validation". *American Educational Research Journal, 12*, 444-447.

Scriven, M. (1988a). The design and use of forms for the student evaluation of teaching. Perth, WA: CTES, University of Western Australia.

Scriven, M. (1988b). Evaluating teachers as professionals: The duties-based approach. In S. J. Stanley & W. J. Popham (Eds.), *Teacher evaluation: Six prescriptions for success*. Alexandria, VA: Association for Supervision and Curriculum Development.

Scriven, M. (1993). *The validity of student ratings in teacher evaluation*: Evaluation & Development Group.

Scriven, M. (1994). Using students ratings in teacher evaluation. *Evaluation Perspectives, 4*, 1-4.

Scriven, M. (1995). *Student ratings offer useful input to teacher evaluations* (No. ERIC Document Reproduction Service ED 398 240). Washington, DC.: ERIC Clearinghouse on Assessment and Evaluation.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 39*(46), A40.

Shinkfield, A. J., & Stufflebeam, D. L. (1995). *Teacher evaluation: Guide to effective practice*. Boston: Kluwer Academic Publishers.

Shulman, L. S., & Sykes, G. (1986). *A national board for teaching? In search of a bold standard.* New York: Carnegie Corporation of New York: Task Force on Teaching as a Profession, Carnegie Forum on Education and the Economy.

Sizemore, R. W. (1979). *A comparison of the perceptions of the characteristics of teachers by black and white secondary school students in an urban school district*.

Sizemore, R. W. (1981). Do black and white students look for the same characteristics in teachers? *Journal of Negro Education, 50*, 48-53.
Smith, L. R. (1984). Effect of teacher vagueness and use of lecture notes on student performance. *Journal of Educational Research, 78*, 69-74.
Solmon, L. C., & Podgursky, M. (1999). The pros and cons of performance-based compensation. Retrieved 11 June, 2001
Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review, 21*, 287-294.
Stahl, J., Shumway, R., Bergstrom, B., & Fisher, A. (1997). On-line performance assessment using rating scales. *Journal of Outcome Measurement, 1*, 173-191.
Stewart, D. W., & Shamdasani, P. N. (1991). *Focus groups: Theory and practice* (Vol. 20). Newbury Park, CA: Sage Publications.
Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational & Behavioral Statistics, 21*, 365-389.
Stone, J. E. (2002). The value-added achievement gains of NBPTS-certified teachers in Tennessee: A brief report. Retrieved 23 January, 2004, from http://www.education-consumers.com/briefs/stoneNBPTS.shtm
Stroh, L. (1991). High school student evaluation of student teachers: How do they compare with professionals? *Illinois School Research & Development, 27*, 81-92.
Stronge, J. H. (2002). *Qualities of effective teachers*. Alexandria, VA: Association for Supervision and Curriculum Development.
Sullivan, A. M., & Skanes, G. R. (1974). Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology, 66*, 584-590.
Svinicki, M. (1998). POD: Chronicle and so on. Retrieved 16 October 2004 from http://www.umanitoba.ca/uts/sigfted/archive1.html.
Sykes, R. C., & Ito, K. (1997). The effects of computer administration on scores and item parameter estimates of an IRT-based licensure examination. *Applied Psychological Measurement, 21*, 51-63.
Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item format tests with the one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 37*, 221-244.
Tagomori, H. T., & Bishop, L. A. (1995). Student evaluation of teaching: Flaws in the instruments. *Thought & Action, 11*, 63-78.
Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research & Development in Education, 28*, 169-173.
Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.
Taylor, F. W. (1911). *The principles of scientific management*. New York: Harper Bros.
Taylor, G. A. (2000). *Teacher change and the National Board for Professional Teaching Standards: A case study of eleven Colorado teachers.* Unpublished doctoral thesis, University of Colorado at Boulder, Boulder.
Taylor, P. G. (1996). Reflections on Students' Conceptions of Learning and Perceptions of Learning Environments. *Higher Education Research and Development, 15*, 223-237.
Teacher Registration Board. (1997). *Handbook: The registration of teachers in Aotearoa New Zealand*. Wellington, New Zealand: Teacher Registration Board.
The Secretary for Education, & The New Zealand Post Primary Teachers' Association. (2002). Secondary Teachers' Collective Agreement 2002-2004. Wellington, NZ.
Theall, M., Abrami, P. C., & Mets, L. A. (Eds.). (2001). *The student ratings debate: Are they valid? How can we best use them?* San Francisco: Jossey-Bass.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, 27*(5), 45-56.
Thissen, D. (1991a). MULTILOG user's guide: Multiple, categorical item analysis and test scoring using Item Response Theory (Version 6.0). Chicago: Scientific Software.
Thissen, D. (1991b). Plotlog (Version 6.0). Chicago: Scientific Software.
Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 837-847.
Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in The Journal of Experimental Education. *The Journal of Experimental Education, 66*, 75-83.
Tiberius, R. G., Sackin, H. D., & Cappe, L. (1987). A comparison of two teaching methods for evaluating teaching. *Studies in Higher Education, 12*, 287-297.
Tod, L. (2000). The effectiveness of secondary school student evaluation of teacher performance. *New Zealand Journal of Educational Administration, 15*, 23-31.
Todd, E. S., & Higgins, S. (1998). Powerlessness in professional and parent partnerships. *British Journal of Sociology of Education, 19*, 227-236.
Toney, J. W. (1973). *The effects of feedback to teachers from student evaluations of the instructional process.*
Traub, R. E., Haertel, E. H., & Shavelson, R. J. (1996). *The effects of measurement error on the trustworthiness of examinee classifications.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.
Triandis, H. C. (1972). *The analysis of subjective culture*. New York: Wiley-Interscience.
Tucker, P. D. (1997). Lake Wobegon: Where all the teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*, 103-126.
Tuckman, B. W., & Oliver, W. F. (1968). Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology, 59*, 297-301.
Upsall, D. (2000). Teacher accountability: Reflective professional or competent practitioner? *New Zealand Annual Review of Education: Te Arotake A Tau O Te Ao O Te Matauranga I Aotearoa, 10*, 167-185.
Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61*, 219-224.
Vandervoort, L. G., Amrein-Beardsley, A. L., & Berliner, D. C. (2004). National Board Certified Teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46), 1-117.
Veldman, D. J. (1970). *Student Evaluation of Teaching*. Austin, TX: Research and Development Center for Teacher Education, Texas University.
Veldman, D. J., & Peck, R. F. (1967). *The Pupil Observation Survey; Teacher characteristics from the students' viewpoint* (No. RMM-2). Austin, TX: Research and Development Center for Teacher Education, Texas University.
Ware, J. E., & Williams, R. G. (1975). The Doctor Fox effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education, 50*, 149-156.
Ware, J. E., & Williams, R. G. (1977). Discriminant analysis of student ratings as a means for identifying lecturers who differ in enthusiasm or information-giving. *Educational & Psychological Measurement, 37*, 627-639.
Watkins, D. A. (1990). Student ratings of tertiary courses for "alternative calendar" purposes. *Assessment & Evaluation in Higher Education, 15*, 12-21.
Watkins, D. A., Marsh, H. W., & Young, D. (1987). Evaluating tertiary teaching: A New Zealand perspective. *Teaching and Teacher Education, 3*, 41-53.

Weick, K. E. (1976). Educational organisations as loosely coupled systems. *Administrative Science Quarterly, 21*, 1-9.

Wenglinsky, H. (2000). *How teaching matters: Bringing the classroom back into discussions of teacher quality* (Policy Information Center Report). Princeton, NJ: Educational Testing Service and the Milken Family Trust.

Whitworth, J. E., Price, B. A., & Randall, C. H. (2002). Factors that affect college of business student opinion of teaching and learning. *Journal of Education for Business, 77*, 282-289.

Wilcox, D. D. (1999). *The National Board for Professional Teaching Standards: Can it live up to its promise?* Washington, DC: Thomas B Fordham Foundation.

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360° feedback for teacher evaluation. *Journal of Personnel Evaluation in Education, 14*, 179-192.

Wilkinson, I. A. G., & Hamilton, R. J. (2003). Learning to read in composite (multigrade) classes in New Zealand: Teachers make the difference. *Teaching and Teacher Education, 19*, 221-235.

Wilkinson, I. A. G., Hattie, J. A., Parr, J. M., Townsend, M. A. R., Thrupp, M., Lauder, H., et al. (2000). *Influence of peer effects on learning outcomes: a review of the literature*. Wellington: Research Division, New Zealand Ministry of Education.

Williams, R. G., & Ware, J. E. (1976). Validity of student ratings under different incentive conditions: A further study of the Doctor Fox effect. *Journal of Educational Psychology, 68*, 48-56.

Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10*, 253-267.

Williams, W. M., & Ceci, S. J. (1997). "How'm I doing?" Problems with student ratings of instructors and courses. *Change, 29*(5), 12-23.

Wilson, L. M. (1998). *Teachers seeking national board certifcation: A multi-state case study.* Unpublished doctoral thesis, Indiana State University, Terre Haute.

Wilson, R. C. (1986). Improving faculty teaching: Effective use of student evaluations and consultants. *Journal of Higher Education, 57*, 196-211.

Wilson, W. R. (1999). Students rating teachers. *Journal of Higher Education, 70*, 562-571.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education, 12*, 236-247.

Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment & Evaluation in Higher Education, 27*, 49-64.

Wragg, E. C., & Wood, E. K. (1984). Pupil appraisals of teaching. In E. C. Wragg (Ed.), *Classroom teaching skills*. Beckenham: Croom Helm.

Young, B. N., Whitley, M. E., & Helton, C. (1998, November). *Students' perceptions of characteristics of effective teachers.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA.

Young, I. P., Delli, D. A., & Johnson, L. (1999). Student evaluation of faculty. Effects of purpose on pattern. *Journal of Personnel Evaluation in Education, 13*, 179-190.

Younger, M., & Warrington, M. (1999). "He's such a nice man, but he's so boring, you have to really make a conscious effort to learn": The views of Gemma, Daniel and their contemporaries on teacher quality and effectiveness. *Educational Review, 51*, 231-241.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19*, 231-240.

Appendices

Appendix One

Assessing High School Mathematics Teachers

Student Questionnaire Form A

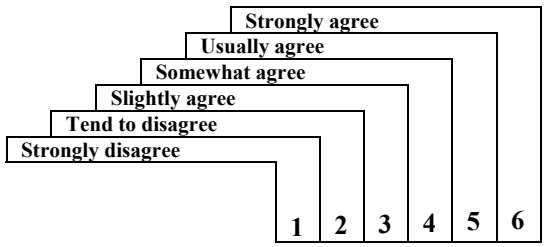
School: _____ M 0 F 0

Class/Year level: _____ M 0 E 0 P 0 A 0 O 0

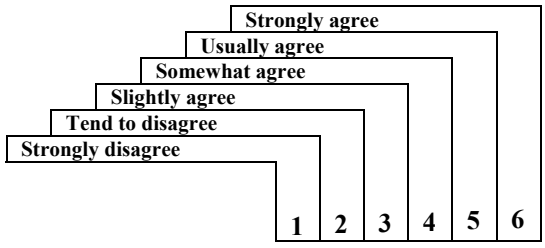
Please indicate the EXTENT of your disagreement/agreement with the following statements by using the following scale:

1 = Strongly disagree **2 = Tend to disagree** **3 = Slightly agree**
4 = Somewhat agree **5 = Usually agree** **6 =Strongly agree**

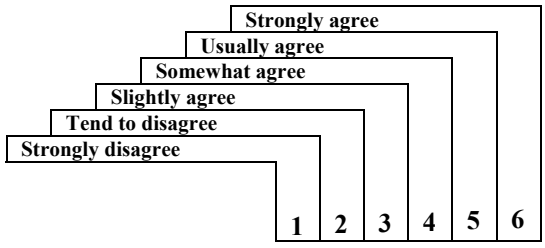
For each question, fill in *one* bubble completely with black/blue pen or pencil. Put a X through any mistake, and fill in the *one* bubble you want to be counted



	My mathematics teacher …		1	2	3	4	5	6
1.	*encourages all students to participate fully in class.*	*1.*	0	0	0	0	0	0
2.	*provides time to apply maths to a broad range of interesting subjects and applications.*	*2.*	0	0	0	0	0	0
3.	*encourages us to explore, confront and challenge new ideas presented in maths.*	*3.*	0	0	0	0	0	0
4.	*uses technology, activities and physical models to help us recognise the connections among different ways of representing ideas in maths.*	*4.*	0	0	0	0	0	0
5.	*helps us to make links between the different strands of maths and other aspects of our lives.*	*5.*	0	0	0	0	0	0
6.	*makes sure that all students participate in class regardless of their gender, ethnicity, cultural background, prior experience and expectations.*	*6.*	0	0	0	0	0	0
7.	*helps us to see the "big picture" by relating the themes in maths.*	*7.*	0	0	0	0	0	0
8.	*recognises settings in the real world where mathematical solutions are worthwhile.*	*8.*	0	0	0	0	0	0
9.	*seems to have a broad and deep understanding of the concepts, principles, techniques and reasoning methods of maths.*	*9.*	0	0	0	0	0	0



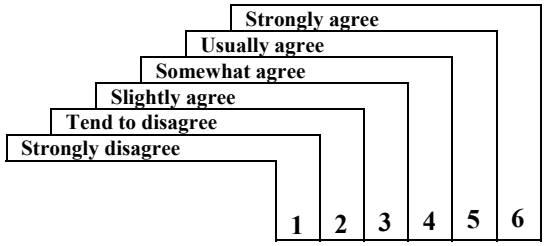
	My mathematics teacher …		1	2	3	4	5	6
10.	*shows us how to use indirect methods (like testing extreme cases, organised searches , etc.) to solve problems.*	*10.*	0	0	0	0	0	0
11.	*helps us to understand & appreciate the powerful relationships between mathematical ideas and problems.*	*11.*	0	0	0	0	0	0
12.	*regards technology as an essential component of teaching maths.*	*12.*	0	0	0	0	0	0
13.	*uses a variety of processes to describe patterns in different kinds of data.*	*13.*	0	0	0	0	0	0
14.	*asks us to explain our solutions to problems and justify our conclusions.*	*14.*	0	0	0	0	0	0
15.	*shows us how we can use geometry to solve problems in the real world.*	*15.*	0	0	0	0	0	0
16.	*has introduced us to a variety of new topics like fractals, linear programming, cracking codes and technology based numerical methods.*	*16.*	0	0	0	0	0	0
17.	*helps us to apply our growing knowledge in both pure and applied settings.*	*17.*	0	0	0	0	0	0
18.	*shows and challenges us to discover and describe patterns in visual, numerical and symbolic data.*	*18.*	0	0	0	0	0	0
19.	*helps us to understand mathematical concepts rather than routine computational procedures and proofs.*	*19.*	0	0	0	0	0	0
20.	*teaches us about the role that maths has in the history of problem-solving and decision-making across time and cultures.*	*20.*	0	0	0	0	0	0
21.	*helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	*21.*	0	0	0	0	0	0
22.	*has a classroom where we are engaged in learning.*	*22.*	0	0	0	0	0	0
23.	*teaches us that proof provides a standard of precision that sets maths apart from other subjects.*	*23.*	0	0	0	0	0	0
24.	*organises tasks that help us see the relationship between different ways of representing mathematical ideas.*	*24.*	0	0	0	0	0	0
25.	*involves us in maths competitions, fairs (e.g., Mathex) and other events that allow us to demonstrate our mathematical knowledge and skills.*	*25.*	0	0	0	0	0	0
26.	*shows us how we can use measurement to solve problems in the real world.*	*26.*	0	0	0	0	0	0
27.	*provides time to develop problem solving skills that we can use both in maths and outside the classroom.*	*27.*	0	0	0	0	0	0
28.	*invites us to question ideas, offer ideas of our own, and argue in support of them.*	*28.*	0	0	0	0	0	0



My mathematics teacher …

			1	2	3	4	5	6
29.	*provides problems and applications to develop the maths we have learned.*	*29.*	0	0	0	0	0	0
30.	*shows us interesting and useful ways of solving problems.*	*30.*	0	0	0	0	0	0
31.	*encourages us to try different techniques to solve problems.*	*31.*	0	0	0	0	0	0
32.	*helps us to effectively apply ideas in maths to solving problems in the everyday world (e.g., the scientific, technical, arts, music worlds).*	*32.*	0	0	0	0	0	0
33.	*helps us to build our own broad and deep understanding of maths.*	*33.*	0	0	0	0	0	0
34.	*encourages us to test mathematical ideas and discover mathematical principles.*	*34.*	0	0	0	0	0	0
35.	*shows us how we can use statistics to solve problems in the real world.*	*35.*	0	0	0	0	0	0
36.	*provides time for us to be involved in peer tutoring.*	*36.*	0	0	0	0	0	0
37.	*uses a variety of teaching methods to represent, solve and make decisions about real problems.*	*37.*	0	0	0	0	0	0
38.	*uses basic skills to solve more complex problems.*	*38.*	0	0	0	0	0	0
39.	*weaves together the pieces of maths to form a comprehensive and flowing mathematical experience.*	*39.*	0	0	0	0	0	0
40.	*distinguishes between different ways of solving a problem to illustrate the most efficient method.*	*40.*	0	0	0	0	0	0
41.	*teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.*	*41.*	0	0	0	0	0	0
42.	*shows us how the different strands of maths are linked together.*	*42.*	0	0	0	0	0	0
43.	*provides frequent opportunity for us to reflect on our own learning.*	*43.*	0	0	0	0	0	0
44.	*encourages us to seek more than one solution to problems.*	*44.*	0	0	0	0	0	0
45.	*shows us how we can use calculus to solve problems in the real world.*	*45.*	0	0	0	0	0	0
46.	*provides tasks that help us to see the many different ways of representing mathematical ideas & problems.*	*46.*	0	0	0	0	0	0
47.	*helps us to communicate better in maths.*	*47.*	0	0	0	0	0	0
48.	*tries out different ways of involving us in our learning of maths*	*48.*	0	0	0	0	0	0

.



	My mathematics teacher …		1	2	3	4	5	6
49.	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	*49.*	0	0	0	0	0	0
50.	*teaches us the fundamental processes of mathematical thinking – exploration, inference, interpretation, representation, modelling, conjecture and analysis.*	*50.*	0	0	0	0	0	0
51.	*encourages us to question our peers when discussing new ideas, and solving problems.*	*51.*	0	0	0	0	0	0
52.	*helps us construct an understanding of the language and processes of maths.*	*52.*	0	0	0	0	0	0
53.	*provides time for us to reflect on and talk about the maths we are learning.*	*53.*	0	0	0	0	0	0
54.	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	*54.*	0	0	0	0	0	0
55.	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	*55.*	0	0	0	0	0	0
56.	*uses a variety of methods to collect, organise, represent and summarise collections of data.*	*56.*	0	0	0	0	0	0
57.	*conveys to the class the idea that maths relates to the real world.*	*57.*	0	0	0	0	0	0
58.	*provides time for us to develop our own personal interests by formulating and solving our own problems.*	*58.*	0	0	0	0	0	0
59.	*uses rules to prove theorems and draw conclusions.*	*59.*	0	0	0	0	0	0
60.	*teaches us to use calculators and computers effectively for both routine and complex problems.*	*60.*	0	0	0	0	0	0
61.	*encourages us to question and discuss the mathematical ideas and concepts we are taught.*	*61.*	0	0	0	0	0	0
62.	*uses a variety of activities to involve each of us in our learning of maths.*	*62.*	0	0	0	0	0	0
63.	*shows us how we can use algebra to represent patterns and solve problems in the real world.*	*63.*	0	0	0	0	0	0
64.	*involves students in decisions about their learning of maths.*	*64.*	0	0	0	0	0	0
65.	*compared with all other maths teachers I have had, is the best.*	*65.*	0	0	0	0	0	0

Thank you for your assistance

Appendix Two

Assessing High School Mathematics Teachers

Student Questionnaire Form B

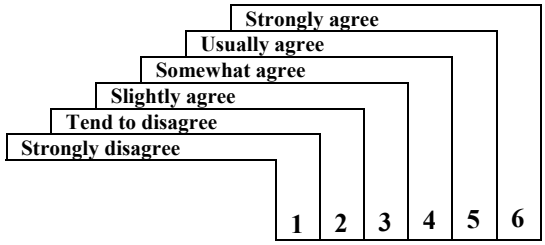
School: _____ M 0 F 0

Class/Year level: _____ M 0 E 0 P 0 A 0 O 0

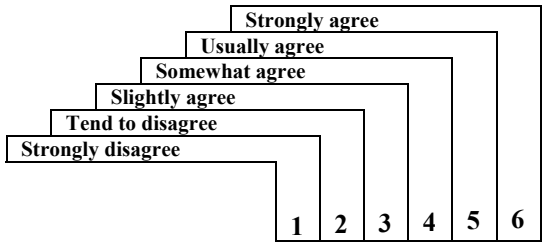
Please indicate the EXTENT of your disagreement/agreement with the following statements by using the following scale:

1 = Strongly disagree **2 = Tend to disagree** **3 = Slightly agree**
4 = Somewhat agree **5 = Usually agree** **6 =Strongly agree**

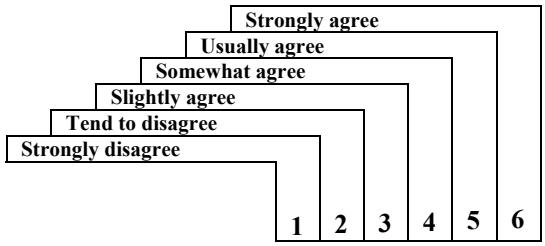
For each question, fill in *one* bubble completely with black/blue pen or pencil. Put a X through any mistake, and fill in the *one* bubble you want to be counted



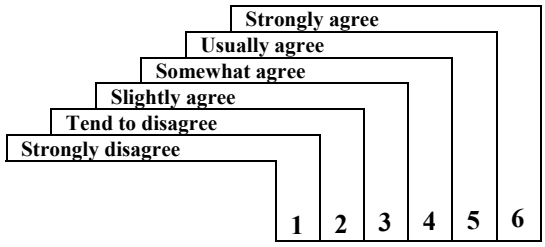
	My mathematics teacher …		1	2	3	4	5	6
1.	*cares about and values each individual in the class.*	*1.*	0	0	0	0	0	0
2.	*shares ideas in an open and positive way.*	*2.*	0	0	0	0	0	0
3.	*is alert and sensitive to our individual differences.*	*3.*	0	0	0	0	0	0
4.	*respects the contributions we make in our maths class.*	*4.*	0	0	0	0	0	0
5.	*makes all topics in maths interesting.*	*5.*	0	0	0	0	0	0
6.	*has good judgment and displays discretion.*	*6.*	0	0	0	0	0	0
7.	*often has new teachers visit our classroom to observe their teaching.*	*7.*	0	0	0	0	0	0
8.	*demonstrates their concern for us through their actions and words.*	*8.*	0	0	0	0	0	0
9.	*creates and maintains a learning environment by being flexible.*	*9.*	0	0	0	0	0	0
10.	*knows the students in this class really well.*	*10.*	0	0	0	0	0	0
11.	*determines and builds on each student's existing mathematical knowledge and understanding.*	*11.*	0	0	0	0	0	0
12.	*realises that not all students in the class have families who can assist them.*	*12.*	0	0	0	0	0	0
13.	*helps us to be confident in learning, doing and understanding maths.*	*13.*	0	0	0	0	0	0
14.	*understands and caters for students with different abilities in maths.*	*14.*	0	0	0	0	0	0
15.	*often has teacher trainees in our classroom.*	*15.*	0	0	0	0	0	0



	My mathematics teacher *…*		1	2	3	4	5	6
16.	*recognises that each student can obtain increased knowledge in maths.*	*16.*	0	0	0	0	0	0
17.	*teaches maths in a lively and enjoyable way.*	*17.*	0	0	0	0	0	0
18.	*uses interesting materials and resources that appeal to different people in the class.*	*18.*	0	0	0	0	0	0
19.	*knows what I can and can't do in maths.*	*19.*	0	0	0	0	0	0
20.	*understands the impact that home life, cultural background, community expectations, and student attitudes can have on our learning.*	*20.*	0	0	0	0	0	0
21.	*allows us to make mistakes without feeling bad.*	*21.*	0	0	0	0	0	0
22.	*creates a positive atmosphere in class where we feel part of a team of learners.*	*22.*	0	0	0	0	0	0
23.	*provides support and encouragement to all of the class.*	*23.*	0	0	0	0	0	0
24.	*holds my interest in class.*	*24.*	0	0	0	0	0	0
25.	*identifies and helps students with special needs or special abilities in maths*	*25.*	0	0	0	0	0	0
26.	*believes that all of the students in the class can learn and use significant mathematics.*	*26.*	0	0	0	0	0	0
27.	*empowers students to think through and solve problems both independently and together as a group.*	*27.*	0	0	0	0	0	0
28.	*recognises the beliefs and attitudes towards maths that each of us brings to the classroom.*	*28.*	0	0	0	0	0	0
29.	*uses their knowledge about each of us to create problems that are interesting and worth solving.*	*29.*	0	0	0	0	0	0
30.	*involves us and our family in exploring career opportunities.*	*30.*	0	0	0	0	0	0
31.	*focuses on the students in the class and their learning in mathematics.*	*31.*	0	0	0	0	0	0
32.	*understands and teaches according to the way that students learn maths.*	*32.*	0	0	0	0	0	0
33.	*is enthusiastic and enjoys teaching us maths.*	*33.*	0	0	0	0	0	0
34.	*stimulates our learning by varying the way we are taught to allow for the strengths and weaknesses of the people in the class.*	*34.*	0	0	0	0	0	0
35.	*uses a variety of ways to encourage and involve our families in our maths work.*	*35.*	0	0	0	0	0	0
36.	*chooses approaches to teaching that work for all students in the class.*	*36.*	0	0	0	0	0	0
37.	*works with other subject teachers to provide for every student in the class.*	*37.*	0	0	0	0	0	0
38.	*allows us to learn maths in different ways*	*38.*	0	0	0	0	0	0



	My mathematics teacher *…*		1	2	3	4	5	6
39.	*creates a welcoming environment that opens the class to family members and members of the community.*	*39.*	0	0	0	0	0	0
40.	*is committed to the learning of all the students in the class.*	*40.*	0	0	0	0	0	0
41.	*expects students to respect the contributions of other students in the class.*	*41.*	0	0	0	0	0	0
42.	*plays a part in keeping the community up to date with what is happening in maths.*	*42.*	0	0	0	0	0	0
43.	*is committed to the principle of equity/fairness in the way they treat all people* .	*43.*	0	0	0	0	0	0
44.	*looks to my family for information about my strengths, interests, habits and home life.*	*44.*	0	0	0	0	0	0
45.	*is able to explain something in different ways to help us understand.*	*45.*	0	0	0	0	0	0
46.	*provides a variety of options to allow for individual interests, aptitudes, knowledge and ways of learning.*	*46.*	0	0	0	0	0	0
47.	*makes maths come alive in the classroom.*	*47.*	0	0	0	0	0	0
48.	*makes learning maths satisfying and stimulating.*	*48.*	0	0	0	0	0	0
49.	*identifies and helps students with special abilities and special needs in maths, including those whose first language is not English.*	*49.*	0	0	0	0	0	0
50.	*listens to what the students have to say.*	*50.*	0	0	0	0	0	0
51.	*enables us to develop confidence and self esteem in maths.*	*51.*	0	0	0	0	0	0
52.	*understands the impact our individual backgrounds have on our learning.*	*52.*	0	0	0	0	0	0
53.	*provides time to build on previous knowledge, interests and understandings.*	*53.*	0	0	0	0	0	0
54.	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.*	*54.*	0	0	0	0	0	0
55.	*keeps my family informed about my progress in maths.*	*55.*	0	0	0	0	0	0
56.	*chooses a variety of approaches to teaching that work for the wide range of students in the class.*	*56.*	0	0	0	0	0	0
57.	*has a classroom where students are respected and feel safe to participate.*	*57.*	0	0	0	0	0	0
58.	*involves families, administrators and teachers in the school and community to help and support student to learn and continue in maths.*	*58.*	0	0	0	0	0	0
59.	*creates and maintains a learning environment by being well planned.*	*59.*	0	0	0	0	0	0
60.	*often attends and contributes to meetings of maths teachers.*	*60.*	0	0	0	0	0	0
61.	*makes everyone in the class believe that maths is for them.*	*61.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
62.	*creates-an-environment-for us to become self-directed and capable of learning maths on our own.*	*62.*	0	0	0	0	0	0
63.	*compared with all other maths teachers I have had, is the best.*	*63.*	0	0	0	0	0	0

Thank you for your assistance

Appendix Three

Assessing High School Mathematics Teachers

Student Questionnaire Form C

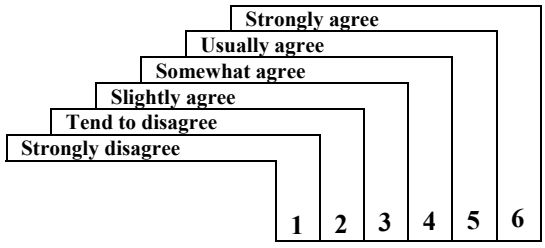
School: _____ M 0 F 0

Class/Year level: _____ M 0 E 0 P 0 A 0 O 0

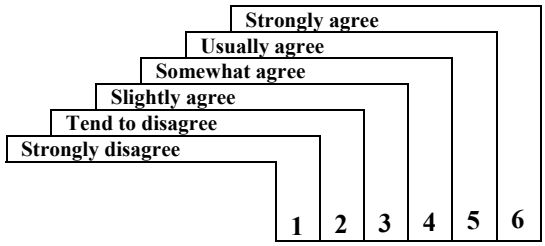
Please indicate the EXTENT of your disagreement/agreement with the following statements by using the following scale:

1 = Strongly disagree **2 = Tend to disagree** **3 = Slightly agree**
4 = Somewhat agree **5 = Usually agree** **6 =Strongly agree**

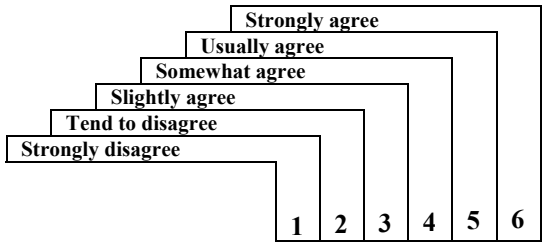
For each question, fill in *one* bubble completely with black/blue pen or pencil. Put a X through any mistake, and fill in the *one* bubble you want to be counted



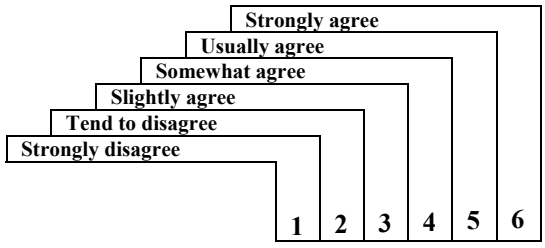
	My mathematics teacher …		1	2	3	4	5	6
1.	*makes maths meaningful for me.*	*1.*	0	0	0	0	0	0
2.	*places a high value on learning maths.*	*2.*	0	0	0	0	0	0
3.	*focuses all of the students on their work.*	*3.*	0	0	0	0	0	0
4.	*is able to use many different ways to get mathematical ideas across, like words, stories, numbers, diagrams, graphs and symbols.*	*4.*	0	0	0	0	0	0
5.	*chooses imaginative examples, problems and situations that motivate us.*	*5.*	0	0	0	0	0	0
6.	*uses group investigations to assess us.*	*6.*	0	0	0	0	0	0
7.	*uses examples from a wide range of fields to show how maths is related and useful.*	*7.*	0	0	0	0	0	0
8.	*explores ideas with us even if the answer is not known in advance.*	*8.*	0	0	0	0	0	0
9.	*integrates the goals of the curriculum and their knowledge of the students in the class.*	*9.*	0	0	0	0	0	0
10.	*gives us the chance to sensitively assess other students' work.*	*10.*	0	0	0	0	0	0
11.	*is not afraid of failure.*	*11.*	0	0	0	0	0	0
12.	*applies concepts in realistic settings.*	*12.*	0	0	0	0	0	0
13.	*presents new ideas they have found in journals and at conferences and meetings to help us expand our learning of maths.*	*13.*	0	0	0	0	0	0
14.	*uses examples from other school subjects and the outside world to help us understand new ideas in maths.*	*14.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
15.	*consistently makes decisions about their teaching that will further our learning.*	*15.*	0	0	0	0	0	0
16.	*teaches us high quality, important and meaningful maths.*	*16.*	0	0	0	0	0	0
17.	*encourages us to advance in maths as far as possible,*	*17.*	0	0	0	0	0	0
18.	*is fair in the way they assess each student in the class.*	*18.*	0	0	0	0	0	0
19.	*tells us that we are expected to do well in maths.*	*19.*	0	0	0	0	0	0
20.	*uses a wide variety of resources (e.g., speakers, historical material, the library, museum visits , etc.) to help us reach our mathematical goals.*	*20.*	0	0	0	0	0	0
21.	*motivates us to do our best work.*	*21.*	0	0	0	0	0	0
22.	*teaches us meaningful and important maths.*	*22.*	0	0	0	0	0	0
23.	*provides the inspiration for student investigations.*	*23.*	0	0	0	0	0	0
24.	*uses items that are in the news and relates them to our classwork.*	*24.*	0	0	0	0	0	0
25.	*gathers information from us and uses it to improve their teaching.*	*25.*	0	0	0	0	0	0
26.	*checks for student understanding before and at the end of each lesson.*	*26.*	0	0	0	0	0	0
27.	*helps us experience success in doing worthwhile maths.*	*27.*	0	0	0	0	0	0
28.	*models their own mathematical reasoning in all tasks, actions and discussions.*	*28.*	0	0	0	0	0	0
29.	*knows and caters for the problems we commonly encounter in learning new topics.*	*29.*	0	0	0	0	0	0
30.	*provides tasks that challenge us to think..*	*30.*	0	0	0	0	0	0
31.	*ensures that all students take courses that lead to increased mathematical knowledge.*	*31.*	0	0	0	0	0	0
32.	*encourages us to place a high value on maths.*	*32.*	0	0	0	0	0	0
33.	*encourages us to set high goals for ourselves in maths.*	*33.*	0	0	0	0	0	0
34.	*recognises and overcomes the barriers that prevent students from achieving success in maths.*	*34.*	0	0	0	0	0	0
35.	*emphasises the points we are expected to understand and learn.*	*35.*	0	0	0	0	0	0
36.	*uses a variety of techniques to maintain control of the students in this class.*	*36.*	0	0	0	0	0	0
37.	*provides useful feedback after each assessment.*	*37.*	0	0	0	0	0	0
38.	*takes calculated risks with the way a lesson might develop if the outcome might be beneficial.*	*38.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
39.	*skillfully combines their knowledge of adolescents, mathematics and how we learn to help us be successful in maths.*	*39.*	0	0	0	0	0	0
40.	*prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.*	*40.*	0	0	0	0	0	0
41.	*helps us to use various performance measures to monitor our progress in maths.*	*41.*	0	0	0	0	0	0
42.	*uses well defined goals to assess our work and learning.*	*42.*	0	0	0	0	0	0
43.	*asks questions and uses skilful probing to help classroom discussion and thinking.*	*43.*	0	0	0	0	0	0
44.	*expects us to learn maths even if we have different backgrounds and previous learning experiences.*	*44.*	0	0	0	0	0	0
45.	*illustrates the way that different cultures have contributed to the development of mathematics.*	*45.*	0	0	0	0	0	0
46.	*seems to modify their plans for the lesson if something interesting comes up.*	*46.*	0	0	0	0	0	0
47.	*provides enough work to keep all students in the class working.*	*47.*	0	0	0	0	0	0
48.	*gives us time to understand new ideas and progress to the next level of understanding.*	*48.*	0	0	0	0	0	0
49.	*intervenes when appropriate to help a student gain better understanding.*	*49.*	0	0	0	0	0	0
50.	*uses a blend of new and traditional methods to teach us.*	*50.*	0	0	0	0	0	0
51.	*keeps the interest of all the students in the class.*	*51.*	0	0	0	0	0	0
52.	*uses cooperative learning strategies and group work to help us learn and tackle substantial mathematical issues.*	*52.*	0	0	0	0	0	0
53.	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.*	*53.*	0	0	0	0	0	0
54.	*does not claim to have all of the answers.*	*54.*	0	0	0	0	0	0
55.	*uses assessment results to provide extra help/extension to appropriate students.*	*55.*	0	0	0	0	0	0
56.	*identifies individual strengths and weaknesses after each assessment.*	*56.*	0	0	0	0	0	0
57.	*uses examples from the history of mathematics to illustrate its development.*	*57.*	0	0	0	0	0	0
58.	*teaches us equally well in all strands of the mathematics curriculum (algebra, number, measurement, geometry, etc.).*	*58.*	0	0	0	0	0	0
59.	*uses an appropriate range of formal and informal assessments to monitor individual and class progress.*	*59.*	0	0	0	0	0	0
60.	*adjusts the lesson if we experience difficulties in learning.*	*60.*	0	0	0	0	0	0
61.	*tells us what the purpose of each lesson is.*	*61.*	0	0	0	0	0	0
62.	*gets us to think about the nature and quality of our work.*	*62.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
63.	*teaches us how to evaluate progress towards our goals.*	*63.*	0	0	0	0	0	0
64.	*uses examples that help us to understand and learn new ideas.*	*64.*	0	0	0	0	0	0
65.	*makes good use of time to optimise learning.*	*65.*	0	0	0	0	0	0
66.	*is fair in the way they assess all students.*	*66.*	0	0	0	0	0	0
67.	*encourages us to take risks and make mistakes.*	*67.*	0	0	0	0	0	0
68.	*compared with all other maths teachers I have had, is the best.*	*68.*	0	0	0	0	0	0

Thank you for your assistance

Appendix Four

Assessing High School Mathematics Teachers

Student Questionnaire Form November

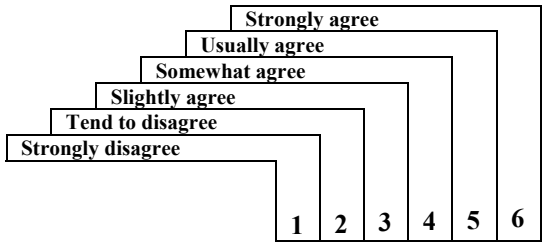
School: _____ M 0 F 0

Class/Year level: _____ M 0 E 0 P 0 A 0 O 0

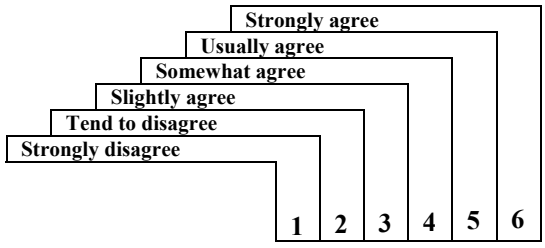
Please indicate the EXTENT of your disagreement/agreement with the following statements by using the following scale:

1 = Strongly disagree **2 = Tend to disagree** **3 = Slightly agree**
4 = Somewhat agree **5 = Usually agree** **6 =Strongly agree**

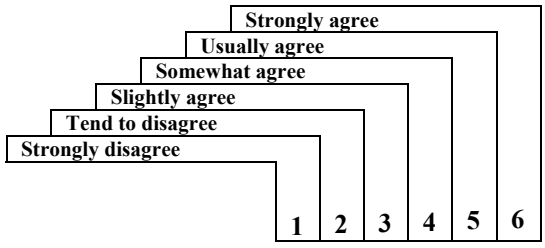
For each question, fill in *one* bubble completely with black/blue pen or pencil. Put a X through any mistake, and fill in the *one* bubble you want to be counted



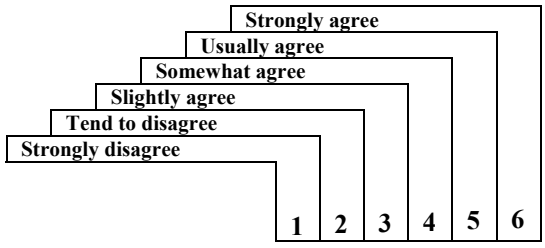
	My mathematics teacher …		1	2	3	4	5	6
1.	*makes maths come alive in the classroom.*	*1.*	0	0	0	0	0	0
2.	*help us to communicate better in maths.*	*2.*	0	0	0	0	0	0
3.	*uses an appropriate range of formal and informal assessments to monitor individual and class progress.*	*3.*	0	0	0	0	0	0
4.	*provides time to develop problem solving skills that we can use both in maths and outside the classroom.*	*4.*	0	0	0	0	0	0
5.	*skilfully asks questions to help classroom discussion and thinking.*	*5.*	0	0	0	0	0	0
6.	*ensures that all students take courses that lead to increased mathematical knowledge.*	*6.*	0	0	0	0	0	0
7.	*regards technology (e.g., calculators and computers) as an essential tool for teaching maths.*	*7.*	0	0	0	0	0	0
8.	*teaches us the fundamental processes of mathematical thinking – for example: exploration, interpretation, representation, modelling, and analysis.*	*8.*	0	0	0	0	0	0
9.	*shows us interesting and useful ways of solving problems.*	*9.*	0	0	0	0	0	0
10.	*understands the impact that home life, cultural background, community expectations and student attitudes can have on our learning.*	*10.*	0	0	0	0	0	0
11.	*enables us to develop confidence and self esteem in maths.*	*11.*	0	0	0	0	0	0
12.	*makes geometry interesting for me.*	*12.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
13.	*creates a positive atmosphere in class where we feel part of a team of learners.*	*13.*	0	0	0	0	0	0
14.	*adjusts the lesson if we experience difficulties in learning.*	*14.*	0	0	0	0	0	0
15.	*helps us make the links between the different strands of maths and other aspects of our lives.*	*15.*	0	0	0	0	0	0
16.	*helps us construct an understanding of the language and processes of maths.*	*16.*	0	0	0	0	0	0
17.	*uses assessment results to provide extra help/extension to appropriate students.*	*17.*	0	0	0	0	0	0
18.	*teaches us that maths is a "science of patterns" with the power to describe significant patterns from the real world.*	*18.*	0	0	0	0	0	0
19.	*makes calculus interesting for me.*	*19.*	0	0	0	0	0	0
20.	*helps the class to understand that maths relates to the real world.*	*20.*	0	0	0	0	0	0
21.	*encourages us to seek more than one solution to problems.*	*21.*	0	0	0	0	0	0
22.	*holds my interest in class.*	*22.*	0	0	0	0	0	0
23.	*makes learning maths satisfying and stimulating.*	*23.*	0	0	0	0	0	0
24.	*provides time for us to reflect and talk about the maths we are learning.*	*24.*	0	0	0	0	0	0
25.	*challenges students to think through and solve problems, either by themselves or together as a group.*	*25.*	0	0	0	0	0	0
26.	*encourages us to try different techniques to solve problems.*	*26.*	0	0	0	0	0	0
27.	*stimulates our learning by varying the way we are taught to allow for the strengths and weaknesses of the people in the class.*	*27.*	0	0	0	0	0	0
28.	*is committed to the learning of all the students in the class.*	*28.*	0	0	0	0	0	0
29.	*involves our families and other teachers in the school to help and support us to learn and continue in maths.*	*29.*	0	0	0	0	0	0
30.	*is able to explain something in different ways to help us understand.*	*30.*	0	0	0	0	0	0
31.	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile maths.*	*31.*	0	0	0	0	0	0
32.	*encourages us to question and discuss the mathematical ideas and concepts we are taught.*	*32.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
33.	*consistently makes decisions about their teaching that will further our learning.*	*33.*	0	0	0	0	0	0
34.	*explores ideas with us even if the answer is not known in advance.*	*34.*	0	0	0	0	0	0
35.	*integrates the goals of the curriculum and their knowledge of the students in the class.*	*35.*	0	0	0	0	0	0
36.	*sometimes involves us and our family in exploring career opportunities.*	*36.*	0	0	0	0	0	0
37.	*provides the inspiration for student investigations.*	*37.*	0	0	0	0	0	0
38.	*teaches us how to evaluate progress towards our goals.*	*38.*	0	0	0	0	0	0
39.	*uses examples that help us to understand and learn new ideas.*	*39.*	0	0	0	0	0	0
40.	*uses a variety of methods to collect, organise, represent and summarise collections of data.*	*40.*	0	0	0	0	0	0
41.	*makes maths meaningful for me.*	*41.*	0	0	0	0	0	0
42.	*uses interesting materials and resources that appeal to different people in the class.*	*42.*	0	0	0	0	0	0
43.	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	*43.*	0	0	0	0	0	0
44.	*knows and caters for the problems we commonly encounter in learning new topics.*	*44.*	0	0	0	0	0	0
45.	*allows us to learn maths in different ways.*	*45.*	0	0	0	0	0	0
46.	*encourages us to place a high value on maths.*	*46.*	0	0	0	0	0	0
47.	*creates a welcoming environment in the classroom for family members and members of the community.*	*47.*	0	0	0	0	0	0
48.	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in maths.*	*48.*	0	0	0	0	0	0
49.	*prepares us for adult life by helping us to see how important maths will be to our careers and to everyday life.*	*49.*	0	0	0	0	0	0
50.	*helps us to realise that maths is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	*50.*	0	0	0	0	0	0
51.	*makes statistics interesting for me.*	*51.*	0	0	0	0	0	0
52.	*uses well defined goals to assess our work and learning.*	*52.*	0	0	0	0	0	0
53.	*works with other subject teachers to provide for students in the class.*	*53.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
54.	*uses their knowledge about each of us to create problems that are interesting and worth solving.*	*54.*	0	0	0	0	0	0
55.	*seeks information from my family about my strengths, interests, habits and home life.*	*55.*	0	0	0	0	0	0
56.	*helps us experience success in doing worthwhile maths.*	*56.*	0	0	0	0	0	0
57.	*teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.*	*57.*	0	0	0	0	0	0
58.	*applies concepts in realistic settings.*	*58.*	0	0	0	0	0	0
59.	*gets us to think about the nature and quality of our work.*	*59.*	0	0	0	0	0	0
60.	*tells us what the purpose of each lesson is.*	*60.*	0	0	0	0	0	0
61.	*encourages us to test mathematical ideas and discover mathematical principles.*	*61.*	0	0	0	0	0	0
62.	*helps us apply our growing knowledge in both pure and applied settings.*	*62.*	0	0	0	0	0	0
63.	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	*63.*	0	0	0	0	0	0
64.	*keeps my family informed on a regular basis about my progress in maths.*	*64.*	0	0	0	0	0	0
65.	*makes algebra interesting for me.*	*65.*	0	0	0	0	0	0
66.	*compared with all other maths teachers I have had, is the best.*	*66.*	0	0	0	0	0	0

Thank you for your assistance.

Appendix Five

Assessing High School Mathematics Teachers

Student Questionnaire Form Technology

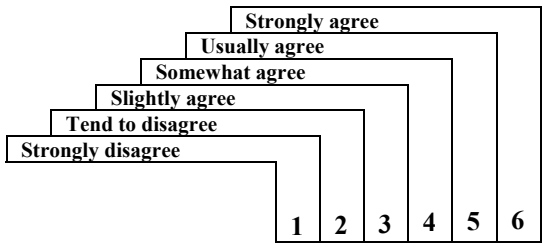
School: _____ M 0 F 0

Class/Year level: _____ M 0 E 0 P 0 A 0 O 0

Please indicate the EXTENT of your disagreement/agreement with the following statements by using the following scale:

1 = Strongly disagree **2 = Tend to disagree** **3 = Slightly agree**
4 = Somewhat agree **5 = Usually agree** **6 =Strongly agree**

For each question, fill in *one* bubble completely with black/blue pen or pencil. Put a X through any mistake, and fill in the *one* bubble you want to be counted



	My mathematics teacher …		1	2	3	4	5	6
1.	enjoys teaching maths using a computer.	*1.*	0	0	0	0	0	0
2.	is very confident when using a computer in our maths lessons.	*2.*	0	0	0	0	0	0
3.	enjoys the challenge of using a computer to solve problems.	*3.*	0	0	0	0	0	0
4.	integrates the use of calculators and computers into their teaching of maths.	*4.*	0	0	0	0	0	0
5.	believes that calculators can help us to learn maths.	*5.*	0	0	0	0	0	0
6.	uses computer and calculator technology to enhance remedial instruction.	*6.*	0	0	0	0	0	0
7.	uses calculators and computers to motivate us.	*7.*	0	0	0	0	0	0
8.	uses computers to help us work with each other.	*8.*	0	0	0	0	0	0
9.	uses modern technology (e.g., computers, calculators, internet) to help us learn maths.	*9.*	0	0	0	0	0	0
10.	makes having computers available in maths fun.	*10.*	0	0	0	0	0	0
11.	uses e-mail and the internet to provide a better learning environment.	*11.*	0	0	0	0	0	0
12.	extends our understanding in maths by using challenging computer-based problems.	*12.*	0	0	0	0	0	0
13.	regards technology (e.g., calculators and computers) as an essential tool for teaching maths.	*13.*	0	0	0	0	0	0
14.	teaches us about the way that maths contributes to technological changes in society, and the way that technology has changed maths.	*14.*	0	0	0	0	0	0

Thank you for your assistance

Appendix Six

Students Evaluating Accomplished Teaching - Mathematics

Student Questionnaire Form

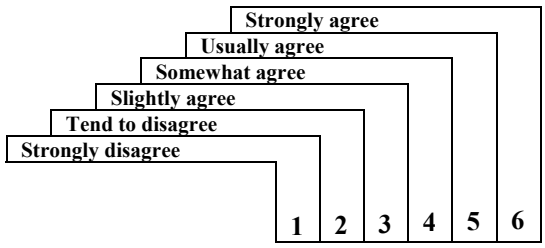
School: _____ M 0 F 0

Class/Year level: _____ C 0 AA 0 H 0 As 0 N 0 O 0

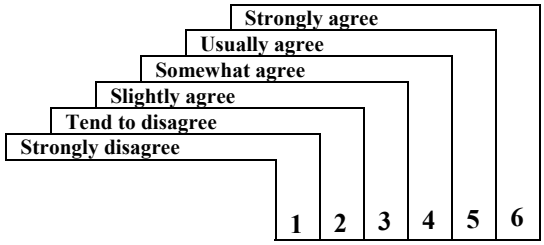
Please indicate the EXTENT of your disagreement/agreement with the following statements by using the following scale:

1 = Strongly disagree **2 = Tend to disagree** **3 = Slightly agree**
4 = Somewhat agree **5 = Usually agree** **6 =Strongly agree**

For each statement, fill in *one* bubble completely with black/blue pen or pencil. If you change your mind, put a cross (X) through that response, and fill in the *one* bubble you want to be counted.

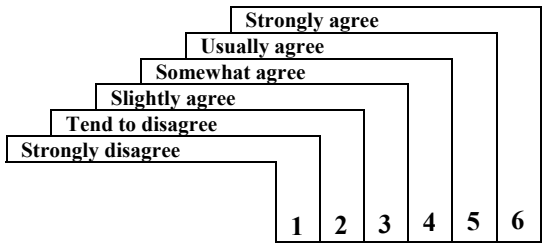


	My mathematics teacher …		1	2	3	4	5	6
1.	*makes math come alive in the classroom.*	*1.*	0	0	0	0	0	0
2.	*skillfully asks questions to help classroom discussion and thinking.*	*2.*	0	0	0	0	0	0
3.	*teaches us the fundamental processes of mathematical thinking – for example: exploration, interpretation, representation, modelling, and analysis.*	*3.*	0	0	0	0	0	0
4.	*shows us interesting and useful ways of solving problems.*	*4.*	0	0	0	0	0	0
5.	*enables us to develop confidence and self esteem in math.*	*5.*	0	0	0	0	0	0
6.	*makes geometry interesting for me.*	*6.*	0	0	0	0	0	0
7.	*creates a positive atmosphere in class where we feel part of a team of learners.*	*7.*	0	0	0	0	0	0
8.	*adjusts the lesson if we experience difficulties in learning.*	*8.*	0	0	0	0	0	0
9.	*helps us make the links between the different strands of math and other aspects of our lives.*	*9.*	0	0	0	0	0	0

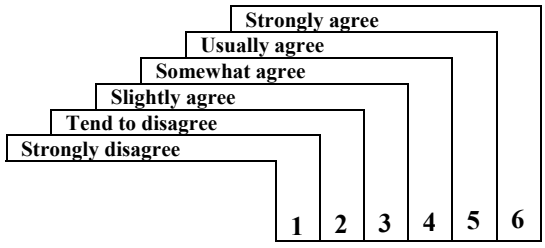


	My mathematics teacher …		1	2	3	4	5	6
10.	*helps us construct an understanding of the language and processes of math.*	*10.*	0	0	0	0	0	0
11.	*uses assessment results to provide extra help/extension to appropriate students.*	*11.*	0	0	0	0	0	0
12.	*teaches us that math is a "science of patterns" with the power to describe significant patterns from the real world.*	*12.*	0	0	0	0	0	0
13.	*makes calculus interesting for me.*	*13.*	0	0	0	0	0	0
14.	*helps the class to understand that math relates to the real world.*	*14.*	0	0	0	0	0	0
15.	*encourages us to seek more than one solution to problems.*	*15.*	0	0	0	0	0	0
16.	*makes learning math satisfying and stimulating.*	*16.*	0	0	0	0	0	0
17.	*provides time for us to reflect and talk about the math we are learning.*	*17.*	0	0	0	0	0	0
18.	*challenges students to think through and solve problems, either by themselves or together as a group.*	*18.*	0	0	0	0	0	0
19.	*encourages us to try different techniques to solve problems.*	*19.*	0	0	0	0	0	0
20.	*is committed to the learning of all the students in the class.*	*20.*	0	0	0	0	0	0
21.	*involves our families and other teachers in the school to help and support us to learn and continue in math.*	*21.*	0	0	0	0	0	0
22.	*uses different ways of teaching to help us understand.*	*22.*	0	0	0	0	0	0
23.	*sequences each lesson in a way that makes sense to us, making it possible for everyone to learn worthwhile math.*	*23.*	0	0	0	0	0	0
24.	*consistently makes decisions about their teaching that will further our learning.*	*24.*	0	0	0	0	0	0

Please turn over for remaining items



	My mathematics teacher …		1	2	3	4	5	6
25.	*explores ideas with us even if the answer is not known in advance.*	*25.*	0	0	0	0	0	0
26.	*integrates the goals of the curriculum and their knowledge of the students in the class.*	*26.*	0	0	0	0	0	0
27.	*sometimes involves us and our family in exploring career opportunities.*	*27.*	0	0	0	0	0	0
28.	*teaches us how to evaluate progress towards our goals.*	*28.*	0	0	0	0	0	0
29.	*uses examples that help us to understand and learn new ideas.*	*29.*	0	0	0	0	0	0
30.	*uses a variety of methods to collect, organize, represent and summarize collections of data.*	*30.*	0	0	0	0	0	0
31.	*uses interesting materials and resources that appeal to different people in the class.*	*31.*	0	0	0	0	0	0
32.	*teaches us about the fundamental role of proof in establishing the truth of mathematical statements.*	*32.*	0	0	0	0	0	0
33.	*knows and caters for the problems we commonly encounter in learning new topics.*	*33.*	0	0	0	0	0	0
34.	*encourages us to place a high value on math.*	*34.*	0	0	0	0	0	0
35.	*creates a welcoming environment in the classroom for family members and members of the community.*	*35.*	0	0	0	0	0	0
36.	*takes extra steps to ensure that all students (regardless of their ability) learn and achieve success in math.*	*36.*	0	0	0	0	0	0
37.	*prepares us for adult life by helping us to see how important math will be to our careers and to everyday life.*	*37.*	0	0	0	0	0	0
38.	*helps us to realize that math is continuously evolving and growing to make sense of the world – its order, chaos, stability and change.*	*38.*	0	0	0	0	0	0
39.	*makes statistics interesting for me.*	*39.*	0	0	0	0	0	0



	My mathematics teacher …		1	2	3	4	5	6
40.	*works with other subject teachers to provide for students in the class.*	*40.*	0	0	0	0	0	0
41.	*seeks information from my family about my strengths, interests, habits and home life.*	*41.*	0	0	0	0	0	0
42.	*teaches us about the way that math contributes to technological changes in society, and the way that technology has changed math.*	*42.*	0	0	0	0	0	0
43.	*applies concepts in realistic settings.*	*43.*	0	0	0	0	0	0
44.	*gets us to think about the nature and quality of our work.*	*44.*	0	0	0	0	0	0
45.	*tells us what the purpose of each lesson is.*	*45.*	0	0	0	0	0	0
46.	*encourages us to test mathematical ideas and discover mathematical principles.*	*46.*	0	0	0	0	0	0
47.	*helps us apply our growing knowledge in both pure and applied settings.*	*47.*	0	0	0	0	0	0
48.	*develops our ability to think and reason mathematically, and have a mathematical point of view.*	*48.*	0	0	0	0	0	0
49.	*keeps my family informed on a regular basis about my progress in math.*	*49.*	0	0	0	0	0	0
50.	*makes algebra interesting for me.*	*50.*	0	0	0	0	0	0
51.	*compared with all other math teachers I have had, is the best.*	*51.*	0	0	0	0	0	0

Thank you for your assistance.